LESSON 14

STATISTICS

Statistics is the science of collecting, organizing and interpreting numerical facts which we often call **data**. Synonyms for data are scores, measurements and observations. The study and collection of data involves classifying data in various heads. The process involves lot of representations of a characteristic by numbers and it is termed as **measurement**. In other words **Data** are measurements of a situation under consideration.

Example: The measurements of heights of all creatures in world is a data. All numerical characteristics are called **variables**. A large number of observations on a single variable can be summarized in a table of frequencies. Any particular pattern of variation is termed as distribution.

1. MEASURES OF CENTRAL TENDENCY

The most commonly used measures of central tendency are

- The mode
- The median
- The arithmetic mean

1.1 THE MODE

The mode is the most frequently occurring value in a distribution of a variable.

1.2 THE MEDIAN

The median is the middlemost point in a rank ordered set of measures.

If the number of observations is odd, then the median is $\frac{n\Gamma 1}{2}$ th observation. If the number of observations is even, then median is the mean of $\frac{n}{2}$ th and $\frac{n}{2}\Gamma 1$ th observations.

1.3 THE ARITHMETIC MEAN

The arithmetic mean is defined as the sum of the values divided by the total number of values. $X \text{ X } \frac{X_i}{N}$. The mean is not necessarily the middle of the distribution, as is the median.

The mean is the point in a distribution about which deviations from it sum to zero. A deviation score (*x*) is defined as the distance between a value and its mean. They can be either positive (+) or negative (-).In this sense the mean is a *centroid*.

It is observed that the measures of central tendency are not sufficient to give complete information about the given data. "Variability" is another factor which is required to be studied under statistics. The single number which describes variability, is known as 'Measures of dispersion'.

2. MEASURES OF DISPERSION

Dispersion means **'Scatteredness'**. Dispersion measures the degree of scatteredness of the variable about a central value.

There are following measures of dispersion:

- Range
- Quartile deviation
- Mean deviation
- Variance
- Standard deviation.

In this chapter, we shall study all of these measures of dispersion, except the Quartile deviation.

RANGE

The range is the simplest measure of variation to find. The usual definition of range is the difference between the maximum and minimum values of a population.

RANGE = MAXIMUM VALUE – MINIMUM VALUE

For example, consider the following series

60 60 60 60 60 60 60 60 60 60 60 Range = 0

0 2 3 15 20 60 89 91 95 99 100 Range = 100

0 49 50 51 54 60 74 75 76 78 100 Range = 100

- Since the range only uses the largest and smallest values, it is greatly affected by extreme values.
- The range of data gives us a rough idea of variability or scatter.

3. MEAN DEVIATION

Mean deviation of a distribution is the arithmetic mean of the absolute deviations of the terms of the distribution from its statistical mean (arithmetic mean, median or mode).

- Mean deviation may be obtained from any measure of central tendency. However, mean deviation from mean and median are commonly used in statistical studies.
- Mean deviation about the median is least.

3.1 MEAN DEVIATION FOR UNGROUPED DATA

Let $x_1, x_2, x_3, ..., x_n$ are n values of a variable X and k be the statistical mean (A.M., median, mode) about which we have to find the mean deviation. The mean deviation (M.D.) about k is given by

$$\text{M.D.}(k) = \frac{|x_1 - k| + |x_2 - k| + |x_3 - k| + ... + |x_n - k|}{n} = \frac{\sum_{i=1}^{n} |x_i - k|}{n}$$

***Illustration* 1**

Question: **Find the mean deviation about the mean for the following data:**

12, 3, 18, 17, 4, 9, 17, 19, 20, 15, 8, 17, 2, 3, 16, 11, 3, 1, 0, 5

Solution: We have, $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{200}{20} = 10$

The respective absolute values of the deviations from mean, i.e., $|x_i - \bar{x}|$ are

2, 7, 8, 7, 6, 1, 7, 9, 10, 5, 2, 7, 8, 7, 6, 1, 7, 9, 10, 5

Therefore $\sum_{i=1}^{20} |x_i - \bar{x}| = 124$ and $M.D. = \frac{124}{20} = 6.2$

***Illustration* 2**

Question: **Following are the marks obtained by 9 students in an examination, find their mean deviation from median.**

49, 68, 21, 32, 54, 38, 41, 66, 59

Solution: Arranging the observations in ascending order of magnitude, we have,

21, 32, 38, 41, 49, 54, 59, 66, 68

Number of observations = 9

Median = 5^{th} term = 49

Calculation of mean deviation:

x_i	$\lvert d_i \rvert = \lvert x_i - 49 \rvert$
21	28
32	17
38	11
41	8
49	0
54	5
59	10
66	17
68	19
Total	115

$\therefore \quad M.D. = \frac{1}{n} \lvert d_i \rvert = \frac{115}{9} = 12.78$

3.2 MEAN DEVIATION FOR GROUPED DATA

(a) Discrete Frequency Distribution:

Let $x_1, x_2, x_3, ..., x_n$ be n observations occurring with frequencies $f_1, f_2, f_3, ..., f_n$ respectively and k be the statistical mean (A.M., median, mode). The mean deviation (M.D.) about k is given by

$$\text{M.D.}(k) = \frac{\lvert x_1 - k \rvert f_1 + \lvert x_2 - k \rvert f_2 + \lvert x_3 - k \rvert f_3 + ... + \lvert x_n - k \rvert f_n}{f_1 + f_2 + f_3 + ... + f_n} = \frac{\sum_{i=1}^{n} \lvert x_i - k \rvert f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} \lvert d_i \rvert f_i}{N},$$

Where $d_i = x_i - k$ and $N = \sum_{i=1}^{n} f_i$ = total frequency

- The mean of given discrete frequency distribution is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_nx_n}{f_1 + f_2 + f_3 + \ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

- To find the median of given discrete frequency distribution, observations are arranged in ascending order. After this, cumulative frequencies are obtained, then the observation is identified whose cumulative frequency is equal to or just greater than $\frac{N}{2}$, this value of the observation lies in the middle of the data, it is the required median.

