Module 6

SKEWNESS AND KURTOSIS OF DISTRIBUTION (DATA)

Introduction

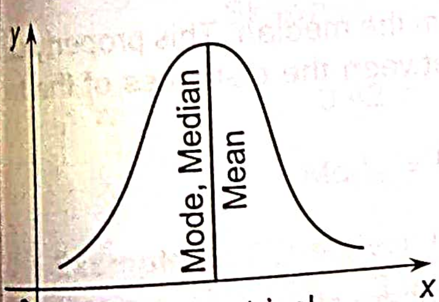
We know that raw data are voluminous and not easy to grasp. We condense them by calculating easure of central tendency such as mean, median, mode, etc. and also by calculating a measure dispersion such as range, quartile deviation, standard deviation, etc. But these two figures are t enough to enable us to draw sufficient inferences about data. A frequency distribution has two re characteristics symmetry and flatness. A graph of a frequency distribution may be symmetrical skew. It can also be flat or peaked. These characteristics are measured from the coefficient of ewness and kurtosis respectively.

Skewness

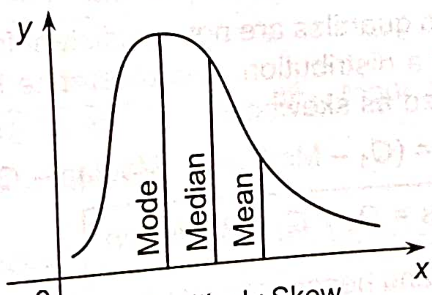
It may happen that two distributions have the same mean and the same standard deviation; t one is symmetrical and the other is not. A distribution which is not symmetrical about the mode, called skew. By skewness we mean the lack of symmetry of the distribution.

For a symmetrical distribution, the values, at equal distances on either side of the mode, have ual frequencies. As a result of this, the mean, mode and median-all coincide for a symmetrical stribution. The frequency curve of a symmetrical distribution rises slowly, reaches a maximum and s *equally* slowly.

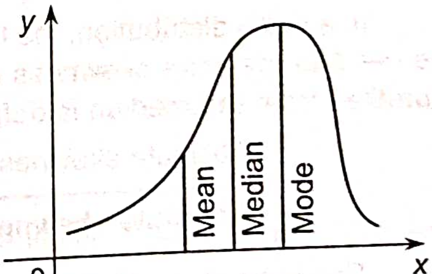
On the other hand, for a skew distribution, the mean, mode and median do not coincide. ewness is called positive or negative according as mean and median are to the right or to the left mode. A positively skew distribution curve rises rapidly, reaches a maximum and falls *slowly*. A egatively skew distribution curve rises slowly, reaches a maximum and falls *rapidly*. These curves re shown in following figures.



Symmetrical
Mode = Median = Mean



Positively Skew
Mode < Median < Mean



Negatively Skew
Mean < Median < Mode

3. Measures of Skewness

Like measures of dispersion, measures of skewness are of two types (i) Absolute measures of kewness and (ii) Relative measures of skewness.

Since for a skew distribution mean and mode do not coincide, the difference between them an be used as a measure of skewness. Similarly, since for a skew distribution the two quartiles are ot equidistant from the median, the difference between the distances of the two quartiles from the edian can be used as a measure of skewness. The first method was suggested by Karl Pearson nd the second by Bowley. The magnitude of the measure of skewness shows the extent of skewness nd the sign shows the nature-positive or negative. The greater the measure, the greater is the skewness. If the measure is positive, skewness is positive, if the measure is negative, the skewness is negative.

(a) Measures of Skewness Based on Mean

Since in a skew distribution mean and mode do not coincide, the distance between them is used to measure skewness.

$$\boxed{\textbf{Absolute Skewness} = \textbf{Mean} - \textbf{Mode}}$$

If mode is ill-defined but the distribution is moderately skew, we use the formula : Mean – Mode = 3 (Mean – Median). In such a case,

$$\boxed{\textbf{Absolute Skewness} = 3\,(\textbf{Mean} - \textbf{Median})}$$

However, since these measures are expressed in the units of distribution, they are not very useful for comparing skewness of distributions which are measured in different units. Hence, for comparing two or more distributions we use another measure of skewness. It is called the relative measure of skewness and is obtained by dividing the above measures by standard deviation (σ). This is Karl Pearson's coefficient of skewness. It is defined as

$$\boxed{\textbf{Karl Pearson's coefficient of skewness} = \frac{\textbf{Mean} - \textbf{Mode}}{\sigma}}$$

In case the mode is ill-defined, the coefficient of skewness is obtained from the following formula.

$$\boxed{\textbf{Karl Pearson's coefficient of skewness} = \frac{3\,(\textbf{Mean} - \textbf{Median})}{\sigma}}$$

The value of this coefficient usually lies between + 1 and – 1. If the distribution is symmetrical, the mean and mode coincide and consequently the coefficient is zero. If the coefficient is positive the distribution is positively skew and if it is negative the distribution is negatively skew.

