Chapter- 15

STATISTICS

Data: - Facts or figures, collected with a definite purpose are called data. Data can be of two types.

Ungrouped Data: - In ungrouped data, data is listed in series e.g. 1, 4, 5, 6, 12, 13, etc. This is also called individual data

Grouped Data:- It is two types

(a) **Discrete data: -** In this type, data is presented in such a way that exact measurements of items are clearly shown. e.g. 15 students of class XI have secured the following marks

Marks	Frequency	Marks	Frequency
11	3	14	4
12	1	15	2
13	5		

(b) **Continuous group data: -** In this type, data is arranged in groups or classes but they are not exactly measurable, they form a continuous series. e.g.

Marks obtained	Number of students
0 – 10	5
10 – 20	7
20 – 30	13
30 – 40	20

Measures of Central tendency:- A certain value that represents the whole data and signifying its characteristics is called a measure of central tendency. Mean or average, median and mode are the measures of central tendency.

(1) **Mean (Arithmetic Mean):-** This arithmetic mean (or simple mean) of a set of observations is obtained by dividing the sum of the values of observations by the number of observations.

Mean of ungrouped data:- The mean of n observation $x_1, x_2, x_3, ...x_n$ is given by

$$\text{Mean}(\bar{x}) = \frac{x_1 + x_2 + x_3 + ...+ x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mean of Grouped Data:-

(i) The direct method, let $x_1, x_2, ...x_n$ be n observations with respective frequencies $f_1, f_2, ...f_n$.

$$\text{Then Mean}(\bar{x}) = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

(ii) Assumed mean method

$$\text{Mean} = a + \frac{\sum f_i d_i}{\sum f_i}$$

Where, a = assumed mean, $d_i = x_i - a =$ deviation from assumed mean

(iii) Step deviation method

$$\text{Mean}(\bar{x}) = a + \frac{\sum f_i u_i}{\sum f_i} \times h$$

Where, a = assumed mean, $u_i = \frac{x_i - h}{b}$ and h = width of the class interval

(2) **Median:-** Median is defined as the middlemost or the central value of the observations when the observations are arranged either in ascending or descending order of their magnitude.

Then

(i) If n is odd, Median = Value of the $\left(\frac{n+1}{2}\right)^{th}$ observation

(ii) If n is even, Median $= \dfrac{\text{sum of values of the} \left(\frac{n}{2}\right)^{th} \text{ and} \left(\frac{n}{2}+1\right)^{th} \text{ observations}}{2}$

Median of Grouped Data:-

(1) For discrete Data first, arrange the data in ascending or descending order and find the cumulative frequency, Now, find $\frac{N}{2}$, where $N = \sum f_i$

After that, find the median by using the following formula.

(i) If $\sum f_i = N$ is even, then

$$\text{Median} = \frac{\text{value of} \left(\frac{N}{2}\right)^{th} \text{ observation} + \text{value of} \left(\frac{N}{2}+1\right)^{th} \text{ observation}}{2}$$

(ii) If $\sum f_i = N$ is odd, then $\text{Median} = \text{value of} \left(\frac{N+1}{2}\right)^{th}$ observation

(2) For continuous data first, arrange the data in ascending or descending order and find the cumulative frequencies of all the classes.

Now, find $\frac{N}{2}$, where, $N = \sum f_i$

Further, find the class interval, whose cumulative frequency is just greater than or equal to N/2.

Then, median = $\ell + \dfrac{\left(\dfrac{N}{2} - cf\right)}{f} \times b$

Where ℓ = Lower limit of the median class

N = Number of observations

cf = Cumulative frequency of class preceding the median class

f = Frequency of the median class

b = Class width (assuming class size to be equal)

The measure of Dispersion:- The measures of central tendency are not sufficient to give complete information about the given data. Variability is another fact which is required to be studied under statistics. The single number that describes variability is called a measure of dispersion. It is the measure of spreading (scatter of the data about some central tendency. The dispersion or scatter in data is measured based on the observations and the types of measures of central tendency used. Three are following measures of dispersion

(a) Range (b) Quartile deviation (c) Mean deviation (d) Standard deviation

Note:- In this chapter quartile deviation will not be discussed. It is not in the syllabus.