***Illustration* 4**

Question: **Find the mean deviation from the median of the following frequency distribution:**

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

Solution: Calculation of mean deviation from the median

Age (in years) x_i	Frequency f_i	Cumulative frequency	f_ix_i	$\lvert x_i - 18 \rvert$	$f_i \lvert x_i - 18 \rvert$
10	3	3	30	8	24
12	5	8	60	6	30
15	4	12	60	3	12
18	10	22	180	0	0
21	8	30	168	3	24
23	4	34	92	5	20
Total	$N = 34$				110

Here, $N = 34$, $\frac{N}{2} = 17$ and $\frac{N}{2} + 1 = 18$

$\therefore$ Median = $\frac{\text{value of } 17^{th} \text{ term} + \text{value of } 18^{th} \text{ term}}{2} = \frac{18 + 18}{2} = 18$ years

$\therefore$ Mean deviation from the median = $\frac{\sum f_i \lvert x_i - 18 \rvert}{N} = \frac{110}{34} = 3.23$ years

Illustration 5

Question: **The mean of 4, 7, 2, 8, 6 and *a* is 7. Find the mean deviation about median of these observations.**

Solution: Here number of observations, $n = 6$

Given $\frac{4 + 7 + 2 + 8 + 6 + a}{6} \times 7$ or $27 + a \times 42$ or $a = 15$

… Arranging the observations in ascending order, we get 2, 4, 6, 7, 8, 15

… Median, k = mean of $\frac{n}{2}$ th observation and $\frac{n}{2} + 1$ th observation

$= \frac{3\text{rd observations} + 4\text{th observation}}{2} \times \frac{6 + 7}{2} \times 6.5$

Calculation of mean:

x_i	$\lvert x_i - k \rvert$
2	4.5
4	2.5
6	0.5
7	0.5
8	1.5
15	8.5
Total	18

… Mean deviation about median = $\frac{\lvert x_i - k \rvert}{n} \times \frac{18}{6} \times 3$

Illustration 6

Question: **Find the mean deviation about the mean for the following data:**

x_i	1	4	9	12	13	14	21	22
f_i	3	4	5	2	4	5	4	3

Solution: For computing mean and then deviation about mean, we construct the following table:

x_i	f_i	$x_i f_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
1	3	3	11	33
4	4	16	8	32
9	5	45	3	15
12	2	24	0	0
13	4	52	1	4
14	5	70	2	10
21	4	84	9	36

22	3	66	10	30
Total	30	360		160

Mean $\bar{x} = \frac{f_i x_i}{f_i} = \frac{360}{30} = 12$

M.D. (about mean) = $\frac{f_i |x_i - \bar{x}|}{f_i} = \frac{160}{30} = 5.33$

(b) Continuous Frequency Distribution:

The mean of a continuous frequency distribution is calculated with the assumption that the frequency in each class is centered at its mid-point.

Let x_i be the mid-value of the i^{th} class, f_i be the frequency of the i^{th} class and k be the statistical mean (A.M., median, mode), then the mean deviation (M.D.) about k is given by

M.D.(k) = $\frac{\sum_{i=1}^{n} |x_i - k| f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} |d_i| f_i}{N}$, Where $d_i = x_i - k$ and $N = \sum_{i=1}^{n} f_i$ = total frequency

SHORTCUT METHOD FOR CALCULATING MEAN DEVIATION ABOUT MEAN

For the given continuous frequency distribution, arithmetic mean can be calculated by shortcut (step-deviation)method. Rest of the procedure is same.

In this method,

(i) Take an assumed mean (just middle data or close to it).

(ii) Calculate deviations of the observations (or mid-point of classes) from the assumed mean.

(iii) If there is a common factor of all the deviations, divide them by this common factor to simplify the deviations.

(iv) Now the arithmetic mean $\bar{x}$ by step-deviation method is given by

$\bar{x} = a + \frac{\sum_{i=1}^{n} f_i d_i}{N} \times h$, where $d_i = \frac{x_i - a}{h}$

a = assumed mean, h = common factor and $N = \sum_{i=1}^{n} f_i$

TO CALCUATE THE MEDIAN FOR A CONTINUOUS FREQUENCY DISTRIBUTION

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

(ii) The class corresponding to cumulative frequency just more than $\frac{N}{2}$ is known as median class.