(d) Measures of Skewness Based on Quartiles

In a skew distribution, the two quartiles are not equidistant from the median. This property can be used to measure skewness of a distribution. The difference between the distances of the two quartiles from the median is defined as skewness.

$$\text{Absolute skewness} = (Q_1 - \text{Median}) - (\text{Median} - Q_3)$$

$$\boxed{\textbf{Absolute skewness} = Q_3 + Q_1 - 2\ \textbf{Median}}$$

This is an absolute measure and hence, cannot be used to compare skewness of two or more distributions. The corresponding relative measure of skewness called Bowley's coefficient of skewness is obtained by dividing the above measure by interquartile range.

$$\boxed{\textbf{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2\ \textbf{Median}}{Q_3 - Q_1}}$$

As before, if this coefficient is zero, the distribution is symmetrical. If it is positive, the distribution is positively skew. If it is negative, the distribution is negatively skew. This coefficient also lies between + 1 and – 1.

If the coefficient of skewness is + 1, then

$$1 = \frac{Q_3 - Q_1 - 2\ \text{Median}}{Q_3 - Q_1} \qquad \therefore\ Q_3 - Q_1 = Q_3 + Q_1 - 2\ \text{Median} \qquad \therefore\ \text{Median} = Q_1.$$

If the coefficient of skewness is – 1, then

$$-1 = \frac{Q_3 + Q_1 - 2\text{ Median}}{Q_3 - Q_1} \qquad \therefore\ -Q_3 + Q_1 = Q_3 + Q_1 - 2\text{ Median} \qquad \therefore\ \text{Median} = Q_3.$$

Thus, when the coefficient of skewness is +1, the median coincides with the first quartile and when the coefficient of skewness is – 1, the median coincides with the third quartile.

Example 1 : Given that A.M. = 160, Mode = 157, S.D. = 50. Find (i) Karl Pearson's coefficient of skewness, (ii) Coefficient of variation.

Sol. : We have Karl Pearson's coefficient of

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

But A.M. = 160, Mode = 157 and σ = 50.

$$\therefore \qquad \text{Skewness} = \frac{160 - 157}{50} = \frac{3}{50} = 0.6$$

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100 = \frac{50}{160} \times 100 = 31{\cdot}25.$$

Example 2 : For a moderately skew distribution the mean, median and Karl Pearson's coefficient of skewness are 86, 80 and 0·42, respectively. Find the mode and coefficient of variation.

Sol. : We have, $\bar{x}$ = 86, Median = 80, Coefficient of skewness = 0.42

$$\text{Now, the coefficient of skewness} = \frac{3\,(\text{Mean} - \text{Median})}{\sigma}$$

Putting the given values

$$0.42 = \frac{3\,(86 - 80)}{\sigma} \qquad \therefore\ \sigma = \frac{18}{0{\cdot}42} = 42{\cdot}85$$

Further, the coefficient of

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$\therefore \qquad 0.42 = \frac{86 - \text{Mode}}{18/0{\cdot}42} \qquad \therefore\ 86 - \text{Mode} = 18$$

$$\therefore \qquad \text{Mode} = 86 - 18 = 68.$$

Example 3 : For a given frequency distribution mean, mode and Karl Pearson's coefficient of skewness are 120, 123, – 3 respectively. Find C.V.

Sol. : We have Karl Pearson's coefficient of

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

But Mean = 120, mode = 123 and skewness = – 3

$$\therefore \qquad -3 = \frac{120 - 123}{\sigma} \qquad \therefore\ -3 = -\frac{3}{\sigma} \qquad \therefore\ \sigma = 1.$$

$$\text{Coefficient of variation} = \frac{\sigma}{\bar{x}} \times 100 = \frac{1}{120} \times 100 = \frac{5}{6}.$$

Example 4 : From the data given below calculate Karl Pearson's coefficient of skewness and interpret your result.

Wages (₹) : 70-80, 80-90, 90-100, 100-110, 110-120, 120-130, 130-140, 140-150

No. of persons : 12, 18, 35, 42, 50, 45, 20, 8.

