CHAPTER IV

MEASURES OF DISPERSION

We have seen that, the whole data is represented by a single value known as average. It cannot describe the data completely. There may be two or more data sets with same mean but data sets may not be identical. For the following three sets of data we observe that mean for three data sets are same but data sets differ in variation.

Data sets	Observations					Total	AM
I	25	25	25	25	25	125	25
II	00	10	20	25	70	125	25
III	23	24	25	26	27	125	25

In first data set, all observations are constant whereas observations in the second data set are scattered or dispersed from mean and observations are close to mean in case of third data set.

To avoid this difficulty, it is necessary to study the variation. The variation is also known as dispersion. It gives the information, how individual observations are scattered or dispersed from the mean of a large series.

Deviation = observation - mean

Different Measures of Dispersion :

We shall study the following measures of dispersion.

1) Range
2) Quartile Deviation
3) Coefficient of Quartile Deviation
4) Mean deviation
5) Standard deviation
6) Variance and
7) Coefficient of Variation.

All these are absolute measures except 3 & 7, since the unit of the measure is same as that of the observations.

ABSOLUTE AND RELATIVE MEASURES OF DISPERSION :

Absolute measures possess units. They are useful for comparison of variability of two sets of data only when both are in the same units and their central values (i.e. average) are nearly equal. But in many problems one or both of these conditions are

not fulfilled. For example, we are interested to compare the variation in weight and variation in height of a group of persons. Weight may be measured in kg. and height may be in cm. Therefore, comparison is not possible until and unless unitless quantity is available. Therefore, we need the measures of dispersion which are independent of units. Such a measure can be obtained by taking the ratio of measure of dispersion and some central value of the data. It is called measure of relative dispersion. The relative measure is called as coefficient of the respective absolute measure.

1) RANGE :

Range is the simplest measure of dispersion. It is defined as the difference between the highest and the lowest values. For example, consider the data sets which are defined earlier.

For the first data set, the range = 25 -25 = 0. For the second data set, the range = 70 - 0 = 70. For the third data set, the range = 27 - 23 = 4. As it is based on two extreme observations, it is considered to be the poorest of the measures of dispersion. For example, the range of 5, 10, 15, 18, 80 is same as that of 5, 70, 72, 75, 80 however, variation patterns are different.

2) QUARTILE DEVIATION (QD) :

The quartile deviation (QD), which is also called semi interquartile range is some what analogous to the range of a distribution. If Q_1 and Q_3 are the first and third quartiles of a distribution, we have -

$$\text{Quartile deviation (QD)} = \frac{Q_3 - Q_1}{2}$$

The quartile deviation takes into account the middle half of the data between Q_3 and Q_1.

3) COEFFICIENT OF QUARTILE DEVIATION :

The relative measure of dispersion corresponding to quartile deviation is known as the coefficient of quartile deviation. It is defined as -

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

This will be always less than one and will be positive as $Q_3 > Q_1$, Smaller value of coefficient of QD indicates lesser variability.

Example : The following data gives the chest measurement of 100 students.

Chest measurement in cm. -	68-74	75-81	82-88	89-95	96-102	103-109
No. of students	5	31	40	20	3	1

Find (i) Range

(ii) Quartile deviation (QD) and

(iii) Coefficient of quartile deviation.

Solution :

Class boundaries	No. of students	Less than cumulative frequencies
67.5 - 74.5	05	05
74.5 - 81.5	**31**	**36**
81.5 - 88.5	**40**	**76**
88.5 - 95.5	20	96
95.5 - 102.5	03	99
102.5 -109.5	01	100
Total	**100**	

i) Range = 109.5 – 67.5 = 42

ii) Quartile deviation = $\frac{Q_3 - Q_1}{2}$

$$Q_1 = l + \frac{\frac{N}{4} - c.f.}{f} \times h$$

Here $\frac{N}{4}$ = 25, l =74.5, f = 31, cf = 5, and h = 7

Therefore, $Q_1 = 74.5 + \frac{25 - 5}{31} \times 7 = 79.02$

$$Q_3 = l + \frac{\frac{3N}{4} - c.f.}{f} \times h$$

Here 3 N/4 = 75 l = 81.5

f = 40 cf = 36 & h = 7

Therefore, $Q_3 = 81.5 + \frac{75 - 36}{40} \times 7 = 88.33$

Thus, $QD = \frac{88.33 - 79.02}{2} = 4.655$

iii) Coefficient of quartile deviation

$$\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{88.33 - 79.02}{88.33 + 79.02} = \frac{9.31}{167.35} = 0.0556319$$

