DISPERSION

Average or measures of central tendency give us an idea of the concentration of the observations about the central part of the distribution.

Let us consider the following three set of data

(i) x_i : 7, 8, 9, 10, 11 $\rightarrow$ $\Sigma x_i = 45$ & $\bar{x} = 9$

(ii) x : 3, 6, 9, 12, 15 $\rightarrow$ $\Sigma x_i = 45$ & $\bar{x} = 9$

(iii) x : 1, 5, 9, 13, 17 $\rightarrow$ $\Sigma x_i = 45$ & $\bar{x} = 9$

In all the above cases we have 5 observation with mean 9. If we are given that the mean of 5 observation is 9, we cannon form an idea as to whether it is the average of 1st set of data or 2nd set of data or 3rd set of data or some other set of data. whose sum is 45.

Thus we see that the measures of central tendency are inadequate to give us a complete idea of the distribution. They must be supported and supplemented by some other measures.

One such measures is Dispersion.

Literal meaning of dispersion is "scatteredness". Dispersion gives us an idea about the homogeneity or heterogeneity of the distribution.

Measures of Dispersion :

Various measures of dispersion are as follows :

(i) Range

(ii) Mean deviation

(iii) Standard deviation.

I. Range :

The simplest measure of the dispersion of a variable is its range, which is defined as the difference between the highest (maximum) and the lowest (minimum) values of the observation/variable.

Let us consider an example here. Suppose two students, A and B of a college received the following marks in eight monthly examination in a particular subject :

Marks obtained by A	Marks obtained by B
63	61
47	54
56	56
44	57
66	60
65	59
80	55
43	62

In this example, average score of both the student A & B is same, i.e., 58.

In this example, the range of the marks obtained by A is 80 - 43 = 37 and that of B is 62 - 54 = 8.

(ii) Mean deviation

Let x_A be the chosen average value of the variable x, then $x_i - x_A$ is the deviation of the ith given value of x from the average. Clearly, the higher the deviations

$x_1 - x_A, x_2 - x_A, \ldots, x_n - x_A$

in magnitude, the higher is the dispersion of x. One may therefore, consider some way of combining the deviations to get a measure of dispersion. It is readily seen that the simple arithmatic mean of the deviations, viz., $\frac{1}{n}\sum_i (x_i - x_A)$, cannot serve this purpose, as the sum sum of the deviations - and proportionally the arithmatic mean - may be quite small even when the individual deviations are large, positive and negative deviations almost cancelling each other. In fact, if x_A is considered to the arithmatic mean of x, then the sum of the deviation vanishes, whatever the deviations are individually. This diticulty may be overcome by considering, instead of the deviations themselves, their absolute values, in which ~~their~~ the magnitude of the deviations (and not their sign) will be considered. The arithmatic mean of the absolute deviations of x_i from x_A ~~will be~~ is the required measure of dispersion and is referred to as the mean deviation of x about x_A, denoted by MD_A and given by

$$MD_A = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_A| \qquad (2.1)$$

It can be shown that MD_A is least when measure about median. Let us consider the same example as discussed above.

Marks obtained by A $(x^{(1)})$	$x_i^{(1)} - \bar{x}^{(1)}$
63	5
47	-11
56	-2
44	-14
66	8
65	7
80	22
43	-15

Marks obtained by B $(x^{(2)})$	$x_i^{(2)} - \bar{x}^{(2)}$
61	3
54	-4
56	-2
57	-1
60	2
59	1
55	-3
62	4

Let

$\bar{x}^{(1)} = \bar{x}^{(2)} = \bar{x} = 58 = x_A$ (Arithmatic Mean).

$\therefore \sum_{i=1}^{8} |x_i^{(1)} - x_A| = 5+11+2+14+8+7+22+15 = 84$

$\therefore \frac{1}{8} \sum_{i=1}^{8} |x_i - x_A| = \frac{84}{8} = 10.5$

& $\sum_{i=1}^{8} |x_i^{(2)} - x_A| = 3+4+2+1+2+1+3+4 = 20$

$\therefore \frac{1}{8} \sum_{i=1}^{8} |x_i^{(2)} - x_A| = \frac{20}{8} = 2.5.$

$\therefore$ Mean deviation of the marks obtained by the students A and B about the arithmatic mean 58 are 10.5 and 2.5, respectively.