Range:- Range is the difference of maximum and minimum values of data

Range = Maximum value – Minimum value

Mean deviation:- Mean deviation is an important measure of dispersion, which depends upon the deviations of the observations from a central tendency. Mean deviation is defined as the arithmetic mean of the absolute deviations of all the values taken from a central value or a fixed number 'a'. The mean deviation from 'a' is denoted as MD (a). Thus, the mean deviation from 'a'.

$$MD(a) = \frac{[\text{sum of absolute values of deviations from a}]}{\text{Number of observaions}}$$

Note:- Mean deviation may be obtained from any measure of central tendency. But in this chapter, we study deviation from mean and median.

Mean Deviation for Ungrouped Data:-

Let $x_1, x_2, ...x_n$ be n observations. Then, the mean deviation about the mean or median can be determined by using the following steps.

Step - I, Find the mean or median of given observations using a suitable formula.

Step – II, Find the deviation of each observation x_i from $\bar{x}$ (mean) or M (medium) and then take their absolute value i.e $|x_i - \bar{x}|$ or $|x_i - M|$.

Step – III, Find the sum of absolute values of deviations obtained in step III

i.e $\sum_{i=1}^{n}|x_i - \bar{x}|$ or $\sum_{i=1}^{n}|x_i - M|$

Step – IV, Now, find mean deviation about mean or median by using the formula

$$\frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n} \text{ or } \frac{\sum_{i=1}^{n}|x_i - M|}{n}$$ where n is the number of observations.

Example:-1

Find the mean deviation from the mean for the following data

6.5, 5, 5.25, 5.5, 4.75, 4.5, 6.25, 7.75, 8.5

Solution:- Given observations are 6.5, 5, 5.25, 5.5, 4.75, 4.5, 6.25, 7.75, 8.5

Here number of observation n = 9

Let $\bar{x}$ be the mean of given data

Then, $\bar{x} = \frac{6.5+5+5.25+5.5+4.75+6.25+7.75+8.5}{9} = \frac{54}{9} = 6$

Let us make the table for deviation and absolute deviation.

x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$
6.5	0.5	0.50
5.0	−1	1.00
5.25	−0.75	0.75
5.5	−0.5	0.50
4.75	−1.25	1.25
4.5	−1.50	1.50
6.25	0.25	0.25
7.75	1.75	1.75
8.5	2.5	2.50
Total		$\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert = 10.00$

$\therefore$ Mean deviation about the mean, $MD(\bar{x}) = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{9} = \frac{10}{9} = 1.1$

Hence, the mean deviation about mean is 1.1

Example:- 2

Find the mean deviation about the median for the data 34, 66, 30, 38, 44, 50, 40, 60, 42, 51.

Solution:- The given data can be arranged in ascending order as 30, 34, 38, 40, 42, 44, 50, 51, 60, 66.

Here, the total number of observations is 10. i.e n = 10, which is even

$\therefore$ median

$$(M)=\frac{\left(\frac{n}{2}\right)\text{th observation}+\left(\frac{n}{2}+1\right)\text{th observation}}{2}$$

$$=\frac{\left(\frac{10}{2}\right)\text{th observation}+\left(\frac{10}{2}+1\right)\text{thobservation}}{2}$$

$$=\frac{(5\text{th observation}+6\text{thobservation})}{2}$$

$$=\frac{42+44}{2}=\frac{86}{2}=43$$

Let us make the table for absolute deviation

x_1	$\lvert x_1-M\rvert$
30	$\lvert 30-43\rvert=13$
34	$\lvert 34-43\rvert=9$
38	$\lvert 38-43\rvert=5$
40	$\lvert 40-43\rvert=3$
42	$\lvert 42-43\rvert=1$
44	$\lvert 44-43\rvert=1$
50	$\lvert 50-43\rvert=7$
51	$\lvert 51-43\rvert=8$
60	$\lvert 60-43\rvert=17$
66	$\lvert 66-43\rvert=23$
Total	$\sum_{i=1}^{10}\lvert x_i-M\rvert=87$