(iii) Median = $l + \frac{h}{f}\left(\frac{N}{2} - c\right)$,

where l = lower limit of median class

f = frequency of the median class

h = width of the median class

c = cumulative frequency of the class just preceding the median class

***Illustration* 7**

Question: **Find the mean deviation from the mean for the data:**

Classes	**0-100**	**100-200**	**200-300**	**300-400**	**400-500**	**500-600**	**600-700**	**700-800**
Frequency	**4**	**8**	**9**	**10**	**7**	**5**	**4**	**3**

Solution: We construct the following table:

Classes	x_i	$d_i = \frac{x_i - 350}{100}$	f_i	$f_i d_i$	$\lvert x_i - k \rvert$	$f_i \lvert x_i - k \rvert$
0-100	50	–3	4	–12	308	1232
100-200	150	–2	8	–16	208	1664
200-300	250	–1	9	–9	108	972
300-400	350	0	10	0	8	80
400-500	450	1	7	7	92	644
500-600	550	2	5	10	192	960
600-700	650	3	4	12	292	1168
700-800	750	4	3	12	392	1176
Total			$N = f_i = 50$	$f_i d_i = 4$		7896

Actual mean,

$$k = A - \frac{f_i d_i}{f_i} \times h, \qquad \text{where } A = \text{assumed mean and } h = \text{class interval}$$

$$\therefore \quad k = 350 - \frac{4}{50} \times 100 = 358$$

Now, mean deviation,

$$\text{M.D. } (k) = \frac{f_i \lvert x_i - k \rvert}{f_i} = \frac{7896}{50} = 157.92$$

***Illustration* 8**

Question: **Find the mean deviation of the following distribution from the median.**

Classes	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**
Frequencies	**10**	**12**	**8**	**16**	**14**	**10**

Solution: We have

Classes	x_i	f_i	Cumulative frequency	$\lvert x_i - 43.125 \rvert$	$f_i \lvert x_i - 43.125 \rvert$
10-20	15	10	10	28.125	281.250
20-30	25	12	22	18.125	217.500
30-40	35	8	30	8.125	65.000
40-50	45	16	46	1.875	300.000
50-60	55	14	60	11.875	166.250
60-70	65	10	70	21.875	218.750
Total		$N = 70$			1248.750

Here $N = f_i = 70$, $\frac{N}{2} = 35$

The cumulative frequency just greater than 35 is 46 and the corresponding class is 40-50. So, 40-50 is the median class.

Now, median $M = l + \frac{\frac{N}{2} - C}{f} \times h$, here $l = 40$, $N = 70$, $C = 30$, $h = 10$, $f = 16$

$\therefore \quad M = 40 + \frac{35 - 30}{16} \times 10 = 43.125$

Mean deviation from median = $\frac{f_i \lvert x_i - 43.125 \rvert}{N} = \frac{1248.750}{70} = 17.83$

3.3 Limitations of mean deviation

(i) The sum of the absolute deviations about the mean is greater than the sum of the absolute deviations from median, in fact mean deviation about median is least. Therefore mean deviation about mean is not very suitable.

(ii) In the series, where the degree of variability is very high, the median is not a representative of central tendency. The mean deviation about the mean is not a very good measure of dispersion.

(iii) Mean deviation is calculated on the basis of absolute values of the deviations and therefore can not be subjected to further algebraic treatment.

4. VARIANCE AND STANDARD DEVIATION

While calculating the mean deviation, the absolute values of the deviations were taken to avoid the difficulty which arose due to the signs of deviation. The another way is to take squares of all the deviations.

4.1 VARIANCE

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

Therefore, if $x_1, x_2, x_3, \ldots\ldots, x_n$ be n given values of a variate and $\bar{x}$ be their mean, then

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

4.2 STANDARD DEVIATION

In the calculation of variance, the units of individual observations x_i and the unit of their mean $(\bar{x})$ are different from that of variance. The proper measure of dispersion about the mean of a set of observations is expressed as positive square root of variance and is known as standard deviation (σ).

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

4.3 VARIANCE AND STANDARD DEVIATION IN DIFFERENT CASES

(a) In case of individual series (ungrouped data):

Let $x_1, x_2, x_3, \ldots\ldots x_n$ are n values of a variable x, then by definition

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

(where $\bar{x}$ is A.M. of $x_1, x_2, x_3, \ldots\ldots x_n$ i.e., $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$)

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

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

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

(b) In case of discrete frequency distribution:

Let $x_1, x_2, x_3, \ldots\ldots x_n$ be n observations having frequency $f_1, f_2, f_3, \ldots\ldots, f_n$ respectively, then

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

(where $\bar{x}$ is A.M. of $x_1, x_2, x_3, ...x_n$ i.e., $\bar{x} = \frac{x_i f_i}{N}$ and $N = \sum_{i=1}^{n} f_i$)

$$= \frac{1}{N}\left[\sum x_i^2 - 2\bar{x}x_i + \bar{x}^2\right]f_i = \frac{x_i^2 f_i}{N} - 2\bar{x}\frac{x_i f_i}{N} + \bar{x}^2\frac{f_i}{N}$$

$$= \frac{x_i^2 f_i}{N} - 2\bar{x}\,\bar{x} + \bar{x}^2 = \frac{x_i^2 f_i}{N} - \bar{x}^2$$

Hence standard deviation, $\sigma = \sqrt{\frac{x_i^2 f_i}{N} - \bar{x}^2}$

(c) In case of continuous frequency distribution:

Let x_i = mid-value of i th class

f_i = frequency of i th class

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

$\bar{x}$ = A.M. of given observations

then variance, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}\left[x_i - \bar{x}\right]^2 f_i = \frac{x_i^2 f_i}{N} - \bar{x}^2$

and standard deviation, $\sigma = \sqrt{\frac{x_i^2 f_i}{N} - \bar{x}^2}$

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

Illustration 9

Question: **Calculate the standard deviation of the first *n* natural numbers.**

Solution: Here, $x_i = i$, where $i = 1, 2, ..., n$

Now, $\sigma = \sqrt{\frac{x_i^2}{n} - \bar{x}^2}$

Mean, $\bar{x} = \frac{x_i}{n} = \frac{1 + 2 + 3 + ... + n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$

and $x_i^2 = 1^2 + 2^2 + 3^2 + ... + n^2 = \frac{n(n+1)(2n+1)}{6}$

$\therefore \sigma = \sqrt{\frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2} = \sqrt{\frac{(n+1)}{2}\left(\frac{2n+1}{3} - \frac{n+1}{2}\right)} = \sqrt{\frac{n+1}{2}\cdot\frac{4n+2-3n-3}{6}}$

$$= \sqrt{\frac{n(n-1)}{12}} \times \sqrt{\frac{n^2-1}{12}}$$

***Illustration* 10**

Question: **If the following are the heights in centimeters of 8 students, find their arithmetic mean and standard deviation.**

162, 163, 160, 164, 160, 170, 161, 164

Solution: We have,

Mean, $\bar{x} = \frac{162 + 163 + 160 + 164 + 160 + 170 + 161 + 164}{8} = 163$

Height (in cm) x_i	$x_i - \bar{x}$	$f(x_i - \bar{x})^2$
160	–3	9
160	–3	9
161	–2	4
162	–1	1
163	0	0
164	1	1
164	1	1
170	7	49
Total		74

Here $n = 8$ and $f(x_i - \bar{x})^2 = 74$

… $\sigma = \sqrt{\frac{1}{n} f(x_i - \bar{x})^2} = \sqrt{\frac{74}{8}} = \sqrt{9.25} = 3.04$ cm

Thus, mean = 163 and standard deviation = 3.04 cm

***Illustration* 11**

Question: **Find the variance and standard deviation of the following frequency distribution:**

x_i	6	10	14	18	24	28	30
f_i	2	4	7	12	8	4	3

Solution: Calculation of variance and standard deviation:

x_i	f_i	$f_i x_i$	$x_i - 19$	$(x_i - 19)^2$	$f_i (x_i - 19)^2$
6	2	12	–13	169	338
10	4	40	–9	81	324
14	7	98	–5	25	175
18	12	216	–1	1	12
24	8	192	5	25	200
28	4	112	9	81	324
30	3	90	11	121	363

	$N = \sum f_i = 40$	$f_i x_i = 760$			$f_i (x_i - \bar{x})^2 = 1736$

Here, $N = 40$, $f_i x_i = 760$

… Mean, $\bar{x} = \frac{1}{N}. \sum f_i x_i = \frac{760}{40} = 19$. We have, $f_i (x_i - \bar{x})^2 = 1736$

… Variance, $\sigma^2 = \frac{1}{N}. \sum f_i (x_i - \bar{x})^2 = \frac{1736}{40} = 43.4$

and standard deviation = $\sigma = \sqrt{43.4} = 6.59$

***Illustration* 12**

Question: **Calculate the mean, variance and standard deviation for the following distribution:**

Class	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90-100**
Frequency	**3**	**7**	**12**	**15**	**8**	**3**	**2**

Solution: Calculation of mean, variance and standard deviation:

Classes	Mid-values x_i	f_i	$f_i x_i$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
30-40	35	3	105	729	2187
40-50	45	7	315	289	2023
50-60	55	12	660	49	588
60-70	65	15	975	9	135
70-80	75	8	600	169	1352
80-90	85	3	255	529	1587
90-100	95	2	190	1089	2178
		$\sum f_i = N = 50$	$\sum f_i x_i = 3100$		$\sum f_i (x_i - \bar{x})^2 = 10050$

Mean, $\bar{x} = \frac{\sum f_i x_i}{N} = \frac{3100}{50} = 62$

Hence, variance $\sigma^2 = \frac{\sum f_i (x_i - \bar{x})^2}{N} = \frac{10050}{50} = 201$

and standard deviation, $\sigma = \sqrt{201} = 14.17$

4.4 SHORTCUT METHOD TO FIND VARIANCE AND STANDARD DEVIATION: (When x_i are large)

Let the assumed mean be A and the width of class interval be h.

Let $u_i = \frac{x_i - A}{h}$ $\quad x_i = A + hu_i$ ..... (1)

The arithmetic mean $\bar{x} = \frac{f_i x_i}{N}$

$$= \frac{f_i (A + hu_i)}{N}$$

$$\bar{x} = A + h\frac{f_i u_i}{N} \quad \text{..... (2)}$$

from (1) and (2),

$$x_i - \bar{x} = h\ u_i - \frac{f_i u_i}{N} \quad \text{..... (3)}$$

Now, variance(σ^2) = $\frac{f_i (x_i - \bar{x})^2}{N} = \frac{h^2}{N}$ $u_i - \frac{f_i u_i}{N}^2 f_i$ {using (3)}

$$= \frac{h^2}{N} \quad (u_i - \bar{u})^2 f_i = h^2 \mid \text{(variance of variable } u_i)$$

$\sigma_x^2 = h^2 \sigma_u^2$ $\quad \sigma_x = h\sigma_u$ .....(4)

$\sigma_x = \frac{h}{N}\sqrt{N \quad f_i u_i^2 - (f_i u_i)^2}$ as $\sigma = \frac{1}{N}\sqrt{N \quad f_i x_i^2 - (f_i x_i)^2}$

***Illustration* 13**

Question: **Calculate the mean, variance and standard deviation for the following distribution:**

Class	30-40	40-50	50-60	60-70	70-80	80-90	90-100
Frequency	3	7	12	15	8	3	2

Solution: Let the assumed mean $A = 65$.

