Measure of dispersion

Dispersion

The extent (limit) to which the values are spread out from an average is called dispersion.

Measure of dispersion

Any formula used to measure the dispersion is called a Measure of dispersion.

Types of Measure of dispersion

There are two main types of Measure of dispersion.

1) Absolute Measure of dispersion
2) Relative Measure of dispersion

Absolute Measure of dispersion

It is the dispersion which measures the variation present among the values in terms of the same unit as the unit of the data.

Types

i) Range ii) Quartile deviation
iii) Mean deviation
iv) Standard deviation
v) Variance

Relative Measure of dispersion

It is the dispersion which measures the variation present in the data relative to their average. It is expressed in the form of ratios, Cofficient or percentage. It is independent of the unit of the data.

Types

i) Coefficient of Range
ii) Coefficient of Quartile deviation
iii) Coefficient of Mean deviation
iv) Coefficient of Standard deviation
v) Coefficient of Variance or variation

Range

Un-group data

The difference between the largest value and the smallest value in the data is called Range. It is denoted by R

R=largest value-smallest value where X_0=smallest value

$R=X_m-X_0$ X_m=largest value

Discrete frequency distribution

The difference between the largest value of the variable X and smallest value of the variable X it is called Range. It is denoted by R

R=largest value-smallest value where X_0=smallest value of the variable X

$R=X_m-X_0$ X_m= largest value of the variable X

Continuous data or Grouped data

The difference between the upper class boundary of the highest class and the lower class boundary of the lowest class it is called Range. It is denoted by R

R=largest value-smallest value where X_0=lower class boundary

$R=X_m-X_0$ X_m= upper class boundary

Coefficient of Range

It is relative measure of dispersion and is based on the value of range. It is also called range coefficient of dispersion. It is defined as

$$Coefficient\ of\ range = \frac{X_m - X_0}{X_m + X_0}$$

The range X_m-X_0 is standardized by the total X_m+X_0

Advantages or merits of Range

1) It easy to calculate
2) It is suitable if the data is homogenous
3) It is useful in small sample inquires
4) It is easy to interpret

Disadvantages or demerits of Range

1) It is highly rough measure of dispersion
2) It gives no idea between the two extreme values
3) It is not based on all the values
4) It is not capable mathematical treatment

Example: 4.1: The marks obtained by 9 students are given below.
45, 32, 37,46,39,36,41,48,36.

Solution: X_0= Lowest marks =32
X_m=highest marks = 48

Range=X_m-X_0=48-32=16

$$Coefficient\ of\ range = \frac{X_m - X_0}{X_m + X_0} = \frac{48-32}{48+32} = \frac{16}{80} = 0.2$$

Example 4.2: Find the range and coefficient of range from the following discrete frequency distribution.

X	10	11	12	13	14	15	16	17	18	19	20	21
f	9	36	75	105	116	107	88	66	45	30	18	5

Solution:

X_0= Smallest value of the variable X =10
X_m= Smallest value of the variable X = 21

Range=X_m-X_0=21-10=11

$$Coefficient\ of\ range = \frac{X_m - X_0}{X_m + X_0} = \frac{21-10}{21+10} = \frac{11}{31} = 0.355$$

Example 4.3: Find the range and coefficient of range from the following frequency distribution.

Classes	5-9	10-14	15-19	20-24	25-29	30-34
f	9	36	75	105	116	107

Solution:

C.b	4.5-9.5	9.5-14.5	14.5-19.5	19.5-24.5	24.5-29.5	29.5-34.5
f	9	36	75	105	116	107

*C.b=class boundary

X_0= lowest class boundary of the highest class=4.5
X_m= upper class boundary of the highest class = 34.5

Range=X_m-X_0=34.5-4.5=30

$$Coefficient\ of\ range = \frac{X_m - X_0}{X_m + X_0} = \frac{34.5 - 4.5}{34.5 + 4.5} = \frac{30}{39} = 0.769$$

Quartile deviation

The half of the difference between the upper quartile and lower quartile is called quartile deviation. It is also called semi inter quartile range. It is denoted by Q.D or S.I.Q.R

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

Q_1=lower quartile

Q_3=upper quartile

Coefficient of Quartile deviation

The relative measure of quartile deviation is called coefficient of quartile deviation or quartile coefficient of dispersion. It is defined as Ratio between Q_3-Q_1 and Q_3+Q_1

$$Coefficien\ of\ Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Which is a pure number and is used for comparing the variation in two or more sets of data

Advantages or merits of Q.D

1) It is easy to calculate
2) It is simple to understands
3) It is not affected by extreme values
4) It is superior to range
5) It is useful badly skewed distributions
6) I is not affected by the dispersion of the individual values of the variable
7) It is specially useful for measuring variation in case of open end distribution

Disadvantages or demerits of Q.D

1) It is not based on all the values
2) Q.D is same for all sets of values having same quartiles
3) It is amenable mathematical treatment
4) The Q.D is not widely used as other measure of dispersion
5) It is ignore first 25% and last 25% of the observations in the data
6) Its value is much affected by sampling fluctuations

Because of these limitations quartile deviation is not a useful measure in statistical inference.

Exampe.4.4: Following are the marks obtained by 20 students, Calculate lower quartile Q_1 and upper quartile Q_3 also Q.D and coefficient of Quartile. (Un-group data)
38,60,41,40,29,40,51,56,40,56,39,40,54,40,53,54,37,53,45,50.

Solution: First we arrange the data in ascending order
Arrange
29,37,38,39,40,40,40,40,40,41,45,50,51,53,53,54,54,56,56,60

Q_1=The value of $(\frac{n+1}{4})th$ item = (21/4) th=5.25th value

=5th value+0.25(6th value-5th value) =40+0.25(40-40) = 41+0.25(0) =40+0=40Ans

Q_3=The value of $3(\frac{n+1}{4})th$ item =3 (21/4)th=3(5.25)th value=15.75th value

=15th value+0.75(16th value-15thvalue)
=53+0.75(5453) =53+0.75(1)=53+0.75=53.75Ans

$$Q.D = \frac{Q_3 - Q_1}{2} = \frac{53.75 - 40}{2} = 6.875$$

$$Coefficien\ of\ Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{53.75 - 40}{53.75 + 40} = \frac{13.75}{93.75} = 0.147$$

Example 4.5: Find the range and coefficient of range from the following discrete frequency distribution.

X	10	11	12	13	14	15	16	17	18	19	20	21
f	9	36	75	105	116	107	88	66	45	30	18	5

Solution:

X	10	11	12	13	14	15	16	17	18	19	20	21
f	9	36	75	105	116	107	88	66	45	30	18	5
c.f	9	45	120	225	341	448	536	602	647	677	695	700

*Q_1 and Q_3 *C.f=Cumulative frequency

$Q_1=(\frac{n+1}{4})th$ item = (701/4) th=175.25th value=13

$Q_3=3(\frac{n+1}{4})th$ item =3 (701/4) th=3(175.25)th value=525.75th value=16

=15th value+0.75(16th value-15thvalue)

=53+0.75(5453) =53+0.75(1)=53+0.75=53.75Ans

$$Q.D=\frac{Q_3-Q_1}{2}=\frac{16-13}{2}=1.5$$

$$Coefficien\ of\ Q.D=\frac{Q_3-Q_1}{Q_3+Q_1}=\frac{16-13}{16+13}=\frac{3}{29}=0.103$$

Example 4.6: Find the semi inters quartile range and coefficient of Q.D from the following frequency distribution.