Now, mean deviation about median, $MD=\frac{\sum_{i=1}^{n}\lvert x_i-M\rvert}{10}=\frac{87}{10}=8.7$

Mean Deviation for Grouped Data:-

(A) For discrete frequency distribution:- Let the given data have n distinct values $x_1, x_2, ...x_n$ and their corresponding frequencies are $f_1, f_2..........f_n$, respectively. Then this data can be represented in the tabular form, as

x_i	x_1	x_2	x_3	....	x_n
f_i	f_1	f_2	f_3	....	f_n

and is called a discrete frequency distribution.

Here, mean deviation about mean or median is given by $\dfrac{\sum_{i=1}^{n} f_i|x_i - A|}{N}$ where $N = \sum_{i=1}^{n} f_i =$ total frequency and $A =$ mean or median

The working rule for finding the mean deviation about mean:-

For finding the mean deviation about the mean, we use the following working steps

Step – 1, Find the mean of given observations using the formula $\bar{x} = \dfrac{\sum f_i x_i}{\sum f_i}$

Step – 2, Find the deviation of each observation x_i from $\bar{x}$ and take their absolute values i.e $|x_i - \bar{x}|$ and then find $f_i|x_i - \bar{x}|$

Step – 3, Find the sum of absolute values of deviations obtained in step II, i.e. $\sum f_i|x_i - \bar{x}|$

Step – 4, Now, find the mean deviation about mean by using the formula, $\dfrac{\sum f_i|x_i - \bar{x}|}{\sum f_i}$

Example:- 3

Find the 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 the following table from the given data

x_i	f_i	$f_i x_i$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
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
Total	40	300		92

Here, $N = \sum f_i = 40, \sum f_i x_i = 300$

Now mean $(\bar{x}) = \dfrac{1}{N}\sum f_i x_i = \dfrac{1}{40} \times 300 = 7.5$

$\therefore$ Mean deviation about the mean.

$$MD(\bar{x}) = \frac{1}{N}\sum f_i|x_i - \bar{x}| = \frac{1}{40} \times 92 = 2.3$$

Hence, the mean deviation about mean is 2.3

The working rule for finding the mean deviation about Median:-

For finding the mean deviation about median, we use the following working steps

Step – 1, Arrange the given data either in ascending order or descending order.

Step – 2, Make a cumulative frequency table.

Step – 3, Find the median by using the formula

(i) If $\sum f_i = N$ is even, then

$$\text{Median} = \frac{\text{value of}\left(\frac{N}{2}\right)^{th}\text{observation} + \text{value of}\left(\frac{N}{2}+1\right)^{th}\text{observation}}{2}$$

(ii) If $\sum f_i = N$ is odd, then Median = value of $\left(\frac{N+1}{2}\right)^{th}$ observation

Step – 4, Find the absolute values of deviations of observations x_i from M.

Step – 5, Find the product of frequency with absolute deviation i.e $f_i|x_i - M|$

Step – 6, Find the mean deviation from the median by using the formula, $MD = \frac{\sum f_i|x_i - M|}{\sum f_i}$

Example:-4

Find the mean deviation from the median of the following frequency distribution.

Age (in year)	10	12	15	18	21	23
Frequency	3	5	4	10	8	4

Solution:- The given observation are already in ascending order. New, let us make the cumulative frequency.

Age (x_i)	Frequency (f_i)	cf
10	3	3
12	5	8
15	4	12
18	10	22
21	8	30
23	4	34
Total	$N = 34$	

Here, $\sum f_i = N = 34$, which is even.

$$\therefore \text{Median} = \frac{\text{value of}\left(\frac{34}{2}\right)^{th}\text{observation} + \text{Value of}\left(\frac{34}{2}+1\right)^{th}\text{observation}}{2}$$

$$= \frac{\text{value of } 17^{th}\text{ observation} + \text{value of } 18^{th}\text{ observation}}{2}$$

$$=\frac{18+18}{2}=18$$

$\therefore$ Both of these observation lies in the cumulative frequency 22 and its corresponding observation is 18. Now, let us make the following table from the given data.

