0302301 - STATISTICS FOR DATA ANALYSIS

UNIT	MODULE	WEIGHTAGE
1.	STATISTICS: OVERVIEW	20%
2.	MEASURE OF DISPERSION	20%
3.	CORRELATION AND REGRESSION	20%
4.	FUNDAMENTALS OF PROBABILITY	20%
5.	STATISTICAL ANALYSIS USING R PROGRAMMING	20%

0302301 - STATISTICS FOR DATA ANALYSIS

Text Book: Business statistics by Padmalochan hazarika

Related Programming Tool: R

UNIT - 2 Measures of Dispersion

- Quartile
- Range
- Quartile Deviation
  - Coefficient of Q.D.
  - Advantage and Disadvantage of Q.D.
- Mean Deviation
  - Coefficient of M.D.
  - Advantage and Disadvantage of M.D.
- Standard Deviation
  - Coefficient of S.D.
  - Advantage and Disadvantage of S.D.
- Relationships among Q.D., M.D., S.D.

RANGE

- The range of a distribution is the difference between the largest and smallest observation of that distribution.
- If L denotes the largest observation and S denotes the smallest observation of a distribution then the range R of the distribution will be

$$\underline{\mathbf{R = L - S}}$$

RANGE

- If the marks obtained by six students are 24, 12, 16, 11, 40 and 42, find the range of these marks.
- **SOLUTION**

  24, 12, 16, 11, 40, 42

  L = 42, S = 11
- **Range R = L - S**

  = 42 - 11

  = 31

RANGE

- Determine range for the following distribution.

Weight (k.g.)	40	47	56	62	70
No. of students	4	7	11	3	1

- L = 70 kg, S = 40 kg,
- **Range R = L - S**

  = 70 - 40 (kg)

  = 30 kg

RANGE

- The following distribution is a distribution of height. Determine range for the following distribution.

Height (c.m.)	120-129	130-139	140-149	150-159
No. of person	10	17	23	8

- **Range R = (Upper limit of 150-159) - (Lower limit of 120-129)**

  = 159 - 120

  = 39 cm

RANGE

- Advantage
  - It is easy to understand and calculate range.
- Disadvantage
  - Depends on only the two extreme values of data.
  - May arrive at wrong conclusion.
  - Can’t obtain range of a distribution with either one or both first and last class interval being open.

QUARTILES

- The three quantities of a distribution in ascending order which divide the distribution into four equal parts are called **quartiles** the distribution.
- Denoted by Q1, Q2, Q3.
- Q1: first quartile / lower quartile (25%)
- Q2: second quartile / median (50%)
- Q3: third quartile / upper quartile (75%)

QUARTILES

- For ungrouped frequency:

In case of *ungrouped frequency distributions*, the quartiles are determined by using the following formula :

$$Q_i = \text{value of } \frac{i\,(N+1)}{4} \text{ th term, } i = 1, 2, 3;$$

or, $$Q_i = \text{value of } \frac{iN}{4} \text{ th term, } i = 1, 2, 3 \qquad \text{.... 5.13}$$

Here N = Total frequency.

- For grouped frequency:

The formulae for determining the three quartiles in case of *grouped frequency distributions* are as follows :

$$Q_i = L + \frac{\frac{iN}{4} - f_c}{f} \times l, \; i = 1, 2, 3 \qquad \text{...(5.13)}$$

where L = Lower limit/lower boundary of the quartile class,

f = Frequency of the quartile class,

f_c = Cumulative frequency of the class preceeding the quartile class,

and l = Length of the quartile class. (The letter h may also be used in place of l)

Note: (*i*) The formula for Q_2 is same as that of median.