Classes	5-9	10-14	15-19	20-24	25-29	30-34
f	9	36	75	105	116	107

Solution:

C.b	4.5-9.5	9.5-14.5	14.5-19.5	19.5-24.5	24.5-29.5	29.5-34.5
f	9	36	75	105	116	107
c.f	9	45	120	225	341	448

*C.b=class boundary *C.f=Cumulative frequency

$$Q_1=l+\frac{h}{f}\left(\frac{n}{4}-c\right)$$ Lower quartile $$(\frac{448}{4})th=112th$$

$$Q_1=l+\frac{h}{f}\left(\frac{n}{4}-c\right)=14.5+\frac{5}{75}(112-45)=18.97$$

$$Q_3=l+\frac{h}{f}\left(\frac{3n}{4}-c\right)$$ $$(\frac{3*448}{4})th=(3*112)th=336th$$

$$=24.5+\frac{5}{116}(336-225)=29.28$$

$$Q.D=\frac{Q_3-Q_1}{2}=\frac{29.28-18.97}{2}=5.157$$

$$Coefficien\ of\ Q.D=\frac{Q_3-Q_1}{Q_3+Q_1}=\frac{29.28-18.97}{29.28+18.97}=\frac{10.314}{48.25}=0.214$$

Mean deviation

It is defined as the mean of the deviations of the values taken from their averages (mean, median, and mode) without algebraic sign. It is denoted by M.D

It is given as **Un-group data**

$$M.D = \frac{\sum |X_i - Mean|}{n}$$ Mean deviation from mean

$$M.D = \frac{\sum |X_i - Median|}{n}$$ Mean deviation from median

$$M.D = \frac{\sum |X_i - Mode|}{n}$$ Mean deviation from mode

Group data

$$M.D = \frac{\sum f|X_i - Mean|}{\sum f}$$ Mean deviation from mean

$$M.D = \frac{\sum f|X_i - Median|}{\sum f}$$ Mean deviation from median

$$M.D = \frac{\sum f|X_i - Mode|}{\sum f}$$ Mean deviation from mode

Note

Without algebraic sign mean absolute value of the deviation i.e. $|X_i - Mean|$ or $|X_i - Median|$.The absolute value of the positive number in the itself where as the absolute value of the negative number is the number without its minus sign i.e. $|X_i| = X_i$, $|-X_i| = X_i$

Advantages or merits of M.D

1) There is no confusion in its definition
2) It is based on all the values
3) It is easy to calculate
4) It is easy to understand
5) It gives weight to the observations according to their size
6) It is less affected by extreme values

Disadvantages or demerits of M.D

1) It has a mathematical flaw of ignoring signs
2) It has no further mathematical treatment
3) It is affected by extreme values
4) It is not generally used in social sciences

Note

We take the deviation about median is a more appropriate average to compute mean deviation because the sum of absolute deviations of items from median minimum. But in practice mean is more commonly used average to calculate mean deviation and this is the reason of calling the measure as mean deviation.

Coefficient of Mean deviation

Ratio between mean deviation and average used in its calculation is called coefficient of mean deviation. It is denoted by

$$\textit{Coefficient of } M.D \textit{ or mean coeffcient of dispersion} = \frac{M.D_{\bar{X}}}{\bar{X}}$$

$$\textit{Coefficient of } M.D \textit{ or median coeffcient of dispersion} = \frac{M.D_{\tilde{X}}}{\tilde{X}}$$

$$\textit{Coefficient of } M.D \textit{ or mode coeffcient of dispersion} = \frac{M.D_{\hat{X}}}{\hat{X}}$$

Example 4.7: Calculate the Mean Deviation and Co-efficient of M.D from i) Mean ii) Median iii) Mode. From the following data
32, 45,37,46,39,36,41,48 and 36

Solution: First we arrange the observations 32, 36, 36,37,39,41,45,46,48

Mean= $\bar{x} = \frac{\sum x}{n}$ =40

Median= 39

Mode=36

X	$\lvert X - Mean \rvert$	$\lvert X - Median \rvert$	$\lvert X - Mode \rvert$
32	8	7	4
36	4	3	0
36	4	3	0
37	3	2	1
39	1	0	3
41	1	2	5
45	5	6	9
46	6	7	10
48	8	9	12
Total	$\sum \lvert X - Mean \rvert = 40$	$\sum \lvert X - Median \rvert = 39$	$\sum \lvert X - Mode \rvert = 44$

$M.D = \frac{\sum \lvert X_i - Mean \rvert}{n}$ =4.44 Mean deviation from mean

$M.D = \frac{\sum \lvert X_i - Median \rvert}{n}$ =4.33 Mean deviation from median

$M.D = \frac{\sum \lvert X_i - Mode \rvert}{n}$ =4.8 Mean deviation from mode

$$\textit{Coefficient of } M.D_{\bar{x}} \textit{ or mean coeffcient of dispersion} = \frac{M.D_{\bar{X}}}{\bar{X}} = 0.111$$

$$\textit{Coefficient of } M.D_{\tilde{x}} \textit{ or median coeffcient of dispersion} = \frac{M.D_{\tilde{X}}}{\tilde{X}} = 0.1111$$

$$\textit{Coefficient of } M.D_{\hat{x}} \textit{ or mode coeffcient of dispersion} = \frac{M.D_{\hat{X}}}{\hat{X}} = 0.136$$

Example 4.8: Calculate the Mean Deviation and Co-efficient of M.D from i) Mean ii) Median iii) Mode. From the following frequency distribution.

Classes	5-9	10-14	15-19	20-24	25-29	30-34
f	9	36	75	105	116	107

Solution:

C.b	4.5-9.5	9.5-14.5	14.5-19.5	19.5-24.5	24.5-29.5	29.5-34.5
f	9	36	75	$105=f_1$	$116=f_m$	$107=f_2$
c.f	9	45	120	225	341	448

*C.b=class boundary *C.f=Cumulative frequency

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{10636}{448} = 23.74$$

$$Median = \tilde{X} = l + \frac{h}{f}\left(\frac{n}{2} - c\right) = 19.5 + \frac{5}{105}(224 - 120) = 24.45 \qquad (\frac{448}{2})th = 224th$$

$$\hat{X} = l + \frac{f_m - f_1}{f_m - f_1 + f_m - f_2} \times h = 24.5 + \frac{116 - 107}{116 - 107 + 116 - 105} \times 5 = 26.75$$

x	f	$f\lvert X - Mean\rvert$	$f\lvert X - Median\rvert$	$f\lvert X - Mode\rvert$
7	9	150.66	157.05	177.75
12	36	422.64	448.2	531
17	75	505.5	558.75	731.25
22	105	182.7	257.25	498.75
27	116	378.16	295.8	29
32	107	883.82	807.85	561.75
	$\sum f$ 448	$\sum f\lvert X - Mean\rvert = 2523.48$	$\sum f\lvert X - Median\rvert = 2524.9$	$\sum f\lvert X - Mode\rvert = 2529.5$

$$M.D = \frac{\sum f\lvert X_i - Mean\rvert}{\sum f}$$ Mean deviation from mean=5.633

$$M.D = \frac{\sum f\lvert X_i - Median\rvert}{\sum f}$$ Mean deviation from median=5.634

$$M.D = \frac{\sum f\lvert X_i - Mode\rvert}{\sum f}$$ Mean deviation from mode=5.646

$$Coefficient\ of\ M.D_{\bar{x}}\ or\ mean\ coeffcient\ of\ dispersion = \frac{M.D_{\bar{X}}}{\bar{X}} = 0.237$$

$$Coefficient\ of\ M.D_{\tilde{x}}\ or\ median\ coeffcient\ of\ dispersion = \frac{M.D_{\tilde{X}}}{\tilde{X}} = 0.230$$

$$Coefficient\ of\ M.D_{\hat{x}}\ or\ \mathrm{mod}\ e\ coeffcient\ of\ dispersion = \frac{M.D_{\hat{X}}}{\hat{X}} = 0.211$$