Sol. :

Calculation of Karl Pearson's Coefficient of Skewness

Wages	No. of persons	Mid-points	$m - 105$	$d / 10$		
x_i	f_i	m_i	d_i	d_i'	$f_i d_i'$	$f_i d_i'^2$
70-80	12	75	−30	−3	−36	108
80-90	18	85	−20	−2	−36	72
90-100	35	95	−10	−1	−35	35
100-110	42	105	0	0	0	0
110-120	50	115	10	1	50	50
120-130	45	125	20	2	90	180
130-140	20	135	30	3	60	180
140-150	8	145	40	4	32	128
Total	230				125	753

$$\text{Mean, } \bar{x} = A + \frac{\Sigma f_i d_i}{N} \times h$$

Here, $A = 105$, $\Sigma f_i d_i = 125$, $N = 230$, $h = 10$

$$\therefore \quad \text{Mean, } \bar{x} = 105 + \frac{125}{230} \times 10 = 105 + 5{\cdot}43 = 110.43$$

$$\text{Mode} = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Modal class is 110-120. Here, $L_1 = 110$, $f_1 = 50$, $f_0 = 42$, $f_2 = 45$, $i = 10$.

$$\therefore \quad \text{Mode} = 110 + \frac{50 - 42}{100 - 42 - 45} \times 10$$

$$= 110 + \frac{8}{13} \times 10 = 116{\cdot}15$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\Sigma f_i d_i'^2}{N} - \left(\frac{\Sigma f_i d_i'}{N}\right)^2} \times h = \sqrt{\frac{753}{230} - \left(\frac{125}{230}\right)^2} \times 10$$

$$= \sqrt{3{\cdot}274 - 0{\cdot}294} \times 10 = \sqrt{2{\cdot}98} \times 10$$

$$= 1.755 \times 10 = 17.55$$

$$\text{Coefficient of skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{110{\cdot}13 - 116{\cdot}5}{17{\cdot}55} = -\frac{5{\cdot}72}{17{\cdot}55} = -0{\cdot}326.$$

Since, the coefficient is negative, the distribution is negatively skewed.

Example 5 : From the following data, calculate Bowley's coefficient of skewness :

Weekly earnings in ₹ : 10-12, 12-14, 14-16, 16-18, 18-20, 20-22, 22-24, 24-26,

No. of employees : 3, 6, 10, 15, 24, 42, 75, 90,

26-28, 28-30, 30-32, 32-34, 34-36, 36-38, 38-40

79, 55, 36, 26, 16, 16, 7.

Sol. : We have to calculate median and the two quartiles first.

Median = size of $(N/2)^{th}$ item

= size of 250^{th} item.

$\therefore$ Median class is 24-26.

Calculation of Median

Weekly earnings in ₹ (x_i)	No. of employees (f_i)	c.f.
10-12	3	3
12-14	6	9
14-16	10	19
16-18	15	34
18-20	24	58
20-22	42	100
22-24	75	175
24-26	90	265
26-28	79	344
28-30	55	399
30-32	36	435
32-34	26	461
34-36	16	477
36-38	16	493
38-40	7	500
	$N = 500$	

$$\text{Median} = L_1 + \frac{(N/2) - \text{c.f.}}{f} \times i$$

$$= 24 + \frac{250 - 175}{90} \times 2$$

$$= 24 + \frac{75}{90} \times 2$$

$$= 24 + 1{\cdot}66$$

$$\therefore \text{Median} = ₹\ 25{\cdot}66$$

The first quartile,

$$Q_1 = \text{Size of } (N/4)^{th} \text{ item}$$

$$= \text{Size of } 125^{th} \text{ item}$$

$$Q_1 = L_1 + \frac{(N/4) - \text{c.f.}}{f} \times i$$

$$= 22 + \frac{125 - 100}{75} \times 2$$

$$= 22 + \frac{25}{75} \times 2$$

$$\therefore Q_1 = ₹\ 22{\cdot}66$$

The third quartile,

$$Q_3 = \text{Size of } (3N/4)^{th} \text{ item} = \text{Size of } 375^{th} \text{ item}$$

$$Q_3 = L_1 + \frac{(3N/4) - \text{c.f.}}{f} \times i = 28 + \frac{375 - 344}{55} \times 2$$

$$\therefore Q_3 = 28 + \frac{31}{55} \times 2 = 28 + 1{\cdot}13 = ₹\ 29{\cdot}13.$$

Now, Bowley's Coefficient of Skewness

$$= \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1} = \frac{29{\cdot}13 + 22{\cdot}66 - 2(25{\cdot}66)}{29{\cdot}13 - 22{\cdot}66}$$