(*ii*) The Q_i (i = 1, 2, 3) is that class in which the $\frac{i\,(N+1)}{4}$ th term or the $\frac{iN}{4}$ th term lies (i = 1, 2, 3)

QUARTILES

Example 28. Determine quartiles from the following distribution :

Weight (lbs) x :	4	5	6	8	11	13	14
No. of children (f) :	2	4	5	7	3	2	1

Solution:

Table 5.15: Table for determining quartiles

x	f	f_c
4	2	2
5	4	6
6	5	11
8	7	18
11	3	21
13	2	23
14	1	24
	$N = 24$	

Q_1 = Value of $\frac{N+1}{4}$ th observation *i.e.*, $\frac{24+1}{4}$ th or 6.25th observation

= Value of 6th item + 0.25 (value of 7th item – value of 6th item)

= 5 + 0.25 (6 – 5) = 5.25

Thus **Q_1 = 5.25 lbs.**

Q_2 = Value of $\frac{2(N+1)}{4}$ th item = Value of $\frac{2(24+1)}{4}$ th item

= Value of 12.5th item

∴ Q_2 = Value of 12th item + 0.5 (Value of 13th item – value of 12th item)

= 8 + 0.5 (8 – 8) = 8

Thus **Q_2 = 8 lbs.**

Q_3 = Value of $\frac{3(N+1)}{4}$ th item = Value of $\frac{3(24+1)}{4}$ th item

= Value of 18.75th item.

∴ Q_3 = Value of 18th item + 0.75 (value of 19th item – value of 18th item)

= 8 + 0.75 (11 – 8) = 8 + 0.75 × 3 = 8 + 2.25 = 10.25

Thus **Q_3 = 10.25 lbs.**

QUARTILES

Example 27. Determine the quartiles for the following distribution :

Class interval :	10–15	15–20	20–25	25–30	30–40
Frequency :	4	12	16	22	10
Class interval :	40–50	50–60	60–70		
Frequency :	8	6	4		

Table 5.14: Table for determining quartiles

Class interval	*Frequency* (f)	*Cumulative frequency* (f_c)
10 – 15	4	4
15 – 20	12	16
20 – 25	16	32
25 – 30	22	54
30 – 40	10	64
40 – 50	8	72
50 – 60	6	78
60 – 70	4	82
	$N = 82$	

$$Q_1 = \text{Value of } \frac{N+1}{4} \text{th observation}$$

$$= \text{Value of } \frac{82+1}{4} \text{th observation} = \text{Value of } 20{\cdot}75\text{th observation}$$

We see from the above cumulative frequency table that the 20·75th observation is included in the class 20–25.

Now, $Q_1 = L + \dfrac{\frac{N}{4} - f_c}{f} \times I$ [Putting $i = 1$ in formula 5.13]

Here $L = 20$, $N = 82$, $f = 16$ and $f_c = 16$.

$\frac{82}{4}$

$$\therefore Q_1 = 20 + \frac{20.5 - 16}{16} \times 5 = 20 + \frac{22.5}{16} = 20 + 1.41 = \mathbf{21.41}$$

Q_2 = Value of $\dfrac{2(N+1)}{4}$ th observation = Value of $\dfrac{2 \times 83}{4}$ th observation value of 41.5th observation. This observation lies in the class 25 – 30.

Now, $$Q_2 = L + \frac{\frac{2N}{4} - f_c}{f} \times I$$

Here $L = 25$, $N = 82$, $f = 22$, $f_c = 32$, $I = 5$

$$\therefore Q_2 = 25 + \frac{41 - 32}{22} \times 5 = 25 + \frac{45}{22} = 25 + 2.05 = \mathbf{27.05.}$$

Q_3 = Value of $\dfrac{3(N+1)}{4}$ th *i.e.* $\dfrac{3 \times 83}{4}$ th or 62.25th observation. This observation is included in the class 30-40.

$$\therefore Q_3 = L + \frac{\frac{3N}{4} - f_c}{f} \times I = 30 + \frac{61.5 - 54}{10} \times 10 = 30 + 7.5 = \mathbf{37.5.}$$

Interquartile Range and Quartile Deviation

- The **interquartile range** of a distribution is the difference between the **third quartile Q3** and **first quartile Q1** of the distribution.
- Half the interquartile range of a distribution is called quartile deviation (Q.D.) of the distribution.
- **Interquartile range = Q3 - Q1**
- **Quartile Deviation: Q3 - Q1 / 2**
- **Coefficient of Quartile Deviation: (Q3 - Q1 / 2) / (Q3 + Q1 / 2)**
    - **Q3 - Q1 / Q3 + Q1**

Interquartile Range and Quartile Deviation

Ungrouped freq.

Example 4. The following is the distribution of wages of some workers. Determine quartile deviation and coefficient of quartile deviation.