4) MEAN DEVIATION (MD) / (AVERAGE DEVIATION) :

We have seen that, the range and quartile deviation are not proper measures of dispersion, since they are not based on all observations. In order to overcome the drawbacks of range and quartile deviation, we define mean deviation (MD).

If $\bar{x}$ be the mean of n observations like $x_1, x_2, \ldots x_n$ then, the difference $x_1 - \bar{x}_1, x_2 - \bar{x}_2, \ldots x_n - \bar{x}_n$ give deviation from the AM.

Any measure of this kind or, $\frac{\text{Sum } (x_1 - \bar{x})}{n}$, derived from these deviations is not going to be useful because some deviations being positive and some deviations being negative their sum is zero. We are removing the algebraic signs of these deviations getting the useful measure of dispersion. This is done by taking the absolute values of these differences. Thus, the mean deviation is given by -

$$\text{Mean deviation (MD)} = \frac{\text{Sum } | x - \bar{x} |}{n}$$

Where $| a |$ stands for absolute value of "a".

Note that $| a | = a$ and $| - a | = a$

In case of frequency distribution it is defined as -

$$MD = \frac{\text{Sum } f | x - \bar{x} |}{N} \quad \text{where } N = \text{Sum } f.$$

Coefficient of mean deviation about mean :

It is obtained by taking the division of MD and average. Coefficient of

$$MD = \frac{\text{MD about mean}}{\text{Mean}}$$

Example : Zaheer Khan bowled the following number of maiden overs in a cricket test series against England. 4,3,5, 2,6,4,7,5,0 & 4. Calculate the mean deviation and coefficient of mean deviation about mean.

Solution :

x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	
4	4 - 4 = 0	0	
3	3 - 4 = -1	1	
5	5 - 4 = 1	1	
2	2 - 4 = -2	2	
6	6 - 4 = 2	2	
4	4 - 4 = 0	0	
7	7 - 4 = 3	3	
5	5 - 4 = 1	1	
0	0 - 4 = -4	4	
4	4 -4 = 0	0	
Total 40	**0**	**14**	**Mean ($\bar{x}$) 40/10 = 4**

$$\text{Mean Deviation} = \frac{\text{Sum } | x - \bar{x} |}{n} = \frac{14}{10} = 1.4$$

$$\text{Coefficient of MD} = \frac{\text{Mean Deviation}}{\text{Mean}}$$

$$= \frac{1.4}{4} = 0.35$$

5) STANDARD DEVIATION (SD) :

The algebraic signs may be eliminated by taking the squares of the deviations instead of taking the absolute values of the deviations. The standard deviation (SD) is then defined as the positive square root of the arithmetic mean of the squares of the deviations taken from the arithmetic mean. It is denoted by symbol σ , a Greek alphabet, read as sigma .

Thus,

$$SD(\sigma) = \sqrt{\frac{\text{Sum } (x - x)^2}{n}}$$

This measures the scatter in the sample but we are interested for measuring the estimate of scatter in the population from which the sample is drawn. When the population mean is known, we can find deviation from it where population has "n" independent observations. The SD is obtained by dividing mean square deviations by no. of observations i.e. "n" . When population mean is not known, we can take sample mean as an estimate of population mean. In this case only n – 1 observations are independent (note that sum of mean deviation from mean is always zero i.e. n -1 observations and total of series are known, we can find the n^{th} observation). Therefore, when there are n observations in the data, the diviser is n – 1. In statistical language, n - 1 is called degree of freedom.

$$SD(\sigma) = \sqrt{\frac{\text{Sum } (x - \bar{x})^2}{n - 1}}$$

Example : During 6th May to 15th May 2002 the maximum daily temperature (in degrees centigrade) in city A happened to be 38 , 40, 42, 41, 39, 42, 40, 28, 41 & 39. Find the SD of the maximum daily temperature.