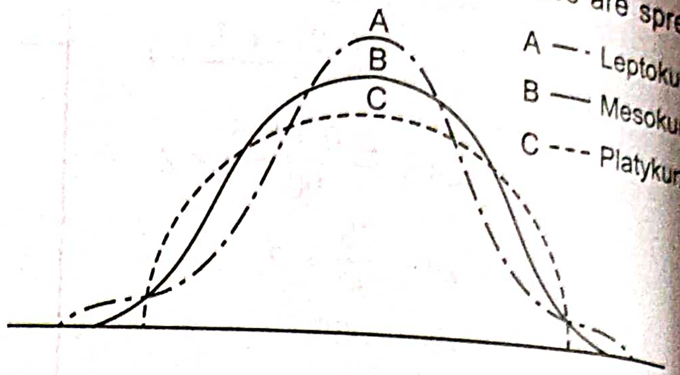
$$= \frac{51{\cdot}79 - 51{\cdot}32}{6{\cdot}47} = \frac{0{\cdot}47}{6{\cdot}47} = 0{\cdot}073.$$

4. Kurtosis

The term kurtosis is derived from a Greek word meaning bulging or convexity. In addition to the characteristics considered so far *viz.* measures of central tendency, measures of dispersion and measures of skewness, we may need to know whether the frequency distribution curve is peaked or flat around the central value, usually the mode. This character of the curve is indicated by the measure of **Kurtosis**.

Two distributions may be perfectly symmetrical about the mode but one may be flat around the mode and the other maybe peaked around it. To understand this nature of a given frequency distribution curve, we compare it with a perfectly symmetrical and in a sense ideal curve which is neither flat nor peaked, called the normal curve. If a curve is more peaked than the normal curv is called **leptokurtic.** If one the other hand it is more flat around the mode than the normal, i called, **platykurtic**. The normal curve itself is called, **mesokurtic**. In a platykurtic curve the valu of the variable are clustered around the mode and in a mesokurtic curve the values are spre evenly around the mode. The name kurtosis was suggested by Karl Pearson.

The adjoining diagram shows the nature of the three types of curves. The curve A is more peaked than the normal curve and is leptokurtic. The curve B is the normal one and is mesokurtic. C is more flat and is platykurtic.



5. Measures of Kurtosis

Kurtosis is measured by the coefficient $\beta_2 = \frac{\mu_4}{\mu_2}$.

For a normal or mesokurtic curve $\beta_2 = 3$ and hence, $\beta_2 > 3$ the curve is leptokurtic and $\beta_2 < 3$ the curve is platykurtic. For this reason, kurtosis can also be measured by the differenc $\beta_2 - 3$ which is denoted by γ_2. Hence, $\gamma_2 = \beta_2 - 3$. For a normal or mesokurtic curve $\gamma_2 = 0$, for leptokurtic curve γ_2 is positive and for platykurtic curve γ_2 is negative.

6. Person's β and γ Coefficients

Karl Pearson defined the following four coefficients, based upon the first four moments abo mean :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \gamma_1 = \sqrt{\beta_1}; \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} \text{ and } \gamma_2 = \beta_2 - 3$$

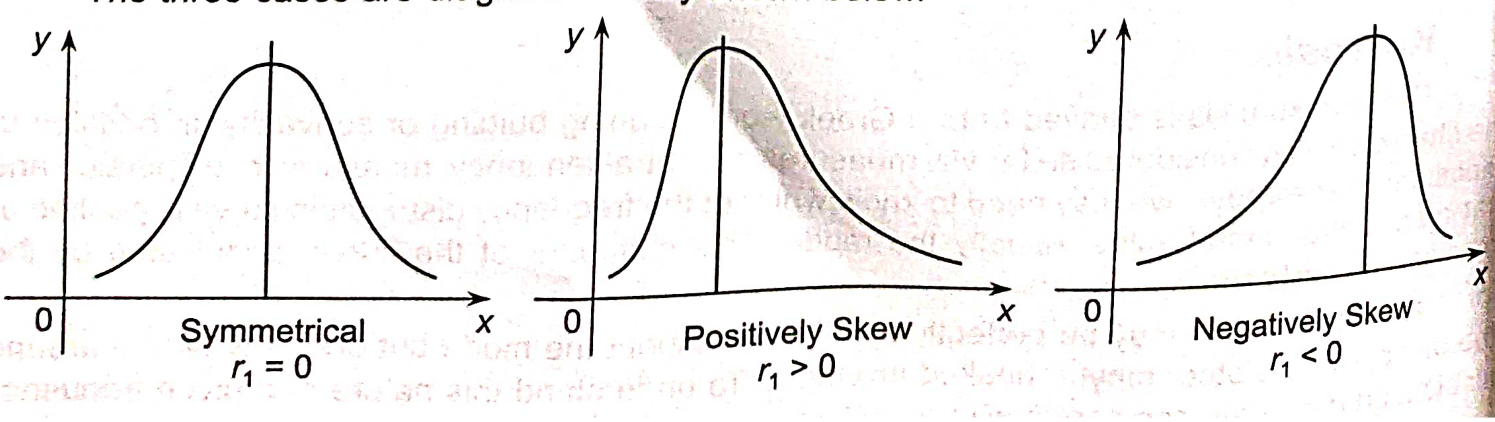
As seen above β_2 and γ_2 are used to measure kurtosis. β_1 and γ_1 are used to measu skewness.