Wages (in ₹) :	20	32	61	75	82	95
No. of workers :	2	4	7	5	4	2

Solution : To determine quartiles the distribution must be in a definite order. The given distribution is in ascending order.

Table 6.1: Table for determining Q. D. and Coeff. of Q. D.

Wages (in ₹) (x)	*No. of Workers* (f)	*Frequencies* (f_c)
20	2	2
32	4	6
61	7	13
75	5	18
82	4	22
95	2	24
	$N = 24$	

$$Q_1 = \text{Value of } \frac{N+1}{4} \text{th term}$$

$$= \text{Value of } \frac{24+1}{4} \text{th term} = \text{value of } 6\frac{1}{4} \text{ th term}$$

$$= \text{Value of 6th term} + \frac{1}{4}\text{(value of 7th term} - \text{value of 6th term)}$$

$$= 32 + \frac{1}{4}(61-32) = 32 + \frac{1}{4} \times 29 = 32 + 7.25 = 39.25$$

i.e., $Q_1 = ₹\,39.25.$

(**N.B.** The place of the $6\frac{1}{4}$ th term will be at the place which is at a distance of $\frac{1}{4}$ th of the distance between the 6th and the 7th terms. It is to be noted that a quartile of a distribution may not be a value of that distribution. The values of the various terms are obtained by adopting the same procedure as in the determination of median *i.e.*, by observing the cumulative frequency table.)

$$Q_3 = \text{Value of } \frac{3(N+1)}{4} \text{ th term}$$

$$= \text{Value of } \frac{3(24+1)}{4} \text{ th term} = \text{Value of } 18\frac{3}{4} \text{ th term}$$

$$= 18 \text{ th term} + \frac{3}{4}\text{(value of 18 th term} - \text{value of 19th term)}$$

$$= 75 + \frac{3}{4}(82-75) = 75 + \frac{3}{4} \times 7 = 75 + 5.25 = 80.25$$

i.e., $Q_3 = ₹\,80.25$

Quartile deviation (Q. D.) $= \dfrac{Q_3 - Q_1}{2} = \dfrac{₹80.25 - ₹39.25}{2} = ₹\dfrac{41}{2} = ₹\,\mathbf{20.50}$

Coeff. of (Q. D.) $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1} = \dfrac{80.25 - 39.25}{80.25 + 39.25} = \dfrac{41}{119.50} = \mathbf{0.34}$ (approx.)

Interquartile Range and Quartile Deviation

Grouped freq.

Example 5. *Determine Q. D. and Coeff. of Q. D. for the following distribution :*

Weight (kg.)	:	30—34	35—39	40—44	45—49	50—54
No. of boys	:	5	11	26	10	8

Solution :

Table 6.2: Table for determining Q. D. & coeff. of Q. D.

Weight (kg.)	*No. of boys (f)*	*Cumulative frequency (f)*
30—34	5	5
35.5 35—39 38.5	11	16
40—44	26	42
45—49	10	52
50—54	8	60
	$N = 60$	

$$Q_1 = \text{Value of } \frac{N}{4} \text{ th term} = \text{Value of } \frac{60}{4} \text{ th } i.e., \text{ 15 th term.}$$

Clearly the 15th term is in the class 35—39.

Now, $$Q_1 = L + \frac{\frac{N}{4} - f_c}{f} \times I$$

Here, $$L = 35.5, \frac{N}{4} = 15, f = 11, f_c = 5, I = 5$$

$$\therefore \quad Q_1 = 35.5 + \frac{15 - 5}{11} \times 5$$

$$= 35.5 + \frac{50}{11} = 35.5 + 4.55 = 40.05$$

i.e., Value of $Q_1 = 40.05$ kg.

$Q_1 =$ Value of $\frac{3N}{4}$ th term $=$ Value of $\frac{3 \times 60}{4}$ th *i.e.,* 45th term. Clearly the 45 th term included in the class 45 – 49.