$\lvert x_i-18\rvert$	8	6	3	0	3	5	Total
$f_i\lvert x_i-18\rvert$	24	30	12	0	24	20	110

$$=\frac{\sum f_i\lvert x_i-M\rvert}{\sum f_i}=\frac{110}{34}=3.24\,\text{year}$$

For continuous frequency distribution:- A continuous frequency distribution is a series in which the data is classified into different class intervals without gaps along with their respective frequencies.

Mean Deviation about Mean:- For calculating the mean deviation from the mean of a continuous frequency distribution, the procedure is the same as for discrete frequency distribution. The only difference is that here we have to obtain the mid-points of the various classes and take the deviations of the mid-points from the mean.

Example:-5

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:- Let us make the following table from the given data

Marks obtained	Number of students (f_i)	Mid – points (x_i)	f_ix_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	47	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
Total	40		1800		400

Here, $N\sum_{i=1}^{7} f_i=40$ and $\sum_{i=1}^{7} f_ix_i=1800$

Therefore, the mean $(\bar{x})=\frac{1}{N}\sum_{i=1}^{7} f_1x_i=\frac{1800}{40}=45$

Now, the mean deviation $MD(\bar{x})=\frac{1}{N}\sum_{i=1}^{7} f_i\lvert x_i-\bar{x}\rvert=\frac{1}{40}\times 400=10$

Hence, the mean deviation about the mean is 10.

Mean Deviation About Median:- For calculating the mean deviation from the median of a continuous frequency distribution, the procedure is the same as about mean. The only difference is that here we replace the mean by median and the median is calculated by the following formula.

$$M = \ell + \frac{\frac{N}{2} - cf}{f} \times h$$

ℓ,f,handcf are respectively the lower limit, the frequency, the width of the median class, and cumulative frequency of class just preceding the median class?

Example:-6

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

Solution:- Let us make the following table from the given data.

Class	Mid – value (x_i)	Frequency (f_i)	Cumulative frequcency(cf)	$\lvert x_i - 14 \rvert$	$f_i \lvert x_i - 14 \rvert$
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
Total		$N = \sum f_i = 44$			$\sum f_i \lvert x_i - 14 \rvert = 278$

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

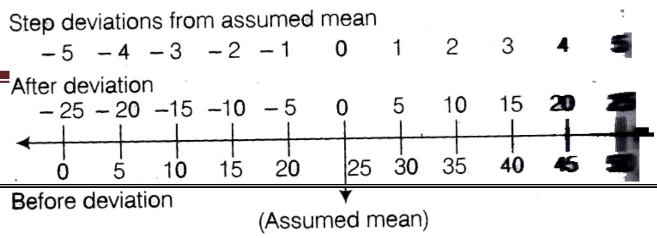
Now median = $\ell + \frac{\frac{N}{2} - cf}{f} \times h$

Where, $\ell = 12$, $f = 12$, $cf = 18$ and $h = 6$

$\therefore$ Median = $12 + \frac{20 - 18}{12} \times 6 = 12 + \frac{4 \times 6}{12} = 14$

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

Shortcut Method for Calculating the Mean Deviation about Mean:-



Sometimes, the data is too large and then the calculation by the previous method is tedious. So, we apply the step deviation method. In this method, we take an assumed mean, which is in the middle or just close to it, in the data. The process of taking step deviation is the change of scale on the number line as shown in the figure given below.

For step deviation method, we denote the new variable by u_i, and it is defined as $u_i = \frac{x_i - a}{h}$

Where a is the assumed mean and h is the common factor or length of the class interval. The mean $\overline{x}$ by step deviation method is given by $\overline{x} = a + \frac{\sum_{i=1}^{n} f_i u_i}{N} \times h$

Example:-7

Find the mean deviation from the mean of the following data by shortcut or step deviation method.