Standard deviation

It is defined as the positive square root of the mean of the squared deviations of the values from their mean. It is denoted by S or δ

Calculation methods

Simple method

$$S=\sqrt{\frac{\sum\left(X-\overline{X}\right)^2}{n}}=\sqrt{\left(\frac{\sum x^2}{n}-\left(\frac{\sum x}{n}\right)^2\right)}$$

Or

$$S=\sqrt{\frac{\sum f\left(X-\overline{X}\right)^2}{\sum f}}=\sqrt{\left(\frac{\sum fx^2}{\sum f}-\left(\frac{\sum fx}{\sum f}\right)^2\right)}$$

Shot-cut method

$$S=\sqrt{\left(\frac{\sum D^2}{n}-\left(\frac{\sum D}{n}\right)^2\right)}$$ Where $D=x-A$ and A= arbitrary value

Or

$$S=\sqrt{\left(\frac{\sum fD^2}{\sum f}-\left(\frac{\sum fD}{\sum f}\right)^2\right)}$$

Coding method

$$S=h\sqrt{\left(\frac{\sum u^2}{n}-\left(\frac{\sum u}{n}\right)^2\right)}$$ Where $u=\frac{x-A}{h}$ And h= class interval

Or

$$S=h\sqrt{\left(\frac{\sum fu^2}{\sum f}-\left(\frac{\sum fu}{\sum f}\right)^2\right)}$$

Advantages or merits of S.D

1) It is rigidly defined
2) It is based on all the observations
3) It is capable of mathematical treatment
4) It is stable in repeated sampling experiments
5) It will be large if the observations are distant from the mean and small if they are close to mean
6) It is possible to calculate combined standard deviation of two or more groups which is not possible with any other measure of dispersion

Disadvantages or demerits of S.D

1) It is affected by extreme values
2) It is not an unbiased estimate of population standard deviation

Variance

It is defined as the arithmetic mean of the squared deviation taken from their mean. It is denoted by $S^2 \; or \; \sigma^2$

It is given as

Simple method or direct method

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

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

Short-cut method

$$S^2 = \frac{\sum D^2}{n} - \left(\frac{\sum D}{n}\right)^2 \qquad S^2 = \frac{\sum fD^2}{\sum f} - \left(\frac{\sum fD}{\sum f}\right)^2$$

Coding method

$$S^2 = h^2\left(\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right)$$

Properties of variance

1) Variance of constant is equal to zero.

   Var (a) =0

   Proof:

   Let by definition of variance

   $$Var(X) = \frac{1}{N}\sum (X-\mu)^2$$

   $Let \quad X = a$

   $$\sum X = Na$$

   $$\frac{\sum X}{N} = \frac{Na}{N}$$

   $$\mu = a$$

   Now

   $$Var(a) = \frac{1}{N}\left(a - a^2\right) = \frac{1}{N}(0)^2 = \frac{1}{N}(0) = 0$$ *hence proved that*

2) The Variance of variable X is independent of origin. It remains unchanged when a Constant is added or Subtracted from each value of the variable X.

$$Var(X \pm a) = \frac{1}{N}\sum (X-\mu)^2 = Var(X)$$

Proof:

$$Var(X) = \frac{1}{N}\sum (X_i - \mu)^2$$

Let $X_i = X_i + a$ *or* $X_i = X_i - a$

$$\sum X_i = \sum X_i + Na$$

$$\frac{\sum X}{N} = \sum X_i + \frac{Na}{N} = \mu + a$$

Now

$$Var(X_i + a) = \frac{1}{N}\sum \left((X_i + a) - (\mu + a)\right)^2 = \frac{1}{N}\sum \left(X_i + a - \mu - a\right)^2 = \frac{1}{N}\sum \left(X_i - \mu\right)^2 = Var(X_i)$$

hence proved that

3) When all the values of variable X are multiplied or divided by a constant. The Variance of these values is multiplied or divided by square of constant.

$$Var(aX) = a^2 Var(X)$$

$$Var(\frac{1}{a}X) = \frac{1}{a^2}Var(X)$$

Proof

$$Var(X) = \frac{1}{N}\sum (X_i - \mu)^2$$

Let $X_i = aX_i$ *or* $X_i = \frac{1}{a}X_i$

$$\sum X_i = a\sum X_i$$

$$\frac{\sum X}{N} = a\sum X_i / N \qquad \mu = a\mu$$

Now

$$Var(aX_i) = \frac{1}{N}\sum \left(aX_i - a\mu\right)^2 = \frac{1}{N}a^2\sum \left(X_i - \mu\right)^2 = a^2\frac{1}{N}\sum \left(X_i - \mu\right)^2 = a^2 Var(X_i)$$

hence proved that

4) If two sets of data consisting of n_1, n_2 values have means $\overline{X}_1, \overline{X}_2$ and variances $S_1^{\,2}, S_2^{\,2}$ respectively then the combined variance of n_1, n_2 values is

$$S^2 = \frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}\left(\overline{X}_1 - \overline{X}_2\right)^2$$

Proof:

By definition of

$$S^2 = \frac{1}{n_1 + n_2}\sum_{i=1}^{n_1+n_2}\left(X_i - \overline{X}\right)^2$$

$$S^2 = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=n_1+1}^{n_1+n_2}\left(X_i - \overline{X}\right)^2\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(X_i \pm \overline{X}_1 - \overline{X})^2 + \sum_{i=n_1+1}^{n_1+n_2}\left(X_i \pm \overline{X}_2 - \overline{X}\right)^2\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(X_i - \overline{X}_1 + \overline{X}_1 - \overline{X})^2 + \sum_{i=n_1+1}^{n_1+n_2}\left(X_i - \overline{X}_2 + \overline{X}_2 - \overline{X}\right)^2\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}((X_i - \overline{X}_1) + (\overline{X}_1 - \overline{X}))^2 + \sum_{i=n_1+1}^{n_1+n_2}\left((X_i - \overline{X}_2) + (\overline{X}_2 - \overline{X})\right)^2\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}((X_i - \overline{X}_1)^2 + (\overline{X}_1 - \overline{X})^2 + 2(X_i - \overline{X}_1)(\overline{X}_1 - \overline{X})) + \sum_{i=n_1+1}^{n_1+n_2}\left((X_i - \overline{X}_2)^2 + (\overline{X}_2 - \overline{X})^2 + 2(X_i - \overline{X}_2)(\overline{X}_2 - \overline{X})\right)\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2 + n_1(\overline{X}_1 - \overline{X})^2 + 2(\overline{X}_1 - \overline{X})\sum_{i=1}^{n_1}(X_i - \overline{X})) + \sum_{i=n_1+1}^{n_1+n_2}\left(X_i - \overline{X}_2)^2 + n_2(\overline{X}_2 - \overline{X})^2 + 2(\overline{X}_2 - \overline{X})\sum_{i=n_1+1}^{n_1+n_2}(X_i - \overline{X}\right)\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2 + n_1(\overline{X}_1 - \overline{X})^2 + 2(\overline{X}_1 - \overline{X})(0) + \sum_{i=n_1+1}^{n_1+n_2}(X - \overline{X}_2)^2 + n_2(\overline{X}_2 - \overline{X})^2 + 2(\overline{X}_2 - \overline{X})(0)\right]$$