Now, $$Q_3 = L + \frac{\frac{3N}{4} - f_c}{f} \times I$$

Here $$L = 45.5, \frac{3N}{4} = 45, f = 1, f_c = 42, I = 5$$

$$\therefore \quad Q_3 = 45.5 + \frac{45 - 42}{10} \times 5$$

$$= 45.5 + 1.5 = 47$$

i.e., $Q_3 =$ **47 kg.**

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{47 - 40.05}{2} \text{ kg}$$

$$= 3.48 \text{ kg (approx.)}$$

$$\mathbf{Q.D. = 3.48 \text{ kg (approx.)}}$$

$$Q.D. = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{47 - 40.05}{47 + 40.05} = \frac{6.95}{87.05} = \mathbf{0.08}.$$

Interquartile Range and Quartile Deviation

- **Advantage:**
    - While only two values of a distribution are involved in the determination of range, 50% of the values of a distribution are involved in the determination of Q.D. Thus, as a measure of dispersion Q.D. is superior to range.
    - Q.D. is not affected by the extreme values since the lowest 25% observations and the highest 25% observations are not taken into account while calculating the Q.D. of a distribution.
    - Q.D. is the only measure which can be used to determine the variation of distribution involving open-end class interval.

Interquartile Range and Quartile Deviation

- **Disadvantage:**
  - Q.D. is based on only 50% of the observations of a distribution. Thus it disregards half of the total observations.
  - It is not amenable for further mathematical treatment.

Mean Deviation

- The arithmetic mean of the absolute deviations of the observations of a distribution from its mean, median or mode is known as mean deviation.
- If a variable x takes n values $x_1, x_2, \ldots, x_n$, then

M.D. = $|x_1 - A| + |x_2 - A| + \ldots + |x_n - A| / n$

Or

M.D. = $\sum|x - A| / n \quad = \sum |d| / n$

- If the frequencies of $x_1, x_2, \ldots, x_n$, are $f_1, f_2, \ldots, f_n$, respectively then,

M.D. = $f_1|x_1 - A| + f_2|x_2 - A| + \ldots + f_n|x_n - A| / n$

Or

M.D. = $\sum f|x - A| / n \quad = \sum f|d| / n$

Mean Deviation

- Coefficient of mean deviation = **M.D. / The avg, from which M.D. is taken**

- Thus,
- Coefficient of mean deviation from mean = **M.D. from mean / Mean**
- Coefficient of mean deviation from median = **M.D. from median / Median**
- Coefficient of mean deviation from median = **M.D. from mode / Mode**

Mean Deviation - Ungrouped Frequency

Example 7. *For the following distribution determine mean deviation (M. D.) from mean and its coefficient.*

x :	10	11	12	13	14
y :	3	12	18	12	3

Solution : First of all we form the following table.

Table 6.4: Computations for M.D. from mean

x	f	fx	$\lvert d\rvert = \lvert x - \bar{x}\rvert$	$f\lvert d\rvert$
10	3	30	2	6
11	12	132	1	12
12	18	216	0	0
13	12	156	1	12
14	3	42	2	6
	$N = 48$	$\Sigma fx = 576$		$\Sigma f\lvert d\rvert = 36$

$$\text{A.M. } (\bar{x}) = \frac{\Sigma fx}{N} = \frac{576}{48} = 12$$

(Usually mean implies arithmetic mean.)

Now, $$\text{M. D. from mean} = \frac{\Sigma f\,|d|}{N} = \frac{36}{48} = 0.75$$

Again, $$\text{coeff. of M. D. from mean} = \frac{\text{M. D. from mean}}{\text{Mean}} = \frac{0.75}{12} = 0.0625$$

Mean Deviation - Grouped Frequency

Example 8. Calculate mean deviation from the mean for the following data. Also find the coefficient of mean deviation.

Class–interval :	*0–4*	*4–8*	*8–12*	*12–16*	*16–20*
Frequency :	*4*	*6*	*8*	*5*	*2*

Table 6.5: Calculations for M.D. from A.M.