Here $h = 10$

We obtain the following table:

Class	Frequency f_i	Mid-point x_i	$y_i = \frac{x_i - 65}{10}$	y_i^2	$f_i y_i$	$f_i y_i^2$
30-40	3	35	–3	9	–9	27
40-50	7	45	–2	4	–14	28
50-60	12	55	–1	1	–12	12
60-70	15	65	0	0	0	0
70-80	8	75	1	1	8	8
80-90	3	85	2	4	6	12
90-100	2	95	3	9	6	18
	$N = 50$				–15	105

Therefore $\bar{x} = A + \frac{f_i y_i}{50} \times h = 65 - \frac{15}{50} \times 10 = 62$

Variance, $\sigma^2 = \frac{h^2}{N^2}\left[N \sum f_i y_i^2 - \left(\sum f_i y_i\right)^2\right]$

$= \frac{100}{2500}\left[50 \times 105 - (-15)^2\right] = 201$

and standard deviation $\sigma = \sqrt{201} = 14.18$

***Illustration* 14**

Question: **Find the mean and standard deviation for the following data using short-cut method.**

x_i	60	61	62	63	64	65	66	67	68
f_i	2	1	12	29	25	12	10	4	5

Solution: Calculation of variance and standard deviation:

x_i	f_i	$d_i = x_i - 64$	d_i^2	$f_i d_i$	$f_i d_i^2$
60	2	–4	16	–8	32
61	1	–3	9	–3	9
62	12	–2	4	–24	48
63	29	–1	1	–29	29
64	25	0	0	0	0
65	12	1	1	12	12
66	10	2	4	20	40
67	4	3	9	12	36
68	5	4	16	20	80
	$f_i = N = 100$			$f_i d_i = 0$	$f_i d_i^2 = 286$

Here, assumed mean = 64

Actual mean, $\bar{x} = a + \frac{1}{N} f_i d_i = 64 + \frac{0}{100} = 64$

Standard deviation, $\sigma = \sqrt{\frac{1}{N} f_i d_i^2 - \left(\frac{f_i d_i}{N}\right)^2} = \sqrt{\frac{286}{100} - 0} = \sqrt{2.86} = 1.69$

5. ANALYSIS OF FREQUENCY DISTRIBUTION

In order to compare the variability of two series with same mean, which are measured in different units, merely calculating the measures of dispersion are not sufficient, but we require such measures which are independent of the units. The measure of variability which is independent of units is called coefficient of variation(C.V.) and defined as

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100, \quad \bar{x} \neq 0$$

Where σ and $\bar{x}$ are standard deviation and mean of data respectively.

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

) COMPARISION OF TWO FREQUENCY DISTRIBUTION WITH SAME MEAN

Let σ_1 and σ_2 be the standard deviation of two series with mean $\bar{x}$, then

C.V. (I^{st} distribution) = $\frac{\exists_1}{\bar{x}} \mid 100, \ \bar{x} \mid 0$

C.V. (II^{nd} distribution) = $\frac{\exists_2}{\bar{x}} \mid 100, \ \bar{x} \mid 0$

Hence, the above two C.V. can be compared on the basis of values of $\exists_1$ and $\exists_2$ only.

Therefore, for the two series with equal means, the series with greater standard deviation (or variance) is called more variable or dispersed than the other, Also the series with lesser value of standard deviation (or variance) is said to be more consistent than the other.

***Illustration* 15**

Question: **The ODI performance of two cricket players of a cricket Team is as follows:**

Player	**Runs in last 10 ODI matches**									
Rahul	**27**	**45**	**31**	**46**	**23**	**87**	**101**	**78**	**24**	**11**
Sachin	**43**	**95**	**5**	**78**	**88**	**103**	**23**	**01**	**41**	**52**

Who is more reliable among these two?

Solution: Clearly it is observed that Sachin had made much more runs than Rahul i.e. 529 vs 473 and hence has a better average i.e. 52.9 vs 47.3. But in order to test the reliability of the two sets of data we need to calculate the Standard Deviation

S.D. of Rahul: s = 30.8

S.D. of Sachin : s = 36.9

C.V For Sachin: = 0.698

C.V. for Rahul: = 0.651

Decision: Since the C.V. of Rahul is less, he is more reliable than Sachin.

PRACTICE PROBLEMS

PP1. Find the mean deviation about median for the following data:

x_i	3	6	9	12	13	15	21	22
f_i	3	4	5	2	4	5	4	3

PP2. 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

PP3. Calculate the mean deviation about median for the following data:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

PP4. Find the variance and standard deviation of the following data:
5, 9, 10, 12, 8, 13, 6

PP5. The following table gives the frequency distribution of marks obtained by a batch of 20 students. Find their mean and standard deviation.