$$= \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + n_1(\overline{X}_1 - \overline{X})^2 + \sum_{i=n_1+1}^{n_1+n_2}(X - \overline{X}_2)^2 + n_2(\overline{X}_2 - \overline{X})^2\right]$$

$$= \frac{1}{n_1 + n_2}\left[n_1 S^2_1 + n_1(\overline{X}_1 - \overline{X})^2 + n_2 S_2^{\,2} + n_2(\overline{X}_2 - \overline{X})^2\right]$$

$$= \frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2} + \frac{n_1(\overline{X}_1 - \overline{X})^2 + n_2(\overline{X}_2 - \overline{X})^2}{n_1 + n_2} \qquad \text{(A)}$$

As we know that

$$\bar{\bar{X}} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Now

$$\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 = \left(\bar{X}_1 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}\right)^2$$

$$\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 = \left(\frac{(n_1 + n_2)\bar{x}_1 - (n_1\bar{x}_1 + n_2\bar{x}_2)}{n_1 + n_2}\right)^2$$

$$\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 = \left(\frac{\bar{x}_1 n_1 + n_2\bar{x}_1 - n_1\bar{x}_1 - n_2\bar{x}_2}{n_1 + n_2}\right)^2$$

$$\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 = \left(\frac{n_2\bar{x}_1 - n_2\bar{x}_2}{n_1 + n_2}\right)^2 = \frac{n^2{}_2(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2}$$

And

$$\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 = \left(\bar{X}_2 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}\right)^2$$

$$\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 = \left(\frac{n_1\bar{x}_2 - n_1\bar{x}_1}{n_1 + n_2}\right)^2 = \frac{n^2{}_1(\bar{x}_2 - \bar{x}_1)^2}{(n_1 + n_2)^2}$$

$$= \frac{n^2{}_1(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2}$$ Put in A

$$S^2 = \frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2} + \frac{n_1\frac{n^2{}_2(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2} + n_2\frac{n^2{}_1(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2}}{n_1 + n_2}$$

$$= \frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2} + \frac{(\frac{n_2(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2} + \frac{n_1(\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2)^2})(n_1 + n_2)}{n_1 + n_2}$$

$$= \frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$$ hence proved

5) The variance of the sum or difference of two independent random variables is equal to the sum of their respective variances

i) Var(x+y) =Var(x) +Var(y)

ii) Var(x+y) =Var(x) +Var(y)

Proof:

i) Var(x+y) =Var(x) +Var(y)

By definition

$$Var(X)=\frac{1}{N}\sum\left(X_i-\mu\right)^2$$

Let $X_i = X + Y_{ii}$

$$\sum X_i = \sum X_i + \sum Y$$

$$\frac{\sum X}{N} = \frac{\sum X}{N} + \frac{\sum Y}{N} \qquad \mu = \mu_x + \mu_y$$

Now

$$Var(X+Y_i)=\frac{1}{N}\sum\ \left((X+Y)-(\mu_x+\mu_y)\right)^2=\frac{1}{N}\sum\ \left(X+y_i-\mu_x-\mu_y\right)^2=a^2\frac{1}{N}\sum\ \left((X-\mu_x)+(y-\mu_y)\right)^2$$

$$=\frac{1}{N}\sum\left((X-\mu_x)^2+(Y-\mu_Y)^2+2(X-\mu_x)(Y-\mu_y)\right)=\frac{1}{N}\left(\sum\left(X-\mu_x\right)^2+\sum\left(Y-\mu_y\right)^2+2\sum\left(X-\mu_x\right)\left(Y-\mu_y\right)\right)$$

$$=\frac{1}{N}\sum\left(X-\mu_x\right)^2+\frac{1}{N}\sum\left(Y-\mu_y\right)^2+2\frac{1}{N}\sum\left(X-\mu_x\right)\left(Y-\mu_y\right)$$

AS we know that X and Y are independent var *iable than*

$$Cov(X,Y)=\frac{1}{N}\sum\left(X-\mu_x\right)\left(Y-\mu_y\right)\ \text{sin}\ ce\ \text{cov}\ ariance\ of\ independent\ \text{var}\ iable\ is\ equal\ to\ zero$$

therefore

$$=\frac{1}{N}\sum\left(X-\mu_x\right)^2+\frac{1}{N}\sum\left(Y-\mu_y\right)^2+2(0)=\frac{1}{N}\sum\left(X-\mu_x\right)^2+\frac{1}{N}\sum\left(Y-\mu_y\right)^2=\text{var}(X)+\text{var}(y)$$

hence proved that

6) If k subgroups of the data consisting of N_1, N_2… N_k ($\sum N_i = N$) observations have respective means μ_1, $\mu_2, \mu_3, \ldots, \mu_k$ and variances $\sigma^2{}_1, \sigma^2{}_2, \sigma^2{}_3, \ldots, \sigma^2{}_k$,than the variance σ^2 of the combined observations is given as

$$\sigma^2 = \frac{1}{N}\sum N_i\left(\sigma^2{}_i + D^2{}_i\right) \qquad \text{i=1,2,3,…,k}$$

Where $D_i = \mu_i - \mu$ and μ is combined mean\

Proof:

By definition

$$\sum_{i=1}^{N_i}\left(X_i - \mu\right)^2 = \sum_{i=1}^{N}\left[X_i - \mu_i + \mu_i - \mu\right]^2 = \sum_{i=1}^{N}\left[(X_i - \mu_i) + (\mu_i - \mu)\right]^2$$

$$= \sum_{i=1}^{N}\left[(X_i - \mu_i)^2 + (\mu_i - \mu)^2 + 2\left(X_i - \mu\right)\left(\mu_i - \mu\right)\right]^2 ==$$

$$= \sum_{i=1}^{N}\ (X_i - \mu_i)^2 + N_i(\mu_i - \mu)^2 + 2\left(\mu_i - \mu\right)\sum_{i=1}^{N}\ \left(X_i - \mu\right)$$

As we know that $\sum_{i=1}^{N}\ \left(X_i - \mu\right)$=0 and $\sigma^2{}_i = \frac{\sum\left(X_i - \mu\right)^2}{N_i}$