Class interval	Mid–value x	f	fx	$\lvert d\rvert = \lvert x - 9.2\rvert$	$f\lvert d\rvert$
0–4	2	4	8	7.2	28.8
4–8	6	6	36	3.2	19.2
8–12	10	8	80	0.8	6.4
12–16	14	5	70	4.8	24.0
16–20	18	2	36	8.8	17.6
		$N = 25$	$\Sigma fx = 230$		$\Sigma f\lvert d\rvert = 96.0$

$$\text{Arithmetic mean } (\bar{x}) = \frac{\Sigma fx}{N} = \frac{230}{25} = 9.2$$

$$\text{Mean deviation from mean } (\text{M.D.}_{\bar{x}}) = \frac{\Sigma f\lvert d\rvert}{N} = \frac{96}{25} = 3.84$$

$$\text{Coefficient of mean deviation} = \frac{M.D._{\bar{x}}}{\bar{x}} = \frac{3.84}{9.2} = 0.42.$$

Mean Deviation

- **Advantage:**
  - It is based on all the observations.
  - It is less affected by extreme values in comparison to standard deviation.
  - Since deviations are taken from average (mean, median and mode), therefore mean deviation is considered to be a good measure for comparing the variability among two or more distributions.

Mean Deviation

- **Disadvantage:**
  - In mean deviation, actual signs of deviations are discarded by taking absolute values of the deviations.
  - Mean deviation from mode is not considered to be a good measure of dispersion.
  - One cannot determine mean deviation for a grouped frequency distribution containing open-end class interval.

Standard Deviation

- The positive square root of the arithmetic mean of the square of the deviations of the values of a variable from its arithmetic mean is called standard deviation of that variable.
- The symbol for S.D.: " σ "
- If variable x takes n values $x_1, x_2, \ldots, x_n$, and if x bar be the arithmetic mean of these values, then

$$\sigma = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n}}$$

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}, \quad \bar{x} = \frac{\Sigma x}{n}$$

Standard Deviation

- In case of the frequency distribution,

Again, in case of the frequency distribution,

x	x_1	x_2	....	x_n
f	f_1	f_2	....	f_n

Standard deviation σ will be :

$$\sigma = \sqrt{\frac{f_1(x_1-\bar{x})^2 + f_2(x_2-\bar{x})^2 + \ldots + f_n(x_n-\bar{x})^2}{N}}, N = \Sigma f$$

i.e.,

$$\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{N}}, \bar{x} = \frac{\Sigma fx}{N} \qquad \ldots(6.11)$$

Instead of the symbol σ we may use the symbol σ_x to clearly signify the standard deviation of the variable x.

Now we shall show that formulae (6.10) and (6.11) are same as the formulae 6.12 (or 6.13) and 6.14 (or 6.15) respectively :

(*i*)

$$\sigma = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} \qquad \ldots(6.12)$$

- Coefficient of S.D. = **σ/xbar** where **σ** = S.D. and X bar - Arithmetic mean

Standard Deviation

Example 9. The following data represent the number of cars entering a gas station between 10 a.m. and 11 a.m. in a city for repairs during the last 8 days of a month.

7, 8, 6, 8, 9, 7, 5, 6.

Calculate the standard deviation for these data.

Solution : Clearly the above data relate to a sample and as such the formula for sample standard deviation should be applied.

Here

$$\bar{x} = \frac{7+8+6+8+9+7+5+6}{8} = \frac{56}{8} = 7$$

Since both the observations and their mean are integers, hence we may easily estimate standard eviation for these data by using the following definitional formula :

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-\cancel{1}}}.$$

Let us first of all calculate the values of $(x-\bar{x})^2$.

Table 6.6: Calculations for S. D.

x	$\bar{x}$	$(x-\bar{x})^2$
7	7	0
8	7	1
6	7	1
8	7	1
9	7	4
7	7	0
5	7	4
6	7	1
		$\Sigma(x-\bar{x})^2 = 12$

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-\cancel{1}}}$$

$$= \sqrt{\frac{12}{8-1}} = \sqrt{\frac{12}{7}} = 1.71.$$

Standard Deviation

Example 10. The following frequency distribution gives the height (in inches) of 100 students selected at random from a college having 3000 students.

Class interval :	60—62	62—64	64—66	66—68	68—70	70—72
No. of students :	5	18	42	20	8	7

Calculate standard deviation.

Solution : We shall use the following formula for calculating standard deviation :

$$s = \sqrt{\frac{N\Sigma fx^2 - (\Sigma fx)^2}{N(N-1)}}.$$

In order to calculate Σfx and Σfx^2 we form the following table :

Table 6.7: Calculations for S. D.