Marks obtained	5	15	25	35	45
Numbers of students	5	4	6	3	2

PP6. If the standard deviation of a set of observations is 4 and if each observation is divided by 4, find the standard deviation of the new set of observations.

PP7. Show that if standard deviation of a set of observations is 8 and if each observation is divided by –2, the standard deviation of the new set of observations will be 4.

PP8. Show that in a discrete series, the relation between standard deviation s and range *r* is $s = r$.

PP9. The mean square deviation of a set of *n* observations $x_1, x_2, \ldots\ldots, x_n$ 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 observation.

PP10. The mean and standard deviation of *n* observations were calculated as *a* and *b*, respectively by a student who took by mistake *c* instead of *d* for one observation. Write a formula for the correct mean and standard deviation?

PP11. Given that *x* is the mean and x^2 is the variance of n observations. If every observation is multiplied by a non zero number *k,* then prove that the mean and variance becomes *kx* and k^2x^2.

PP12. Calculate the mean and standard deviation for the following data:

Wages upto (in Rs.)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of workers	9	17	32	23	40	18	1

PP13. The mean and standard deviation of 100 observations were calculated as 40 and 5.1, respectively by a student who took by mistake 50 instead of 40 for one observation. What are the correct mean and standard deviation?

SOLVED SUBJECTIVE EXAMPLES

***Example* 1:**

If $\bar{x}$ is the mean and mean deviation from mean is M.D $(\bar{x})$, then find the number of observations lying between $\bar{x} - M.D.(\bar{x})$ and $\bar{x} + M.D.(\bar{x})$ for the following data:

34, 66, 30, 38, 44, 50, 40, 60, 42, 51.

Solution:

Given data arranged in ascending order is 30, 34, 38, 40, 42, 44, 50, 51, 60, 66.

$$\bar{x} = \frac{30 + 34 + 38 + 40 + 42 + 44 + 50 + 51 + 60 + 66}{10} = \frac{455}{10} = 45.5$$

Now, $|x_i - \bar{x}| = 15.5,\ 11.5,\ 7.5,\ 5.5,\ 3.5,\ 1.5,\ 4.5,\ 5.5,\ 14.5,\ 20.5$

Mean deviation from the mean = $M.D.(\bar{x}) = \dfrac{\sum |(x_i - \bar{x})|}{10}$

$$= \left(\frac{1}{10}\right)[15.5 + 11.5 + 7.5 + 5.5 + 3.5 + 1.5 + 4.5 + 5.5 + 14.5 + 20.5] = \frac{1}{10}[90.0] = 9$$

Now, $\bar{x} - M.D.(\bar{x}) = 45.5 - 9 = 36.5$

and $\bar{x} + M.D.(\bar{x}) = 45.5 + 9 = 54.5$

Given observations which lie between 36.5 and 54.5 are

38, 40, 42, 44, 50, 51, which are six in number.

$\therefore$ six entries of given data lie between $\bar{x} - M.D.(\bar{x})$ and $\bar{x} + M.D.(\bar{x})$

***Example* 2:**

The mean and standard deviation of 20 observations are found to be 10 and 2 respectively. On rechecking, it was found that an observations was incorrectly written as 8 . Calculate the correct mean and standard deviation in each of the following case:

(i) If the wrong item is omitted.

(ii) If it is replaced by 12.

Solution:

(i) We have, $n = 20,\ \bar{x} = 10$ and $\sigma = 2$

$\therefore \quad \bar{x} = \dfrac{1}{n}\sum x_i \ \Rightarrow\ \sum x_i = n\bar{x} = 20 \times 10 = 200$

$\therefore$ incorrect $\sum x_i = 200$ and $\sigma = 2 \ \Rightarrow\ \sigma^2 = 4$

or $\dfrac{1}{n}\sum x_i^2 - \bar{x}^2 = 4$ or $\dfrac{1}{20}\sum x_i^2 - 100 = 4$ or $\sum x_i^2 = 2080$

$\therefore$ incorrect $\sum x_i^2 = 2080$

(i) If we omit the wrong item, 8 from the observations, then 19 observations are left

Correct $\sum x_i + 8$ = incorrect $\sum x_i$

$\therefore$ correct $\sum x_i = 200 - 8 = 192$

$\therefore$ correct mean = $\frac{192}{19} = 10.10$

and correct $\sum x_i^2 = 2080 - 64 = 2016$

Correct variance = $\frac{1}{19}$ (correct $\sum x_i^2$) – (correct mean)^2

$= \frac{2016}{19} - \left(\frac{192}{19}\right)^2 = \frac{1440}{361}$

$\therefore$ Correct standard deviation = $\sqrt{\frac{1440}{361}} = 1.997$

(ii) If we replace the wrong item by 12

Incorrect $\sum x_i - 8 + 12$ = correct $\sum x_i$

or correct $\sum x_i = 200 + 4 = 204$ and correct mean = $\frac{204}{20} = 10.2$

and incorrect $\sum x_i^2 - 8^2 + 12^2 = 2160$

Correct variance = $\frac{1}{20}$ (correct $\sum x_i^2$) – (correct mean)^2

$= \frac{2160}{20} - \left(\frac{204}{20}\right)^2 = \frac{1584}{400}$

$\therefore$ correct standard deviation = $\sqrt{\frac{1584}{400}}$ = 1.9899

***Example* 3:**

The variance of *n* observations is $\pm^2$. Show that if each observation is increased by *a*, then the variance of the new set of observations remains $\pm^2$.