Therefore cross product term vanish and we have

$$= \sum_{i=1}^{N}\ (X_i - \mu_i)^2 + N_i(\mu_i - \mu)^2 + 2\left(\mu_i - \mu\right)(0)$$

$$= \sum_{i=1}^{N}\ (X_i - \mu_i)^2 + N_i(\mu_i - \mu)^2 = N_i\sigma^2{}_i + N_i\ (\mu_i - \mu)^2 = N_i\sigma^2{}_i + N_i D_i{}^2$$

$$N_i\sigma^2{}_i = N_i(\sigma^2{}_i + D_i{}^2)$$

Now $\sum_{i=1}^{N_i}\ N_i\sigma^2{}_i = \sum_{i=1}^{N_i}\ N_i(\sigma^2{}_i + D_i{}^2)$

$$N\sigma^2 = \sum_{i=1}^{N_i}\ N_i(\sigma^2{}_i + D_i{}^2)$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N_i}\ N_i(\sigma^2{}_i + D_i{}^2)$$

Or

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} N_i\left(\sigma^2{}_i + D^2{}_i\right)}$$

Coefficient of Variance or variation

The relative measure of dispersion of variance is called the Coefficient of variation. The coefficient of variation expressed the standard deviation as percentage in terms of arithmetic mean. It is denoted by C.V

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

Uses of coefficient of variation

1) It is used to compare the dispersion of two or more set of data
2) It is also used for the consistence performance. The smallest the coefficient of variation the more consistence is the performance

Example.4.9: Find the mean S.D and Co-efficient of S.D and Co-efficient of variation from following ungroup data by Direct, short-cut and coding method.
5, 10,15,20,25
Solution:

X	X^2	D=X-A	D^2	U= $\frac{X-A}{h}$	U^2
5	25	-10	100	-2	4
10	100	-5	25	-1	1
15	225	0	0	0	0
20	400	5	25	1	1
25	625	10	100	2	4
$\sum x$=75	$\sum x^2$ =1375	$\sum D = 0$	$\sum D^2 = 250$	$\sum U = 0$	$\sum U^2 = 10$

i) $\bar{x} = \frac{\sum x}{n} = \frac{75}{5} = 15$ $\quad S^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{1375}{5} - (\frac{75}{5})^2 = 50$

ii) $\bar{x} = A + \frac{\sum D}{n} = 15 + \frac{0}{5} = 15$ $\quad S^2 = \frac{\sum D^2}{n} - \left(\frac{\sum D}{n}\right)^2 = \frac{250}{5} - (\frac{0}{5})^2 = 50$

iii) $\bar{x} = A + \frac{\sum U}{n} \times h = 15 + \frac{0}{5} \times 5 = 15$

$$S^2 = \frac{\sum U^2}{n} - \left(\frac{\sum U}{n}\right)^2 \times h^2 = \left(\frac{10}{5} - (\frac{0}{5})^2\right) \times 25 = 50$$

$$S.D = \sqrt{Var(x)} = \sqrt{50} = 7.071$$

Co-efficient of S.d

$$Co-efficient\ of\ S.D = \frac{S}{\bar{X}} = \frac{7.071}{15} = 0.4714$$

Co-efficient of variation

$$Co-efficient\ of\ \text{var}iation = \frac{S}{\bar{X}} \times 100 = \frac{7.071}{15} \times 100 = 47.14\%$$

Example.4.10: Find S.D and Co-efficient of S.D and Co-efficient of variation from the following data by 1) Direct ii) Short-cut iii) Coding method

Classes	10-15	15-20	20-25	25-30	30-35
f	2	4	6	8	3

Solution:

x	f	fx	fD	fu	fx^2	fD^2	fu^2
12.5	2	25	-20	-4	312.5	200	8
17.5	4	70	-20	-4	1225	100	4
22.5	6	135	0	0	3037.5	0	0
27.5	8	220	40	8	6050	200	8
32.5	3	97.5	30	6	3168.75	300	12
Total	23	547.5	30	6	13793.75	800	32

$$\bar{x} = \frac{\sum fx}{\sum f} = 23.80$$

$$S^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 = \frac{13793.75}{23} - \left(\frac{547.5}{23}\right)^2 = 599.728 - 566.65 = 33.08$$

$$\bar{x} = A + \frac{\sum fD}{\sum f} = 22 + \frac{30}{23} = 23.80$$

$$S^2 = \frac{\sum fD^2}{\sum f} - \left(\frac{\sum fD}{\sum f}\right)^2 = \frac{800}{23} - \left(\frac{30}{23}\right)^2 = 34.78 - 1.701 = 33.08$$

$$\bar{x} = A + \frac{\sum fU}{\sum f} \times h = 22 + \frac{6}{23} \times 5 = 23.80$$

$$S^2 = h^2\left(\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right) = 5^2\left(\frac{32}{23} - \left(\frac{6}{23}\right)^2\right) = 25(1.39 - 0.068) = 33.05$$

Co-efficient of S.d

$$Co - efficient\ of\ S.D = \frac{S}{\bar{X}} = \frac{5.7515.}{23.80} = 0.2417$$

Co-efficient of variation

$$Co - efficient\ of\ \text{var}iation = \frac{S}{\bar{X}} \times 100 = \frac{5.7515}{23.80} \times 100 = 24.17\%$$

Example: 4.11: The following table gives mean scores and S.D of two groups of students.

$n_1 = 50 \qquad \bar{x}_1 = 59.5 \qquad n_2 = 20 \qquad \bar{x}_2 = 45.7 \qquad s_1 = 10.73 \qquad s_2 = 7.04$

Find the mean and standard deviation for the combined group of 50 students.

Solution: $\bar{X}_c = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \dfrac{50(59.5) + 20(45.7)}{50 + 20} = \dfrac{3889}{70} = 55.557$

$$S_c^{\,2} = \frac{n_1 S_1^{\,2} + n_2 S_2^{\,2}}{n_1 + n_2} + \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^2 = \frac{50(115.1329) + 20(49.5616)}{50 + 20} + \frac{20(50)}{(50 + 20)^2}(59.5 - 45.7)^2 = 96.3982 + 0.2041(190.44) = 135.267$$

$S_c = \sqrt{135.267} = 11.63$

Example: 4.12: A distribution consists of 3 groups with frequencies 40,25 and 35 having means 65,66 and 72 and S.D 10.5,9.2 and 8.3 respectively. Find the combined S.d.

Solution: $\bar{X}_c = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{n_1 + n_2 + n_3} = \dfrac{40(65) + 25(66) + 35(72)}{40 + 25 + 35} = \dfrac{6770}{100} = 67.7$

n_i	$\bar{x}_i$	s_i	$s_i^{\,2}$	$D_i = (\bar{x}_i - \bar{x}_c)$	$D_i^{\,2}$	$S_i^{\,2} + D_i^{\,2}$	$n_i(S_i^{\,2} + D_i^{\,2})$
40	65	10.5	110.25	-2.7	7.27	117.54	2701.60
25	66	9.2	84.64	-1.7	2.89	87.53	2188.25
35	72	8.3	68.89	4.3	18.49	87.38	3058.30
100							9948.15

$$S_c^{\,2} = \frac{1}{n}\sum_{i=1}^{N_i} n_i(s^2{}_i + D_i^{\,2}) = \frac{9948.15}{100} = 99.4815$$

Or

$$S_c = \sqrt{\frac{1}{n}\sum_{i=1}^{N} n_i\left(s^2{}_i + D^2{}_i\right)} = \sqrt{99.4815 = 9.97}$$

Example: 13: Two students obtained the following marks in 10 papers of M.sc examination.