Class	0 – 100	100 – 200	200 – 300	300 – 400	400 – 500	500 – 600	600 – 700	700 – 800
Frequency	04	08	09	10	07	05	04	03

Solution:- Let us make the following table for step deviation and product of frequency with absolute deviation.

Class	(f_i)	Mid – points (x_i)	$u_i = \frac{x_i - 450}{100}$ $(a = 450, n = 100)$	$f_i u_i$	$\lvert x_i - \overline{x} \rvert = \lvert x_i - 358 \rvert$	$f_i \lvert x_i - \overline{x} \rvert$
0 – 100	4	50	–4	–16	308	1232
100 – 200	8	150	–3	–24	208	1664
200 – 300	9	250	–2	–18	108	972
300 – 400	10	350	–1	–10	8	80
400 – 500	7	450	0	0	92	644
500 – 600	5	550	1	5	192	960
600 – 700	4	650	2	8	292	1168
700 – 800	3	750	3	9	392	1176
Total	$N = \sum f_i = 50$			–46		7896

Here, $a = 450$, $\sum f_i u_i = -46$ and $h = 100$

$$\overline{x} = a + h\left(\frac{1}{N}\sum f_i u_i\right) = 450 + 100 \times \left(-\frac{46}{50}\right) = 358$$

Now, the mean deviation = $\frac{1}{N}\sum f_i \lvert x_i - \overline{x} \rvert = \frac{7896}{50} = 157.92$

Limitations of Mean deviation:-

The following limitations of mean deviations are given below

(a) If the data is more scattered or the degree of variability is very high, then the median is not a valid representative. Thus, the mean deviation about the median is not fully relied on.

(b) The sum of the deviations from the mean is more than the sum of the deviations from the median. Therefore, the mean deviation about mean is not very scientific

(c) The mean deviation is calculated based on absolute values of the deviations and so cannot be subjected to further algebraic treatment. Sometimes, it gives unsatisfactory results.

Variance and Standard Deviation:-

Due to the limitations of mean deviation, some other method is required for the measure of dispersion. Standard deviation is such a measure of dispersion.

Variance:- The absolute values are considered in calculating the mean deviation about mean or median, otherwise the deviation being negative or positive and may cancel among themselves. To overcome this difficulty of the signs of the deviations, we take the squares of all the deviations, so that all deviations become one-negative.

Let $x_i, x_2 ....x_n$ be n observations and $\bar{x}$ be their mean. Then, $(x_1-\bar{x})^2+(x_2-\bar{x})^2+.........+(x_n-\bar{x})^2$

$$=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Definition:- The mean of the squares of the deviations from mean is called the variance and it is denoted by the symbol σ^2.

Standard Deviation:-

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

Or

The square root of the variance, is called the standard deviation. i.e $\sqrt{\sigma^2}$ or σ

It is also known as the root mean square deviation.

Variance and standard deviation of ungrouped data:-

The variance of n observations $x_1, x_2, x_3, ...x_n$ is given by

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

We know that, standard deviation = $\sqrt{\text{variance}}$

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

Or $\quad \frac{1}{n}\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

Working rule to find variance and standard deviation:-

To find the standard deviation or variance when deviations are taken from the actual mean, we use the following working steps.

Step – 1, Calculate the mean $\bar{x}$ of the given observation $x_1, x_2, ...x_n$.

Step – 2, Square the deviations obtained in step II and then find the sum i.e $\sum_{i=1}^{n}(x_i - \bar{x})^2$

Step – 3, Find the variance and standard deviation by using the formula, variance, $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ and standard deviation $= \sqrt{\sigma^2}$.

Example:- 1

Find the variance and standard deviation for the following data, 6, 7, 10, 12, 13, 4, 8, 12.