Standard Deviation

If (x_i, f_i), $i = 1, 2, \ldots, n$ be the frequency distribution of a variable x, the mean deviation of x about the average x_A (may be mean, median or mode) is given by

$$MD_A = \frac{1}{N} \sum_{i=1}^{n} f_i |x_i - x_A| \; ; \; \sum_{i=1}^{n} f_i = N \quad \text{—— (2.2)}$$

Example 9.

Calculate the mean deviation from mean (A.M) for the following data:

Marks :	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of students :	6	5	8	15	7

Marks :	50 – 60	60 – 70
No. of students :	6	3

Soln.

Marks	Mid point (x_i)	No. of students (f_i)	$d_i = \frac{x_i - A}{h}$	$d_i f_i$	$\lvert x_i - \bar{x} \rvert = \lvert x_i - 33.4 \rvert$
0 – 10	5	6	−3	−18	28.4
10 – 20	15	5	−2	−10	18.4
20 – 30	25	8	−1	−8	8.4
30 – 40	35 (= A)	15	0	0	1.6
40 – 50	45	7	1	7	11.6
50 – 60	55	6	2	12	21.6
60 – 70	65	3	3	9	31.6
		50 (= N)		−8	

$N = \sum_{i=1}^{7} f_i = 50$

Let $A = 35$ & $h = 10$

$\therefore \quad A.M. = \bar{x} = A + \frac{h}{N} \sum_{i=1}^{7} f_i d_i = 35 + \frac{10}{50}(-8)$

$= 35 - 1.6 = \boxed{33.4 = \bar{x}}$

$$MD_{\bar{x}} = \frac{1}{N}\sum_{i=1}^{7} f_i |x_i - \bar{x}|$$

$$= \frac{1}{50}\Big[6 \times 28.4 + 5 \times 18.4 + 8 \times 8.4 + 15 \times 1.6$$
$$+ 7 \times 11.6 + 6 \times 21.6 + 3 \times 31.6\Big]$$

$$= \frac{659.2}{50}$$

$$= 13.184$$

Standard deviation and Root Mean square deviation

Standard deviation, usually denoted by the Greek letter small sigma (σ), is the positive square root of the arithmatic mean of the squares of the deviation of the values of the variable x, i.e., x_i $(i = 1, 2, \dots, n)$ from their arithmatic mean $\bar{x}$. and is given by

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{—— (2.3)}$$

For the frequency distribution (x_i, f_i), $i = 1, 2, \dots, k$

s.d. of the variable x is given by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}, \quad N = \sum_{i=1}^{k} f_i \quad \text{—— (2.4)}$$

The square of the standard deviation (s.d.) is called the variance and is given by

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{—— (2.5)}$$

OR,

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 \quad \text{—— (2.6)}$$

Root mean square deviation, denoted by 's' is given by :

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - A)^2}, \quad N = \sum_{i=1}^{k} f_i \quad \text{—— (2.7)}$$

OR

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2}, \quad \text{—— (2.8)}$$

where A is an arbitrary number.

s^2 is called the mean square deviation.

Relation between σ and s.

By definition

$$s^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - A)^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \bar{x} + \bar{x} - A)^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i \{(x_i-\bar{x})^2 + (\bar{x}-A)^2 + 2(\bar{x}-A)(x_i-\bar{x})\}$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i (x_i-\bar{x})^2 + (\bar{x}-A)^2 + 2(\bar{x}-A)\frac{1}{N}\sum_{i=1}^{k} f_i (x_i-\bar{x})$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i (x_i-\bar{x})^2 + (\bar{x}-A)^2 + 0 \quad \because \sum_{i=1}^{k} f_i (x_i-\bar{x}) = 0$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i (x_i-\bar{x})^2 + (\bar{x}-A)^2$$

$$\therefore s^2 = \sigma^2 + d^2, \text{ where } d = \bar{x} - A.$$

$\therefore s^2 \geq \sigma^2$

$\Rightarrow s^2$ will be least when ~~da~~ $d = 0$, i.e., $\bar{x} = A$.