In which student was there greater i) Absolute dispersion ii) Relative dispersion

X	39	45	39	40	46	52	49	41	45	44	$\sum X$	440
X^2	1521	2025	1521	1600	2116	2704	2401	1681	2025	1936	$\sum X^2$	19530
Y	45	48	62	44	38	48	42	54	49	50	$\sum Y$	480
Y^2	2025	2304	3844	1936	1444	2304	1746	2916	2401	2500	$\sum Y^2$	23438

Solution:

i) Absolute dispersion for this we calculate S.D

$$S_X^{\,2} = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{19530}{10} - (\frac{440}{10})^2 = 17 \qquad \bar{X} = \frac{\sum X}{10} = \frac{440}{10} = 44$$

$S_x = \sqrt{17} = 4.12$

$$S_y^2 = \frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2 = \frac{23438}{10} - (\frac{480}{10})^2 = 39.8 \qquad \bar{Y} = \frac{\sum Y}{10} = \frac{480}{10} = 48$$

$$S_y = \sqrt{39.8} = 6.31$$

Since S.D of 2nd student is greater than the first student so 2nd student is grater absolute dispersion.

ii) Relative dispersion for this we calculate C. V

$$Co - efficient\ of\ \text{var}iation = \frac{S_x}{\bar{X}} \times 100 = \frac{4.12}{44} \times 100 = 9.36\%$$

$$Co - efficient\ of\ \text{var}iation = \frac{S_y}{\bar{X}} \times 100 = \frac{6.31}{48} \times 100 = 13.15\%$$

Since C.V of 2nd student is greater than the first student so 1st student is grater relative dispersion or more consistent performance.

Example: 14: A student calculated the values of mean and S.D of 25 observations as 20 and 4 respectively. It was later discovered at the time of checking that he had copied down two values as 7 and 18 while the correct values were 13 and 17.Find the correct values of

i) Mean ii) S.D iii) C.V

Solution: here n=25 $\bar{X} = 20$ $\bar{X} = 20$ $S = 4$ $S^2 = 16$

$$\bar{X} = \frac{\sum X}{n} \qquad n\bar{X} = \sum X \qquad 500 = \sum X(incorrected)$$

$$\sum X(corrected) = \sum X(incorrected) - wrong\ values + corrected\ values$$

$$\sum X(corrected) = 500 - (7+18) + (13+17) = 505$$

$$\bar{X}(corrected) = \frac{\sum X(incorrected)}{n} = \frac{505}{25} = 20.2$$

$$S_X^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 \qquad S_X^2 = \frac{\sum x^2}{25} - \left(\frac{500}{25}\right)^2 \qquad S_X^2 = \frac{\sum x^2}{25} - (20)^2$$

$$16 + 400 = \frac{\sum x^2}{25} \qquad \frac{\sum x^2}{25} = 416 \quad \sum X^2(incorrected) = 416 \times 25 = 10400$$

$$\sum X^2(corrected) = \sum X^2(incorrected) - (wrong\ values)^2 + (corrected\ values)^2$$

$$\sum X^2(corrected) = 10400 - \left[7^2 + 18^2\right] + \left[13^2 + 17^2\right] = 10485$$

$$S_X^2(corrected) = \frac{\sum x^2(corrected)}{n} - \left(\frac{\sum x(corrected)}{n}\right)^2 = \frac{10485}{25} - \left(\frac{505}{25}\right)^2 = 11.36$$

$$S.D = \sqrt{11.36} = 3.37$$

$$C.V(corrected) = \frac{S(corrected)}{\bar{X}(corrected)} \times 100 = \frac{3.37}{20.2} \times 100 = 16.69\%$$

Example: 15: The weight of 20-students in a college is given in the following data.
138,164,150,132,144,125,149,157,146,158,140,147,136,148,152,144,168,126,138,176
Find the percentage of weights falling within the limits
i) $\bar{X} \pm S$ ii) $\bar{X} \pm 2S$ iii) $\bar{X} \pm 3S$ iv) $\bar{X} \pm 0.6745S$
Solution: First of all we arrange the observations

X	125	126	132	136	138	138	140	144	144	
X^2	15625	15876	17424	18496	19044	19044	19600	20736	20736	
X	146	147	148	149	150	152	157	158	164	168	176
X^2	21609	21904	21904	22201	22500	23104	24649	24964	26896	28224	30976

$$\bar{X} = \frac{\sum X}{n} = \frac{2938}{20} = 146.9 \qquad S_X{}^2 = \frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2 = \frac{434924}{20} - (146.9)^2 = 166.59$$

$$S.D = \sqrt{166.59} = 12.91$$

i) The interval $\bar{X} \pm S$

$\bar{X} - S = 146.9 - 12.91 = 133.99$ And $\bar{X} + S = 146.9 + 12.91 = 159.81$

This interval 133.99--------159.81 contains 14 values

Then the percentage of interval is $= \frac{14}{20} \times 100 == 70\%$ students

ii) The interval $\bar{X} \pm 2S$

$\bar{X} - 2S = 146.9 - 25.82 = 121.08$ And $\bar{X} + S = 146.9 + 25.82 = 172.72$

This interval 121.08--------172.72 contains 19 values

Then the percentage of interval is $= \frac{14}{20} \times 100 == 70\%$ students

iii) The interval $\bar{X} \pm 3S$

$\bar{X} - 3S = 146.9 - 38.73 = 108.17$ And $\bar{X} + 3S = 146.9 + 38.73 = 185.63$

This interval 108.17--------185.63 contains 20 values

Then the percentage of interval is $= \frac{20}{20} \times 100 == 100\%$ students

iv) The interval $\bar{X} \pm 0.6745S$

$\bar{X} - 0.6745S = 146.9 - 8.71 = 138.19$ And $\bar{X} + 0.6745S = 146.9 + 8.71 = 155.61$

This interval 138.19--------155.61 contains 10 values

Then the percentage of interval is $= \frac{10}{20} \times 100 == 50\%$ students

Moments

The moments are the arithmetic mean of the power to which the deviations are raised before averaging them.

Moments about mean

The moments about the mean are the arithmetic mean of the power of the deviation (about the arithmetic mean) is raised before averaging them. They are denoted by m_1, m_2,m_3 and m_4

For un-group data		**group data**
$m_1 = \frac{\sum (X_i - \bar{X})}{n}$	1st moment about mean	$m_1 = \frac{\sum f(X_i - \bar{X})}{\sum f}$
$m_2 = \frac{\sum (X_i - \bar{X})^2}{n}$	2nd moment about mean	$m_2 = \frac{\sum f(X_i - \bar{X})^2}{\sum f}$
$m_3 = \frac{\sum (X_i - \bar{X})^3}{n}$	3rd moment about mean	$m_3 = \frac{\sum f(X_i - \bar{X})^3}{\sum f}$
$m_4 = \frac{\sum (X_i - \bar{X})^4}{n}$	4th moment about mean	$m_4 = \frac{\sum f(X_i - \bar{X})^4}{\sum f}$

Moment about any value A or Assumed mean or Provisional mean

The moments about any value A are denoted by m_1', m_2', m_3', m_4' where (read as prime)