Solution:

Let the observations be $x_1, x_2, x_3, ..., x_n$

$\therefore \quad \sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left[\frac{\sum_{i=1}^{n} x_i}{n}\right]^2$ …(i)

When each observation is increased by *a*, the observations become

$x_1 + a, x_2 + a, x_3 + a, ..., x_n + a$

Sum of these observations

$$= \sum_{i=1}^{n} (x_i + a) = \sum_{i=1}^{n} x_i + na \qquad \ldots(ii)$$

and sum of squares of these observations

$$= \sum_{i=1}^{n} (x_i + a)^2 = \sum_{i=1}^{n} \left(x_1^2 + a^2 + 2x_i a\right)$$

$$= \sum_{i=1}^{n} x_i^2 + na^2 + 2a\sum_{i=1}^{n} x_i \qquad \ldots(iii)$$

Hence, the variance of the new set of observations

$$= \frac{\sum_{i=1}^{n} x_i^2 + na^2 + 2a\sum_{i=1}^{n} x_i}{n} - \left[\frac{\sum_{i=1}^{n} x_i + na}{n}\right]^2 \qquad \text{(using (ii) and (iii))}$$

$$= \frac{\sum_{i=1}^{n} x_i^2}{n} + a^2 + 2a\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] - \left[\frac{\sum_{i=1}^{n} x_i}{n}\right]^2 - a^2 - \frac{2na\sum_{i=1}^{n} x_i}{n^2}$$

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

***Example* 4:**

A sample of 25 variates has the mean 40 and standard deviation 5 and a second sample of 35 variates has the mean 45 and standard deviation 2. Find the mean and standard deviation of the two samples of variates taken together.

Solution:

Here, $n_1 = 25, n_2 = 35$

$\bar{x}_1 = 40, \bar{x}_2 = 45$

$\sigma_1 = 5$ and $\sigma_2 = 2$

Let $\bar{x}$ be the mean and σ, the standard deviation of the two samples taken together, then

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{25 \times 40 + 35 \times 45}{25 + 35} = \frac{2575}{60} = 42.917$$

Also $\sigma^2 = \frac{1}{n_1 + n_2}\left[n\sigma_1^2 + n_2\sigma_2^2 + \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^2\right]$

$$= \frac{1}{25 + 35}\left[25(5)^2 + 35(2)^3 + \frac{25 \times 35}{25 + 35}(40 - 45)^2\right] = \frac{1}{60}[625 + 140 + 364.58]$$

$$= \frac{1129.58}{60} = 18.83 \text{ nearly}$$

$\Rightarrow \sigma = 4.34$ nearly.

***Example* 5:**

If $\bar{x}$ is the mean and mean deviation from mean is M.D. $(\bar{x})$, then find the number of observations lying between $\bar{x} - MD(\bar{x})$ and $\bar{x} + MD(\bar{x})$ for the following data:

34, 66, 30, 38, 44, 50, 40, 60, 42, 51

Solution:

Data arranged in ascending order is 30, 34, 38, 40, 42, 44, 50, 51, 60, 66.

$$\bar{x} = \frac{30 + 34 + 38 + 40 + 42 + 44 + 50 + 51 + 60 + 66}{10} = \frac{455}{10} = 45.5$$

$\left| x_i - \bar{x} \right| = 15.5,\ 11.5,\ 7.5,\ 5.5,\ 3.5,\ 1.5,\ 4.5,\ 5.5,\ 14.5,\ 20.5$

Mean deviation from the mean $= M.D.(\bar{x}) = \dfrac{\sum \left| x_i - \bar{x} \right|}{10}$

$$= \left(\frac{1}{10}\right)[15.5 + 11.5 + 7.5 + 5.5 + 3.5 + 1.5 + 4.5 + 5.5 + 14.5 + 20.5] = \frac{1}{10}[90.0] = 9$$

Now $\bar{x} - M.D.(\bar{x}) = 45.5 - 9 = 36.5$

and $\bar{x} + M.D.(\bar{x}) = 45.5 + 9 = 54.5$

Observations of arranged data which lie between 36.5 and 54.5 are 38, 40, 42, 44, 50, 51 which are six in number.

***Example* 6:**

Find the mean and standard deviation of the first *n* terms of an A.P. whose first term is a and common difference is *d*.

Solution:

Here the first *n* terms of the A.P. would be $a,\ a + d,\ a + 2d + ...,\ a + (n-1)d$ respectively.