Class interval	*Mid-value* x	*Frequency* f	fx	fx^2
60—62	61	5	305	18605
62—64	63	18	1134	71442
64—66	65	42	2730	177450
66—68	67	20	1340	89780
68—70	69	8	552	38088
70—72	71	7	497	35287
Total		$N = \Sigma f = 100$	$\Sigma fx = 6558$	$\Sigma fx^2 = 430652$

Now standard deviation $s = \sqrt{\frac{N\ \Sigma fx^2 - (\Sigma fx)^2}{N(N-1)}} = \sqrt{\frac{100 \times 430652 - (6558)^2}{100 \times 99}}$

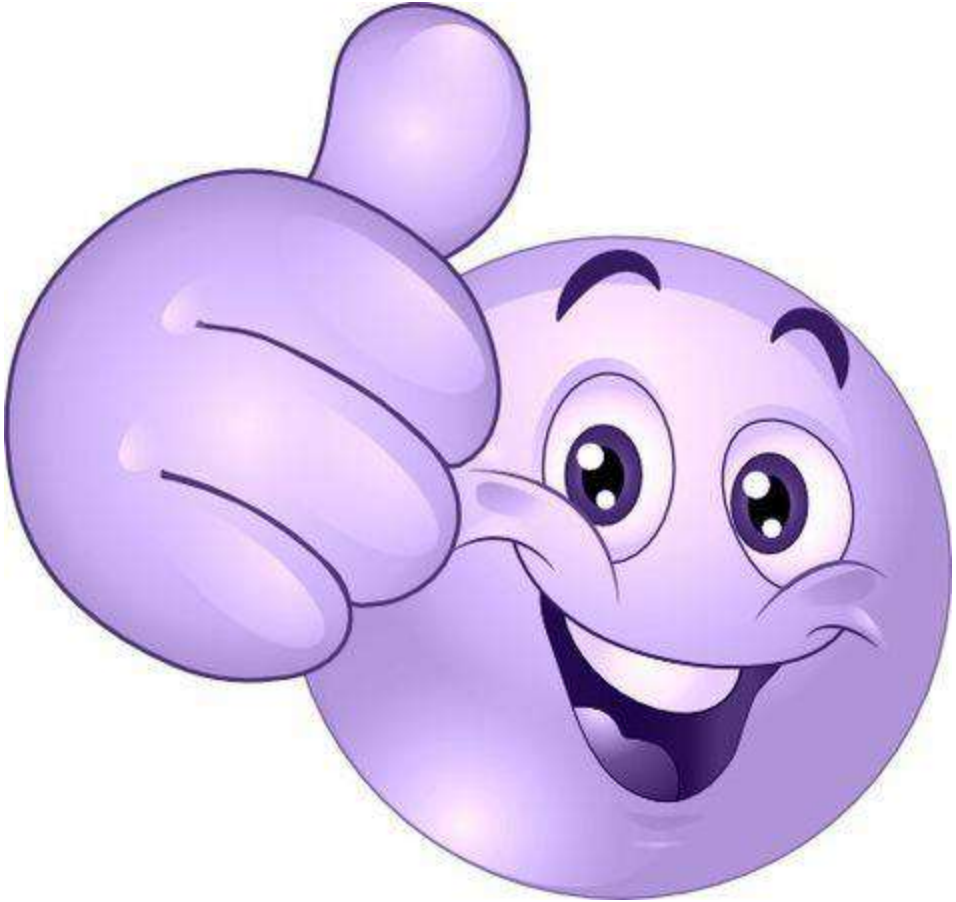
$$= \sqrt{\frac{43065200 - 43007364}{9900}} = \sqrt{\frac{57836}{9900}} = \sqrt{5.842} = 2.42.$$

Standard Deviation

- **Advantage:**
  - Considered to be best measure among all the measures.
  - It is based on all the observations.
  - Based on sampling and correlation analysis.
  - Formula of S.D. is used for further mathematical treatment.
  - Widely used technique of dispersion.
- **Disadvantage:**
  - Difficult to calculate.
  - More affected by extreme values.

Relationships among Q.D., M.D., S.D.

- Q.D. = $\frac{5}{6}$ M.D. = $\frac{2}{3}$ S.D.
- 6 Q.D. = 5 M.D. = 4 S.D.



THANK

YOU...