For un-group data		**group data**
$m_1' = \frac{\sum (X_i - A)}{n} = \frac{\sum D}{n}$	1st moment about "A"	$m_1' = \frac{\sum f(X_i - A)}{\sum f} = \frac{\sum fD}{\sum f}$
$m_2' = \frac{\sum (X_i - A)^2}{n} = \frac{\sum D^2}{n}$	2nd moment about "A"	$m_2' = \frac{\sum f(X_i - A)^2}{\sum f} = \frac{\sum fD^2}{\sum f}$
$m_3' = \frac{\sum (X_i - A)^3}{n} = \frac{\sum D^3}{n}$	3rd moment about "A"	$m_3' = \frac{\sum f(X_i - A)^3}{\sum f} = \frac{\sum fD^3}{\sum f}$
$m_4' = \frac{\sum (X_i - A)^4}{n} = \frac{\sum D^4}{n}$	4th moment about "A"	$m_4' = \frac{\sum f(X_i - A)^4}{\sum f} = \frac{\sum fD^4}{\sum f}$

For continuous data

$$m_1' = \frac{\sum fu}{\sum f} \times h$$

$$m_2' = \frac{\sum fu^2}{\sum f} \times h^2$$

$$m_3' = \frac{\sum fu^3}{\sum f} \times h^3$$

$$m_4' = \frac{\sum fu^4}{\sum f} \times h^4$$

Now we find the moment about mean by using the relationship between mean and moment about any value "A" or moment about zero.

$$m_1 = m_1' - m_1'$$

$$m_2 = m_2' - (m_1')^2$$

$$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3$$

$$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 + 2(m_1')^3$$

Example: 4.16: Find the first four moments about i) zero ii) the value of 7 iii) the mean from the following values**.4,**7,9,5,8,3,6.

Solution:

Moment about zero

X	D=(X-0)	D^2	D^3	D^4
4	4	16	64	256
7	7	49	343	2401
5	5	25	125	625
9	9	81	729	6561
8	8	64	512	4096
3	3	9	27	81
6	6	36	216	1296
Total=42	42	280	2016	15316

$$m_1' = \frac{\sum (X_i - 0)}{n} = \frac{\sum D}{n} = \frac{42}{7} = 6$$

$$m_2' = \frac{\sum (X_i - 0)^2}{n} = \frac{\sum D^2}{n} = \frac{280}{7} = 40$$

$$m_3' = \frac{\sum f(X_i - 0)^3}{\sum f} = \frac{\sum fD^3}{\sum f} = \frac{2016}{7} = 288$$

$$m_4' = \frac{\sum (X_i - 0)^4}{n} = \frac{\sum D^4}{n} = \frac{15316}{7} = 2188$$

Moment about A=7

X	D=(X-7)	D^2	D^3	D^4
4	-3	9	-27	81
7	0	0	0	0
5	-2	4	-8	16
9	2	4	8	16
8	1	1	1	1
3	-4	16	-64	256
6	-1	1	-1	1
Total=42	-7	35	-91	371

***D=X-A**

$$m_1' = \frac{\sum (X_i - 7)}{n} = \frac{\sum D}{n} = \frac{-7}{7} = -1$$

$$m_2' = \frac{\sum (X_i - 7)^2}{n} = \frac{\sum D^2}{n} = \frac{35}{7} = 5$$

$$m_3' = \frac{\sum f(X_i - 7)^3}{\sum f} = \frac{\sum fD^3}{\sum f} = \frac{-91}{7} = -13$$

$$m_4' = \frac{\sum (X_i - 7)^4}{n} = \frac{\sum D^4}{n} = \frac{371}{7} = 53$$

Now we find the moment about mean by using the relationship between mean and moment about any value "A" or moment about zero.

$m_1 = m_1' - m_1' = -1 + 1 = 0$ 1^{st} moment about mean equal to zero

$m_2 = m_2' - (m_1')^2 = 5 - (-1)^2 = 4$ ^{2nd} moment about mean equal to variance

$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3 = -13 - 3(5)(-1) + 2(-1)^3 = 0$

$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 + 2(m_1')^3 = 53 - 4(-13)(-1) + 6(5)(-1)^2 - 3(-1)^4 = 28$

Moment about mean

$$\bar{X} = \frac{\sum X}{n} = \frac{42}{7} = 6$$

X	$D = \sum(X - \bar{X})$	$D^2 = \sum(X - \bar{X})^2$	$D^3 = \sum(X - \bar{X})$	$D^4 = \sum(X - \bar{X})^4$
4	-2	4	-8	16
7	1	1	1	1
5	-1	1	-1	1
9	3	9	27	81
8	2	4	8	16
3	-3	9	-27	81
6	0	0	0	0
Total=42	0	28	0	196

$$m_1 = \frac{\sum(X_i - \bar{X})}{n} = \frac{0}{7} = 0$$ 1st moment about mean

$$m_2 = \frac{\sum(X_i - \bar{X})^2}{n} = \frac{28}{7} = 4$$ 2nd moment about mean

$$m_3 = \frac{\sum(X_i - \bar{X})^3}{n} = \frac{0}{7} = 0$$ 3rd moment about mean

$$m_4 = \frac{\sum(X_i - \bar{X})^4}{n} = \frac{196}{7} = 28$$ 4th moment about mean

Example.4.17: To calculate the moments about the mean first calculate the moment about the value A with common class interval i.e. h=5 then show that by moment about mean are the same result by coding methods.

Marks	20-24	25-29	30-34	35-39	40-44	45-49	50-54
f	1	4	8	11	15	9	2
X	22	27	32	37	42	47	52

Solution: About Arbitrary value

X	f	$D = X - 37$	$fD = f(X - 37)$	fD^2	fD^3	fD^4
22	1	-15	-15	225	-3375	50625
27	4	-10	-40	400	-4000	40000
32	8	-5	-40	200	-1000	5000
37	11	0	0	0	0	0
42	15	5	75	375	1875	9375
47	9	10	90	900	9000	90000
52	2	15	30	450	6750	101250
Total	50		100	2550	9250	296250

Solution:

$$m_1' = \frac{\sum f(X_i - 37)}{\sum f} = \frac{\sum fD}{\sum f} = \frac{100}{50} = 2$$

$$m_2' = \frac{\sum f(X_i - 37)^2}{\sum f} = \frac{\sum fD^2}{\sum f} = \frac{2550}{50} = 51$$

$$m_3' = \frac{\sum f(X_i - 37)^3}{\sum f} = \frac{\sum fD^3}{\sum f} = \frac{9250}{50} = 185$$

$$m_4' = \frac{\sum f(X_i - 37)^4}{\sum f} = \frac{\sum fD^4}{\sum f} = \frac{296250}{50} = 5925$$

Now we find the moment about mean by using the relationship between mean and moment about any value "A" or moment about zero.