Solution

x	$x - \bar{x}$	$(x - \bar{x})2$
38	38 - 40 = -2	4
40	40 - 40 = 0	0
42	42 - 40 = 02	4
41	41-40 = 1	1
39	39 - 40 = -1	1
42	42-40 = 2	4
40	40 - 40 = 00	0
38	38 - 40 = 02	4
41	41 - 40 = 1	1
39	39 - 40 = -1	1
Total 400	**0**	**20**

Here $\bar{x} = \frac{\text{Sum } x}{n}$

$= \frac{400}{10}$

$= 40$

$\sigma = \sqrt{\frac{20}{9}}$

$= \sqrt{2.222222}$

$= 1.490712$

$\simeq 1.49$

SIMPLIFIED FORM :

If the mean ($\bar{x}$) of the series comes out to be integer, the above formula is easy for finding the SD But more often $\bar{x}$ is likely to contain decimal point. As a result, the deviations will also contain decimal places. Thus, it is laborious for calculations of S.D. Therefore, we simplify the above formula.

$$\sigma = \sqrt{\frac{\text{Sum } (x - \bar{x})^2}{n - 1}}$$

Squaring on both sides,

Therefore, $S^2 = \frac{\text{Sum } (x - \bar{x})^2}{n - 1}$

$= \frac{1}{n - 1} [\text{Sum } (x^2 - 2x\bar{x} + \bar{x}^2)]$

$= \frac{1}{n - 1} [\text{Sum } x^2 - \text{Sum } 2x\bar{x} + \text{Sum } \bar{x}^2)]$

$= \frac{1}{n - 1} [\text{Sum } (x^2 - 2\bar{x} \text{ Sum } x + n\bar{x}^2)]$

(Note that, Sum $\bar{x}^2 = n\bar{x}^2$)

$$= \frac{1}{n-1}[\text{Sum } x^2 - 2\bar{x}\, n\bar{x} + n\bar{x}^2)]$$

$$(\because \bar{x} = \frac{\text{Sum } x}{n}$$

$$\therefore \text{Sum } x = n\bar{x}$$

$$s^2 = \frac{1}{n-1}[\text{Sum } x^2 - 2n\bar{x}^2 + n\bar{x}^2)]$$

$$= \frac{1}{n-1}[\text{Sum } x^2 - n\bar{x}^2)]$$

Example : The foliowing numbers give the incubation period in days in ten consecutive patients of infectious hepatitis : 26, 22, 36, 15, 27, 19, 24, 18, 23 & 25. Calculate standard deviation.

Solution :

Here $\bar{x} = \frac{\text{Sum } x}{n} = \frac{235}{10} = 23.5$

x	x^2
26	676
22	484
36	1296
15	225
27	729
19	361
24	576
18	324
23	529
25	625
Sum x = 235	**Sum x^2 = 5825**

$$\sigma = \sqrt{\frac{\text{Sum } x^2 - n\bar{x}^2}{n-1}}$$

$$= \sqrt{\frac{5825 - 10 \times (23.5)2}{9}} = \sqrt{\frac{302.5}{9}} = 33.61$$

When observations are large in size, the formula which is used for finding SD is laborious. In order to overcome this difficulty we, use short cut method.

Short cut (Deviation) Method :

Step 1: Decide assumed mean "a"

Step 2 : Obtain deviation values i.e. $d = x - a$

Step 3 : Compute mean deviation

Step 4 : Apply formula and find SD

$$\sigma = \sqrt{\frac{\text{Sum } d^2 - n\bar{d}^2}{n-1}}$$

Example :

The prices (in Rs.) of 1 kg of edible oil for 7 days in a certain week are 38, 40, 42, 39, 41, 42 & 41. Find standard deviation.