Solution:- Given observations are 6, 7, 10, 12, 13, 4, 8, 12

Number of observations = 8

$$\therefore \text{Mean}(\bar{x}) = \frac{6+7+10+12+13+4+8+12}{8} = \frac{72}{8} = 9$$

Now, let us make the following table for deviation.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
6	−3	9	13	4	16
7	−2	4	4	−5	25
10	1	1	8	−1	1
12	3	9	12	3	9
Total		74	Total		74

∴ Sum of squares of deviations = $\sum_{i=1}^{n}(x_i-\bar{x})^2=74$

Hence, variance, $\sigma^2=\frac{\sum_{i=1}^{8}(x_i-\bar{x})^2}{n}=\frac{74}{8}=9.25$ and standard deviation = $\sqrt{\sigma^2}=\sqrt{9.25}=3.04$

Variance and Standard Deviation of a Discrete Frequency Distribution:-

Let the discrete frequency distribution be $x: x_1, x_2, x_3.......x_n$ and $f: f_1, f_2, f_3..........f_n$. Then by

Direct Method:-

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

Or $(\sigma^2)=\frac{1}{N}\sum f_i x_i^2-\left(\frac{\sum f_i x_i}{N}\right)^2$

And standard deviation, $\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{n}f_i(x_i-\bar{x})^2}$ Or $\sigma=\frac{1}{N}\sqrt{N\sum f_i x_i^2-\left(\sum f_i x_i\right)^2}$ Where $N=\sum_{i=1}^{n}f_i$

Shortcut Method:-

Variance, $\sigma^2=\frac{1}{N}\sum_{i=1}^{n}f_i d_i^2-\left(\frac{\sum f_i d_i}{N}\right)^2$ and standard deviation, $\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{n}f_i d_i^2-\left(\frac{\sum f_i d_i}{N}\right)^2}$

Where, $d_i=x_i-a$, deviation from assumed mean and a = assumed mean.

Example:- 2

Find the variance and standard deviation of the following data.

x_i	4	8	11	17	20	24	32
f_i	3	5	9	5	4	3	1

Solution:-

Let us make the following table from the given data

x_i	f_i	$f_i x_i$	$x_i-\bar{x}$ $=x_i-14$	$(x_i-\bar{x})^2$	$f_i(x_i-\bar{x})^2$
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
Total	30	420			1374

Here, we have, $N=\sum f_i=30, \sum f_i x_i=420$.

$$\therefore \quad \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N} = \frac{420}{30} = 14$$

Hence, the variance $\left(\sigma^2\right) = \frac{1}{N}\sum_{i=1}^{7} f_i\left(x_i - \bar{x}\right)^2 = \frac{1}{30} \times 1374 = 45.8$

And standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{45.8} = 6.77$

Variance and standard deviation of continuous frequency distribution:-

Direct method:- In this method, we first replace each class by its mid-point, then this method becomes similar to the discrete frequency distribution. If there is a frequency distribution of n classes and each class defined by its mid-point x_i with corresponding frequency f_i then variance and standard deviation are respectively.

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

Or $\quad \sigma^2 = \frac{1}{N^2}\left[N\sum f_i x_i^2 - \left(\sum f_i x_i\right)^2\right]$ and $\sigma = \frac{1}{N}\sqrt{N\sum f_i x_i^2 - \left(\sum f_i x_i\right)^2}$

Example:-3

Calculate the 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:- Let us construct the following table.

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	265	529	1587
90 – 100	2	95	190	1089	2178
Total	50		3100		10050

Here, $N = 50$ and $\sum f_i x_i = 3100$

$\therefore$ Mean, $\bar{x} = \frac{1}{N}\sum f_i x_i = 62$

Now, variance $\sigma^2 = \frac{1}{N}\sum f_i\left(x_i - \bar{x}\right)^2 = \frac{1}{50} \times 10050 = 201$ and standard deviation, $\sigma = \sqrt{201} = 14.18$

Shortcut Method or step-deviation method:-

Sometimes the values of 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. For this, we use the step deviation method to find mean and variance.

Variance, $\sigma^2 = h^2\left[\frac{1}{N}\sum_{i=i}^{n} f_i u_i^2 - \left(\frac{\sum_{i=1}^{n} f_i u_i}{N}\right)\right]$ and standard deviation, $\sigma = h\sqrt{\frac{1}{N}\sum_{i=1}^{n} f_i u_i^2 - \left(\frac{\sum_{i=1}^{n} f_i u_i}{N}\right)^2}$

Where $u_i = \frac{x_i - a}{h}$, $a =$ assumed mean and h = width of the class interval.