$m_1 = m_1' - m_1' = 2 - 2 = 0$ 1st moment about mean equal to zero

$m_2 = m_2' - (m_1')^2 = 51 - (2)^2 = 47$ 2nd moment about mean equal to variance

$m_3 = m_3' - 3m_2'm_1' + 2(m_1')^3 = 185 - 3(51)(2) + 2(2)^3 = -105$

$m_4 = m_4' - 4m_3'm_1' + 6m_2'(m_1')^2 + 3(m_1')^4 = 5925 - 4(185)(2) + 6(51)(2)^2 - 3(2)^4 = 5621$

X	f	fX	$D = (X - 39)$	fD	fD^2	fD^3	fD^4
22	1	22	-17	-17	289	-4913	83521
27	4	108	-12	-48	576	-6912	82944
32	8	256	-7	-56	392	-2744	19208
37	11	407	-2	-22	44	-88	176
42	15	630	3	45	135	405	1215
47	9	423	8	72	576	4608	36864
52	2	104	13	26	338	4394	57122
259	50	1950	-14	0	2350	-5250	281050

$$m_1 = \frac{\sum f(X_i - \bar{X})}{\sum f} = \frac{\sum fD}{\sum f} = \frac{0}{50} = 0 \qquad \bar{X} = \frac{\sum fX}{\sum f} = \frac{1950}{50} = 39$$

$$m_2 = \frac{\sum f(X_i - \bar{X})^2}{\sum f} = \frac{\sum fD^2}{\sum f} = \frac{2350}{50} = 47$$

$$m_3 = \frac{\sum f(X_i - \bar{X})^3}{\sum f} = \frac{\sum fD^3}{\sum f} = \frac{-5250}{50} = -105$$

$$m_4 = \frac{\sum f(X_i - \bar{X})^4}{\sum f} = \frac{\sum fD^4}{\sum f} = \frac{281050}{50} = 5621$$

Hence proved moment about direct method and indirect method both are same

Calculate first four raw moments about the value of "A" by coding method

X	f	u	fu	fu2	fd3	fd4
22	1	-3	-3	9	-27	81
27	4	-2	-8	16	-32	64
32	8	-1	-8	8	-8	8
37	11	0	0	0	0	0
42	15	1	15	15	15	15
47	9	2	18	36	72	144
52	2	3	6	18	54	162
259	50	0	20	102	74	474

$$u = \frac{X - 37}{5}$$

A=37

h=5

For continuous data

$$m_1' = \frac{\sum fu}{\sum f} \times h = 2$$

$$m_2' = \frac{\sum fu^2}{\sum f} \times h^2 = 51$$

$$m_3' = \frac{\sum fu^3}{\sum f} \times h^3 = 185$$

$$m_4' = \frac{\sum fu^4}{\sum f} \times h^4 = 5925$$

Symmetry

When the two tails of frequency curve are equal in length from the central value then it is called symmetry. In this distribution

Mean=median=mode

In a symmetrical distribution deviation below the mean are exactly equals corresponding the deviation above the mean. Two quartiles are at equal distance fro the median

Q_3-median=median-Q_1

Skewness

The lack of symmetry in a distribution around some central value is called skewness.

Or

A distribution is said to skewed if it is not symmetrical Skewness is the degree of asymmetry. In skewed distribution $mean \neq median \neq \text{mod}\, e$ .The two tails of the frequency curve are not equal in length.

Types of skewness

a) Positive skewness

If the right tail of a frequency curve is longer than the left tail of the distribution is said to be positively skewed. In positively skewed distribution mean greater than median and median is greater than mode

Mean>median>mode

b) Negative skewness

If the left tail of a frequency curve is longer than the right tail of the distribution is said to be negatively skewed. In negatively skewed distribution mode greater than median and median is greater than mean or mean less than median and median less than mode
Mean>median>mode

Measures of Skewness

1) Karl Pearson 1st coefficient of skewness

$$S_k = \frac{Mean - Mode}{S.D}$$

2) Karl Pearson 2nd coefficient of skewness

$$S_k = \frac{3(Mean - Median)}{S.D}$$

3) Bow ley's Quartile Coefficient of Skewness

$$S_k = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q}$$

Moment Coefficient of Skewness

$$b_1 = \frac{m_3^{\ 2}}{m_2^2}$$

a) Interpretation of Coefficient of Skewness

The measures No.1, 2, and 3 are varies between -1 and +1
i) If $S_K = 0$ the distribution is symmetrical
ii) If $S_K > 0$ the skewness is positive, so distribution is positively skewed
iii) If $S_K < 0$ the skewness is negative, so distribution is negatively skewed

b) For moment of skewness

i) If $b_1=0$ Then the distribution is symmetrical
ii) If m_3 is positive the distribution is positively skewed
iii) If m_3 is negative the distribution is negatively skewed

Kurtosis

kurtosis is the degree of peakedness or flatness of a unimodal frequency curve.
A distribution whose frequency curve is relatively high peaked is called **leptokurtic**
A distribution whose frequency curve is flat-topped peaked is called **Platykurtic**
A distribution whose frequency curve is neither very peaked nor flat-topped is called **Mesokurtic or normal distribution**

Measure of kurtosis

Kurtosis is measured by mean of moment ratio i.e. b_2

i) Moment coefficient of kurtosis $b_2 = \frac{m_4}{m_2^2}$

ii) Percentile coefficient of kurtosis $K = \frac{Q.D}{p_{90} - p_{10}}$

where $0 \le k \le 0.5$ for Normal Distribution k=0.263
If $b_2=3$ The distribution is Mesokurtic or normal or symmetrical

If $b_2>3$ The distribution is leptokurtic
If $b_2<3$ The distribution is Platykurtic

Describing the frequency distribution

To describe the major characteristics frequency distribution we need the calculations of the following five quantities.
i) The number of observations that describes the size of the data
ii) A measure of central tendency such as the mean or median that provides information about the centre or average value
iii) A measure of dispersion such as S.D that indicates the variability of the data
iv) A measure of skewness that shows the lack of symmetry in the frequency distribution
v) A measure of kurtosis that gives information about its peakedness
It is interesting to note that all these quantities can be derived from the first four moments. For example the first moment about x=0 is the arithmetic mean the second moment about mean is the variance and third moment is a measure of skewness while the fourth central moment is used to measure kurtosis. Thus the first four moments play a key role in describing frequency distribution.

Example:18: Find the i) Pearson's first and second coefficient of skewness ii) Bowely's Quartile coefficient of skewness iii) moment coefficient of skewness iv) moment coefficient of kurtosis. From the following information
Mean=156.17 S.D=19.03 Mode=147.36 Median=153.50
$\mu_2 = 362.28$ $\mu_3 = 4298.08$ $\mu_4 = 3\sigma^4 = 3(19.03)^4 = 393438.09$
Q_1=142.36 Q_3=167.83 and interpret the results
Solution:

1) Karl Pearson 1st coefficient of skewness

$$S_k = \frac{Mean - Mode}{S.D} = \frac{156.17 - 147.36}{19.03} = 0.46$$ Positively skewed

2) Karl Pearson 2nd coefficient of skewness

$$S_k = \frac{3(Mean - Median)}{S.D} = \frac{3(156.17 - 153.50)}{19.03} = 0.14$$ Positively skewed

3) Bow ley's Quartile Coefficient of Skewness

$$S_k = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1} = \frac{167.83 + 142.36 - 2(153.50)}{167.83 - 142.36} = \frac{3.19}{25.47} = 0.13$$ +ve skewed

Moment Coefficient of Skewness

$$b_1 = \frac{m_3^2}{m_2^2} = \frac{(4298.08)^2}{(362.28)^3} = 0.40 = 0$$ The distribution is symmetrical

Moment coefficient of kurtosis

$$b_2 = \frac{m_4}{m_2^2} = \frac{393438.09}{(362.28)^2} = 3.0$$ The distribution is symmetrical or Mesokurtic