Five Number Summary:-

One way of Summarising data is to give values which provide essential information about the data set.

Minimum value, Q_1, Q_2 = Median, Q_3 and maximum value.

Box and whisker plot:-

0 2 5 2 0 4 4 8 9 8 8

Sorted data

0 0 2 2 4 4 5 8 8 8 9

Minimum value = 0

Maximum value = 9

$Q_1 = 2$

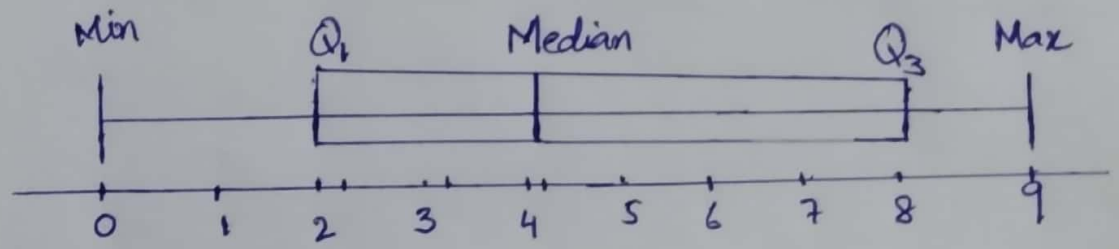
$Q_2 = 4$

$Q_3 = 8$

⇒ Draw a box or rectangle above Q_1 and Q_3

⇒ Make a line inside box above median.

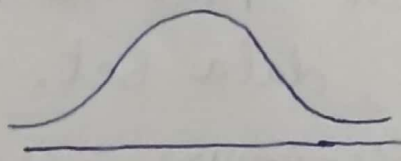
⇒ Draw two whiskers. Left whisker extends from lower Quartile to minimum value and right whisker from upper Quartile to maximum value.



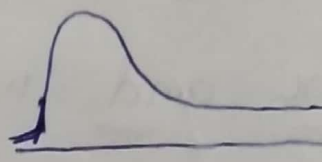
upper fence $= Q_3 + 1.5(Q_3 - Q_1)$

Lower fence $= Q_1 - 1.5(Q_3 - Q_1)$

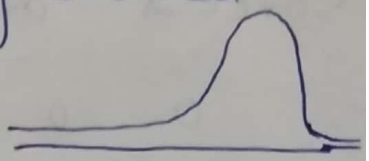
$\Rightarrow$ If $Q_3 - Q_2 = Q_2 - Q_1$ it shows symmetrical distribution.



$\Rightarrow$ If $Q_3 - Q_2 > Q_2 - Q_1$ then Positive skewed distribution.



$\Rightarrow$ If $Q_3 - Q_2 < Q_2 - Q_1$ then the distribution is Negatively skewed



upper fence $= 8 + 1.5(8 - 2)$

$= 17$

Lower fence $= 2 - 1.5(8 - 2)$

$= -7$

So no outliers in this data.

As

$Q_3 - Q_2 > Q_2 - Q_1$

$8 - 4 > 4 - 2$

$4 > 2$

So it is positively skewed.

Measures of dispersion

It is quite possible that two or more sets of data may have the same average but their indivisual observations may differ considerably from Average. Thus a value of central tendency does not adequately describe the data. We therefore need some additional information concerning with how the data values are dispersed about the average. This is done by measuring the dispersion by which we mean the extend to which the observations in a sample or population vary about their mean. A quantity that measures this characteristic is called a measure of dispersion, Scatter or variability.

e.g consider the two sets of data A and B.

A :	48	52	60	60	60	68	72
B:	0	10	60	60	60	110	120

For both sets Mean = Median = Mode = 60

but B is much more spread out than set A.

⇒ Absolute measure of dispersion

⇒ Relative measure of dispersion.

⇒ An absolute measure of dispersion is one that measures the dispersion in terms of the same units or squares of units, as the units of data.

⇒ A relative measure of dispersion is one that is expressed in the form of ratio, co-efficient or percentage and is independent of the units of measurement. It is useful for comparison of data of different nature.

Range :- $R = x_m - x_0$

Range = Largest - Smallest value

Range of set A = $72 - 48 = 24$

" " B = $120 - 0 = 120$

set B is more spread out than A.

Relative measure of Range :-

Co-efficient of Dispersion = $\frac{x_m - x_0}{x_m + x_0}$

This is a pure (i.e Dimensionless) number and used for comparison.

(Example 4.1)

Interquartile Range :-

The interquartile range is the difference b/w first and third Quatile- $Q_3 - Q_1$ and half of the range is called semi-interquartile range, or the quartile deviation.

$$Q.D = \frac{Q_3 - Q_1}{2}$$

IQ Range is just range of middle 50% of the distribution.

⇒ Its relative measure of dispersion is called co-efficient of Q.D.

$$\text{co-efficient of } Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

which is a pure number and is used for comparison. (Example 4.2)

Mean Deviation (OR Average deviation) :-

The M.D of a set of data is defined as the arithmatic mean of the deviations measured either from the mean or from the median, all deviation being counted as positive.

$$M.D = \frac{\Sigma |x - \bar{x}|}{n} \quad \text{(ungroup data)}$$

$$M.D = \frac{\Sigma f|x - \bar{x}|}{\Sigma f} \quad \text{(Group data)}$$

Co-efficient of M.D = $\frac{M.D}{Mean}$ or $\frac{M.D}{Median}$

(Example 4.3 & 4.4)

The variance and standard deviation

Variance is defined as the mean of the squares of deviations of all the observations from their mean.

$$S^2 = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n}$$ Sample variance

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x-\mu)^2}{N}$$ Population variance

⇒ An alternative measure of spread which does take into account the spread of all the values can be devised by finding how far each value is from the mean.

⇒ variance is measured in units^2.

⇒ S.D is square root of variance and measured in same units as the original data values.

$$S.D = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

$$S.D = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$$ (Grouped data)

$$Var = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$$

Co-efficient of variation :- (C.V)

The variability of two or more data sets is compared by C.V.

$$C.V = \frac{S}{\bar{x}} \times 100$$

⇒ A large value of C.V indicates that the variability is greater and consistency is less.

⇒ C.V is used to compare the performance of two players or candidates.

(Example 4.9)
& 4.10

Properties of variance and S.D :-

1) $Var(a) = 0$	$S.D(a) = 0$
2) $Var(x+a) = Var(x)$	$S.D(x+a) = S.D(x)$
3) $Var(ax) = a^2 Var(x)$	$S.D(ax) = \lvert a \rvert S.D(x)$ as S.D cannot be negative.
4) $Var(X \pm Y) = Var(x) + Var(Y)$	$S.D(X \pm Y) = \sqrt{Var(x) + Var(Y)}$

)