Let the assumed mean be a

x_i	$d_i = x_i - a$	d_i^2
a	0	0
$a + d$	d	d^2
$a + 2d$	$2d$	$4d^2$
…….	……..	……..
……..	……..	…….
$a + (n-1)d$	$(n-1)d$	$(n-1)^2 d^2$

Now $\sum d_i = 0 + d + 2d + ... + (n-1)d = [1 + 2 + ... + (n-1)d = \dfrac{n(n-1)d}{2}$

and $\sum d_i^2 = 0 + d^2 + 4d^2 + ... + (n-1)^2 d^2 = [1^2 + 2^2 + ... + (n-1)^2 d^2 = \dfrac{n(n-1)(2n-1)d^2}{6}$

EXERCISE – I

1. Find the mean deviation from the mean for the following data:

(i) 30, 40, 70, 60, 20, 10, 50

(ii) 0, 2, 7, -5, 8, 11, 4, -3

2. Find the mean deviation from the median for the following data:

(i) 22, 24, 30, 27, 29, 31, 25, 41, 42

(ii) 3, 9, 21, 12, 5, 3, 18, 4, 7, 10, 19

3. Find the mean deviation about the mean for the following data:

(i)

x_i	2	5	6	8	10	12
f_i	2	8	10	7	8	5

(ii)

x_i	10	30	50	70	90
f_i	4	24	28	16	8

4. Find the mean deviation about the median for the following data:

(i)

x_i	15	21	27	30	35
f_i	3	5	6	7	8

(ii)

x_i	5	7	9	11	13	15	17
f_i	2	4	6	8	10	12	8

5. Find the mean deviation about median for the following data:

Class	0–10	10–20	20–30	30–40	40–50	50–60
Frequency	6	7	15	16	4	2

6. Find the mean deviation from the mean for the following observations:

Class	0-100	100 -200	200-300	300-400	400-500	500-600	600-700	700-800
Frequency	4	8	9	10	7	5	4	3

7. The height distribution of 100 children is as follows:

Height (cm)	95-105	105–115	115–125	125–135	135–145	145–155
No. of students	9	13	26	30	12	10

Calculate the mean deviation from the mean height.

8. Find the variance of the data:

(i) 6, 8, 10, 12, 14, 16, 18, 20, 22, 24.

(ii) 65, 58, 68, 44, 48, 45, 60, 62, 60, 50.

9. Find the mean, variance, and standard deviation of the following marks scored by 10 students:
45, 70, 62, 60, 50, 48, 67, 34, 65, 58.

10. Find the standard deviation for the following data:

x_i	3	8	13	18	23
f_i	7	10	15	10	6

EXERCISE – II

1. Find the mean and standard deviation for the following data:

x_i	92	93	97	98	102	104	109
f_i	3	2	3	2	6	3	3

2. The following frequency table gives the ages of a group of 50 children invited to a birthday party. Find the mean and standard deviation of the distribution:

Age (in years)	5-7	7-9	9-11	11-13	13-15
Frequency	16	13	10	6	5

3. The measurements (in *mm*) of the diameters of the heads of 107 screws gave the following frequency distribution.

Diameter (in mm)	33–35	36–38	39–41	42–44	45–47
Frequency	17	19	23	21	27

Find the variance and the standard deviation of this distribution.

4. Variance of *n* observations $x_1, x_2, x_3, ..., x_n$ is σ^2. Find the variance, if each observation is (i) decreased by *a* (ii) multiplied by *a*

5. If N = 10, $\bar{x} = 12$, $\sum x_i^2 = 1530$, find the coefficient of variation.

6. For a distribution, the coefficient of variation is 22.5% and the value of the arithmetic mean is 7.5. Find out the value of standard deviation.

7. From the data given below, state which group is more variable:

Class	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Group A (freq.)	9	17	32	23	40	18	1
Group B (freq.)	18	22	40	18	32	8	2

8. Three factories F_1, F_2 and F_3 show following results about the number of workers and the wages paid to them.

	F_1	F_2	F_3
No. of workers	5000	6000	4500
Average monthly wages	Rs 2500	Rs 2500	Rs. 3000
Variance of distribution	81	100	120

In which factory, F_1, F_2 or F_3 is there greater variability in individual wages?

9. The sum and sum of squares corresponding to length x (in cm) and weight y (in gm) of 50 plant products are given below:

$$\sum_{i=1}^{50} x_i = 212,\ \sum_{i=1}^{50} x_i^{\,2} = 902.8,\ \sum_{i=1}^{50} y_i = 261,\ \sum_{i=1}^{50} y_i^{\,2} = 1457.6$$

Find which is more varying, the length or the weight?

10. The mean and the standard deviation of a group of 100 observations were found to be 20 and 3 respectively. Later on it was found that three observations were incorrect, which were recorded as 21, 21 and 18. Find the mean and standard deviation, if the incorrect observations are omitted.

ANSWERS

ANSWERS TO PRACTICE PROBLEMS

PP1. 4.99

PP2. 10

PP3. 9.56

PP4. 2.72 , 7.4

PP5. 12.7, 21.5

PP6. 1

PP9. 3

PP10. mean = $a + \left(\frac{d-c}{n}\right)$; standard deviation = $\sqrt{nb^2 + \frac{(d-c)^2}{n}}$

PP12. 14.60

PP13. 39.9, 5

ANSWERS TO EXERCISE – I

1. (i) 17.1 (ii) 4.5

2. (i) 5.11 (ii) 5.27

3. (i) 2.26 (ii) 16

4. (i) 5.1 (ii) 2.72

5. 10.16

6. 157.92

7. 11.296

8. (i) 33 (ii) 66

9. 56.1, 115.93, 10.76

10. 6.12

ANSWERS TO EXERCISE – II

1. 100, 5.39

2. 8.84 years, 2.62 years

3. 17.79, 4.22

4. (i) σ^2 (ii) $a^2 \sigma^2$

5. 25%

6. 1.6875

7. Group B

8. Factory F_2

9. Weight

10. 20 and 3.036