Hence, we can conclude that mean square deviation and consequently root mean square deviation is least when the deviations are taken about $A = \bar{x}$, i.e., standard deviation is the least value of root mean square deviation.

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

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

$$\therefore \boxed{\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \qquad (2.9)$$

Result 2.

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - 2\bar{x}\frac{1}{N}\sum_{i=1}^{k} f_i x_i + \bar{x}^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2 \qquad (2.10)$$

Variance of the combined series

Let us suppose that we have two series of data of sizes n_1 and n_2, with mean $\bar{x}_1$ and $\bar{x}_2$, and standard deviations σ_1 and σ_2, respectively. Then the standard deviation σ of the combined series of data of size $n_1 + n_2$ is given by

$$\sigma^2 = \frac{1}{n_1+n_2}\left[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)\right], \qquad (2.11)$$

where $d_1 = \bar{x}_1 - \bar{x}$, $d_2 = \bar{x}_2 - \bar{x}$ and $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$, is the mean of the combined series.

Proof: Let x_{1i}; $i = 1, 2, \ldots, n_1$ and x_{2j}; $j = 1, 2, \ldots, n_2$ be two series of data, then

$$\left.\begin{aligned}\bar{x}_1 &= \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}\\ \bar{x}_2 &= \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}\end{aligned}\right\} \quad \text{and} \quad \left.\begin{aligned}\sigma_1^2 &= \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2\\ \sigma_2^2 &= \frac{1}{n_2}\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2\end{aligned}\right\}$$

Then the mean $\bar{x}$ of the combined series is given by

$$\bar{x} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j}\right) = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2}$$

The variance σ^2 of the combined series of data is given by

$$\sigma^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x})^2\right]$$

Now

$$\sum_{i=1}^{n_1}(x_{1i}-\bar{x})^2 = \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1+\bar{x}_1-\bar{x})^2$$

$$= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + n_1(\bar{x}_1-\bar{x})^2 + 2(\bar{x}_1-\bar{x})\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)$$

$$= \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + n_1(\bar{x}_1-\bar{x})^2 + 0 \quad \because \sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1) = 0$$

$$= n_1\sigma_1^2 + n_1(\bar{x}_1-\bar{x})^2$$

$$= n_1\sigma_1^2 + n_1 d_1^2, \text{ where } d_1 = \bar{x}_1 - \bar{x}$$

Similarly, we get

$$\sum_{j=1}^{n_2}(x_{2j}-\bar{x})^2 = n_2\sigma_2^2 + n_2 d_2^2, \text{ where } d_2 = \bar{x}_2 - \bar{x}$$

Hence, $$\sigma^2 = \frac{1}{n_1+n_2}\left[n_1(\sigma_1^2+d_1^2) + n_2(\sigma_2^2+d_2^2)\right]$$

Proved.

Variance and consequently standard deviation is independent of change of origin but not of scale.

Proof: ~~$s^2 = \frac{1}{N}\sum_{i=1}^{k} f_i x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{k} f_i x_i\right)^2$~~

Let $d_i = x_i - A$, A is an arbitrary value

$$\therefore \frac{1}{N}\sum_{i=1}^{k} f_i d_i = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - A) = \frac{1}{N}\sum_{i=1}^{k} f_i x_i - A$$

$$\Rightarrow \bar{d} = \bar{x} - A.$$

$$\therefore d_i - \bar{d} = (x_i - A) - (\bar{x} - A) = x_i - \bar{x}$$

$$\therefore \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (d_i - \bar{d})$$

$$\boxed{\sigma_x^2 = \sigma_d^2} \qquad \text{——— (2.12)}$$

$$\text{OR, } \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} (d_i - \bar{d})^2 = \sigma_d^2$$

$\Rightarrow$ variance is independent of change of origin.