Working rule to find variance and standard deviation by shortcut method:-

To find variance and standard deviation, use the following steps.

Step – 1, Select an assumed mean, say a and then calculate $u_i = \frac{x_i - a}{h}$, where h = width of class interval or common factor.

Step – 2, Multiply the frequency of each class with the corresponding u_i and obtain $\sum f_i u_i$.

Step – 3, Square the values of u_i and multiply them with the corresponding frequencies and obtain $\sum f_i u_i^2$.

Step – 4, Substitute the values of $\sum f_i u_i$, $\sum f_i u_i^2$ and $\sum f_i = N$ in the formula,

Variance $(\sigma^2) = h^2\left[\frac{1}{N}\sum f_i u_i^2 - \left(\frac{1}{N}\sum f_i u_i\right)^2\right]$ and standard deviation = $\sqrt{\sigma}$ to get the required values.

Example:- 4

Calculate the mean and standard deviation of the following cumulative data.

Wages (in Rs.)	0 – 15	15 – 30	30 – 45	45 – 60	60 – 75	75 – 90	90 – 105	105 – 120
Number of workers	12	30	65	107	157	202	222	230

Solution:- We are given the cumulative frequency distribution. So, first, we will prepare the frequency distribution is given below.

Class interval	cf	Mid value(x_i)	f_i	$u_i = \frac{x_i - 67.5}{15}$	$f_i u_i$	$f_i u_i^2$
0 – 15	12	7.5	12	–4	–48	192
15 – 30	30	22.5	18	–3	–54	162
30 – 45	65	37.5	35	–2	–70	140
45 – 60	107	52.5	42	–1	–42	42
60 – 75	157	67.5	50	0	0	0
75 – 90	202	82.5	45	1	45	45
90 – 105	222	97.5	20	2	40	80
105 – 120	230	112.5	8	3	24	72
Total			230		–105	733

Here, a = 67.5, h = 15, $N = \sum f_i = 230$, $\sum f_i u_i = -105$ and $\sum f_i u_1^2 = 733$

Now, Mean = $a + h\left(\frac{1}{N}\sum f_i u_i\right) = 67.5 + 15\left(\frac{-105}{230}\right) = 67.5 - 6.85 = 60.65$

Standard deviation, $(\sigma) = \sqrt{h^2\left[\frac{1}{N}\sum f_i u_1^2 - \left(\frac{1}{N}\sum f_i u_i\right)^2\right]}$

$= \sqrt{225\left[\frac{733}{230} - \left(-\frac{105}{230}\right)^2\right]} = \sqrt{225[3.187] - (0.46)^2}$

$= \sqrt{225(3.1870 - 0.2116)} = \sqrt{669.465} = 25.87$

Analysis of frequency Distribution:- We have seen that the mean deviation and standard deviation have the same unite in which the data is given. The measures of dispersion are unable to compare the variability of two or more series which are measured in different units. So, we require those measures which are independent of the units. The measures of variability which is independent of units are called the coefficient of variation denoted as CV and it is given by $CV = \frac{\sigma}{\overline{x}} \times 100$

Where $\overline{x}$ and σ are respectively the mean and the standard deviation of the data. For comparing the variability of two series, we calculate the coefficient of variations for each series.

Comparison of two frequency distributions with the same mean:-

Let us consider two frequency distributions with standard deviations σ_1 and σ_2 and having the same mean $\overline{x}$, then

CV (1st distribution) = $\frac{\sigma_1}{\overline{x}} \times 100$ and CV (2nd distribution) = $\frac{\sigma_2}{\overline{x}} \times 100$

$$\therefore \quad \frac{CV(1\text{st distribution})}{CV(2\text{nd distribution})} = \frac{\frac{\sigma_1}{\bar{x}} \times 100}{\frac{\sigma_2}{\bar{x}} \times 100} = \frac{\sigma_1}{\sigma_2}$$

Two CVs and be compared based on values σ_1 and σ_2. Thus, if two series have equal means, then the series with greater standard deviation (or variance) is said to be more variable (or dispersed) then the other. Also, the series with the lesser value of the standard deviation (or variance) is said to be more consistent than the other.