Solution :

Let us consider a = 40

x	d = x - 40	d^2
38	38-40 = -2	4
40	40-40 = 0	0
42	42-40 = 2	4
39	39-40 = -1	1
41	41 - 40 = 1	1
42	42-40 = 2	4
41	41-40 = 1	1
	Sum d = 3	**Sum d^2 = 15**

$$\sigma = \sqrt{\frac{\text{Sum } d^2 - n\bar{d}^2}{n-1}}$$

$$= \sqrt{\frac{15 - 7 \times (3/7)^2}{7-1}} = \sqrt{2.2857143} = 1.51$$

Step deviation method : This method is used when deviation values are multiple of some number. The steps are -

Step 1 : Decide assumed mean "a"

Step 2 : Obtain the deviation values i.e. $d = x - a$

Step 3 : Find the value of d_1, $d_1 = \frac{d}{h}$, where h - class width.

Step 4 : Apply the following formula for finding SD

$$\sigma = \sqrt{\frac{\text{Sum } d_1^2 - n\bar{d}_1^2}{n-1}} \times h \text{ for ungrouped data}$$

$$\sigma = \sqrt{\frac{\text{Sum } (fd_1^2) - N\bar{d}_1^2}{N-1}} \times h, \text{ for grouped data}$$

Example : The following data provides the chest measurement in cm of 50 MBBS students.

Chest Measurement :	61-70	71-80	81-90	91-100	101-110
No. of students	2	10	20	17	1

Find the mean and standard deviation.

Solution :

Here a = 85.5 and h = 10

Class Interval	Mid values (x)	frequency (f)	$d_1 = \frac{x-a}{h}$	fd_1	$f.d_1^2$
61-70	65.5	2	–2	– 4	8
71-80	75.5	10	–1	–10	10
81-90	85.5	20	0	00	00
91-100	95.5	17	1	17	17
101-110	105.5	1	2	2	4
	Sum f = 50			**+5**	**39**

$$\bar{d}_1 = \frac{\text{Sum } fd_1}{\text{Sum } f} = \frac{5}{50} = 0.1$$

$$\text{Mean} = a + \bar{d}_1 \times h$$

$$= 85.5 + 0.1 \times 10 = 86.5$$

$$\sigma = \sqrt{\frac{\text{Sum } f \cdot \bar{d}_1^2 - N\bar{d}_1^2}{N-1}} \times h$$

$$= \sqrt{\frac{39 - 50 \times (.1) \times (.1)}{50 - 1}} \times 10 = 8.86$$

6) VARIANCE :

The square of the standard deviation of a set of observations is called the variance.. Obviously it is denoted by the symbol σ^2.

Thus,

$$\sigma^2 = \frac{\text{Sum } (x - \bar{x})^2}{n-1}$$ for ungrouped data

$$\sigma^2 = h \times \frac{\text{Sum } f (x - \bar{x})^2}{N-1}$$, (where h = class interval) for grouped data

MERITS AND DEMERITS OF S.D. :

Merits of S.D. :

1) It is rigidly defined.
2) It is based on all observations.
3) It does not ignore the algebraic signs of deviations.
4) It is capable of further mathematical treatment.
5) It is not much affected by sampling fluctuations.

Demerits of S.D. :

1) It is difficult to understand and calculate.
2) It cannot be calculated for qualitative data and distribution with open end classes.
3) It is unduly affected due to extreme deviations.

7) COEFFICIENT OF VARIATION (CV) :

The absolute measure i.e. standard deviation is not useful for comparing the variability of two frequency distributions measured in different units or with widely differing means. To overcome these difficulties, the relative measure known as coefficient of variation is defined, which is given as -

$$CV = \frac{\text{Standard deviation}}{\text{mean}} \times 100$$

Coefficient of variation is always expressed in percentage.

Example - The pulse and respiratory rates recorded per minute for the same group of 62 inpatients of a certain public hospital are as given below :-

A) Pulse rate	70-74	75-79	80-84	85-89	90-94	95-99
No. of patients	8	5	24	4	14	7
B) Respiratory rate	14-16	17-19	20-22	23-25	26-28	29-31
No. of patients	10	19	17	9	4	3

Which of the two rates is more consistent?

Solution : a = 82 and h = 5