Let $d_i = \frac{x_i - A}{h} \Rightarrow x_i = A + h d_i$

$$\therefore \bar{x} = A + \frac{h}{N}\sum_{i=1}^{k} f_i d_i = A + h\bar{d}$$

$$\therefore x_i - \bar{x} = h(d_i - \bar{d})$$

$$\therefore \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \bar{x})^2 = \frac{h^2}{N}\sum_{i=1}^{k} f_i (d_i - \bar{d})^2 = h^2 \sigma_d^2$$

$$\text{OR, } \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} (d_i - \bar{d})^2 = h^2 \sigma_d^2$$

$$\therefore \boxed{\sigma_x^2 = h^2 \sigma_d^2} \qquad \text{——— (2.13)}$$

$\Rightarrow$ Variance is independent of change of origin but not of scale.

Example 10.

Calculate the mean and standard deviation for the following table giving the the age distribution of 542 members.

Age (in years) :	20-30	30-40	40-50	50-60	60-70
No. of members :	3	61	132	153	140

Age (in years) :	70-80	80-90
No. of members :	51	2

Soln. Let $d_i = \frac{x_i - A}{h} = \frac{x_i - 55}{10}$.

Age group	Mid Value (x_i)	No. of members (f_i)	$d_i = \frac{x_i - A}{h}$	$f_i d_i$	$f_i d_i^2$
20-30	25	3	-3	-9	27
30-40	35	61	-2	-122	244
40-50	45	132	-1	-132	132
50-60	55(=A)	153	0	0	0
60-70	65	140	1	140	140
70-80	75	51	2	102	204
80-90	85	2	3	6	18
Total		542	0	-15	765

$$\bar{x} = A + \frac{h}{N}\sum_i f_i d_i = 55 + \frac{10 \times (-15)}{542} = 55 - 0.28$$

$$= 54.72 \text{ (years)}. \quad \simeq 55$$

$$\therefore \sigma_x^2 = h^2 \sigma_d^2 = h^2 \left[\frac{1}{N}\sum_i f_i d_i^2 - \left(\frac{1}{N}\sum f_i d_i\right)^2\right]$$

$$= 100\left[\frac{765}{542} - \left(\frac{-15}{542}\right)^2\right] = 141.07.$$

$\therefore$ standard deviation $= \sigma_x = \sqrt{141.07}$ years $= 11.80$ yrs.

Example 11.

~~Calculation~~ Calculate the s.d. of the marks obtained by two students A and B of a college in eight monthly examination in a particular subject.

Marks obtained by A : 63 47 56 44 66 65 80 43

Marks obtained by B : 61 54 56 57 60 59 55 62

Soln. Let us consider series I $(x^{(1)})$ as marks obtained by A and series II $(x^{(2)})$ as the marks obtained by B.

$x_i^{(1)}$	$d_i^{(1)} = x_i^{(1)} - \bar{x}$	$(d_i^{(1)})^2$	$x_i^{(2)}$	$d_i^{(2)} = x_i^{(2)} - \bar{x}$	$(d_i^{(2)})^2$
63	5	25	61	3	9
47	-11	121	54	-4	16
56	-2	4	56	-2	4
44	-14	196	57	-1	1
66	8	64	60	2	4
65	7	49	59	1	1
80	22	484	55	-3	9
43	-15	225	62	4	16
Total	0	1168	464	0	60

$\therefore \overline{x^{(1)}} = 58 = \overline{x^{(2)}} = \bar{x}$

$\therefore$ Arithmatic mean of the marks obtained by both the students A and B are same, i.e., 58.

$$\sigma_{x^{(1)}}^2 = \sigma_{d^{(1)}}^2 = \frac{1}{n}\sum_i (d_i^{(1)})^2 - \left(\frac{1}{n}\sum_i d_i^{(1)}\right)^2$$

$$= \frac{1}{8} \times 1168 - \left(\frac{1}{8} \times 0\right) = 146$$

$$\therefore \sigma_{x^{(1)}} = \sqrt{146} = 12.08$$

$$\& \quad \sigma_{x^{(2)}}^2 = \sigma_{d^{(2)}}^2 = \frac{1}{n}\sum_i (d_i^{(2)})^2 - \left(\frac{1}{n}\sum_i d_i^{(2)}\right)^2$$

$$= \frac{1}{8} \times 60 - \left(\frac{1}{8} \times 0\right) = 7.5$$

$$\therefore \sigma_{x^{(2)}} = 2.739$$