Example:-1

Two plants A and B of a factory show the following results about the number of workers and the wages paid to them.

	A	B
Number of workers	4000	4500
Average monthly wages	Rs.3000	Rs.3000
Variance of distribution	16	25

Which plant, A or B shows greater variability in individual wages?

Solution:- Here, we observe that average monthly wages in both the plants are the same i.e Rs. 3000. Therefore, the plant with a greater variance will have more variability. Hence, plant B has greater variability in individual wages.

Example:-2

Goals scored by two terms A and B in a football session were as follows.

Number of goals scored in the match	**Number of Matches**	
	Team A	**Team B**
0	24	25
1	9	9
2	8	6
4	5	5
	4	5

Which team is more consistent?

Solution:- For Team a let assumed mean a = 2

Now, let us make the following table from the given data

x_i	f_i	$d_i = x_i - 2$	d_1^2	f_1d_1	$f_1d_1^2$
0	24	−2	4	−48	96
1	9	−1	1	−9	9
2	8	0	0	0	0
3	5	1	1	5	5
4	4	2	4	8	16
Total	50			−44	126

Here, $\sum f_i = 50, \sum f_i d_i = -44$ and $\sum f_i d_i^2 = 126$

$\therefore$ Mean = $a + \dfrac{\sum f_i d_i}{\sum f_i} = 2 - \dfrac{44}{50} = 2 - 0.88 = 1.12$ and standard deviation,

$$\sigma_A = \sqrt{\frac{1}{N}\sum f_i d_i^2 - \left(\frac{1}{N}\sum f_i d_i\right)^2} = \sqrt{\frac{126}{50} - \left(\frac{-44}{50}\right)^2}$$

$$= \sqrt{252 - 0.7744} = \sqrt{1.7456} = 1.32$$

For Team B let assumed mean a = 2, Now let us make the following table from the given data.

x_i	f_i	$d_i = x_i - 2$	d_i^2	$f_i d_i$	$f_i d_1^2$
0	25	−2	4	−50	100
1	9	−1	1	−9	9
2	6	0	0	0	0
3	5	1	1	5	5
4	5	2	4	10	20
Total	50			−44	134

Here, $\sum f_i d_i = -44, \sum f_i = 50$ and $\sum f_i d_1^2 = 134$

$\therefore$ Mean $= a + \dfrac{\sum f_i d_i}{\sum f_i} = 2 - \dfrac{44}{50} = 2 - 0.88 = 134$

Standard deviation, $\sigma_B = \sqrt{\frac{1}{N}\sum f_i d_i^2 - \left(\frac{1}{N}\sum f_i d_i\right)^2} = \sqrt{\frac{134}{50} - \left(\frac{-44}{50}\right)^2}$

$$= \sqrt{2.68 - 0.7744} = \sqrt{1.9056} = 1.38$$

Here, the means of both teams are equal but the standard deviation of team A is less than the standard deviation of team B. Hence, team A is more consistent.

Example:-3

If each of the observations $x_1, x_2, \ldots\ldots\ldots x_n$ is increased by a, where a is a negative or positive number, then show that the variance remains unchanged.

Solution:- Let $\bar{x}$ be the mean of $x_1, x_2, \ldots\ldots\ldots\ldots\ldots x_n$

Then, the variance is given by $\sigma_1^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ .... (1)

If a is added to each observation, then the new observation will be $y_i = x_i + a$ .... (2)

Let the mean of the new observation be $\overline{y}$. Then,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + a) = \frac{1}{n}\left[\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} a\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i + na\right] = \overline{x} + a \text{ i.e } \overline{y} = \overline{x} + a \text{ .... (3)}$$

Now, new variance $\sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i + a - \overline{x} - a)^2$ (using equation (2) and (3)

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sigma_1^2$$

Hence, the variance remains unchanged.