Coefficient of skewness = $\gamma_1 = \sqrt{\beta_1}$ (with the sign of μ_3)

Three special cases

Three cases of r_1 deserve special attention. If $r_1 = 0$, *i.e.*, if $\mu_3 = 0$, it means skewness is zero Hence, if $\gamma_1 = 0$, the curve is perfectly symmetrical. If $r_1 > 0$, skewness is greater than zero. Hence if $\gamma_1 > 0$, the curve is positively skew. If $\gamma_1 < 0$, skewness is less than zero. Hence, if $\gamma_1 < 0$, the curv is negatively skew.

The three cases are diagrammatically shown below.



Further, since for symmetric distribution $\gamma_1 = 0$, which means $\beta_1 = 0$, *i.e.*, $\mu_3 = 0$. Now, from the relation for μ_3', we have

$$\mu_3' = \mu_3 + 3\,\mu_2\,\mu_1' + \mu_1'^{\,3}$$

Since, for symmetrical distribution $\mu_3 = 0$, we get

$$\mu_3' = 3\,\mu_2\,\mu_1' + \mu_1'^{\,3}, \quad i.e., \quad \frac{\mu_3}{\mu_1'} = 3\,\mu_2 + \mu_1'^{\,2}$$

Note ....

We first see that μ_r has the dimension of (variate)^r. Hence, μ_2^3 has 6th dimension and μ_2^3 has also 6th dimension. Hence, $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$ is of zeroth dimension and hence, is a pure number. Similarly, $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$ is also of zeroth dimension. Hence, β_1, β_2 (and also γ_1 and γ_2) are pure numbers and as such are independent of scale and origin.

Example 1 : Second, third and fourth central moments of a variable are 19, 97, 29, 26, 866 respectively. Calculate the beta coefficients correct to three decimal places.

Sol. : We are given that $\mu_2 = 19{\cdot}67$, $\mu_3 = 29{\cdot}26$, $\mu_4 = 866$.

By definition,
$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(29{\cdot}26)^2}{(19{\cdot}67)^2} = 0{\cdot}113$$
$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{866}{(19{\cdot}67)^2} = 2{\cdot}24$$

Example 2 : For any frequency distribution, show that $\beta_2 \geq 1$.

Sol. : Let us consider a frequency distribution $\dfrac{x_i}{f_i}$, $i = 1, 2, 3, \ldots, n$.

We have to show that $\beta_2 = \dfrac{\mu_4}{\mu_2^2} \geq 1$, *i.e.*, $\mu_4 \geq \mu_2^2$.

$$i.e., \quad \frac{1}{N}\Sigma f_i\,(x_i - \overline{x})^4 \geq \left[\frac{1}{N}\Sigma f_i (x_i - \overline{x})^2\right]^2$$

If we put $(x_i - \overline{x})^2 = z_i$, then we have to prove that

$$\frac{1}{N}\Sigma f_i\, z_i^2 \geq \left[\frac{1}{N}\Sigma f_i\, z_i\right]^2$$

$$i.e., \quad \frac{1}{N}\Sigma f_i\, z_i^2 - \left[\frac{1}{N}\Sigma f_i\, z_i\right]^3 \geq 0, \quad i.e., \quad \sigma_z^2 > 0$$

which is true because variance is always positive. Hence, the result.

Example 3 : The first four moments of a distribution are 1, 4, 10 and 46 respectively. Compute the first four central moments and Beta constants. Comment upon the nature of the distribution.

Sol. : We are not given the value of A about which the moments are calculate, however, this value is not required since, we do not need the actual mean.

We are given, $\mu_1' = 1$, $\mu_2' = 4$, $\mu_3' = 10$, $\mu_4' = 46$.

Now, $\mu_2 = \mu_2' - \mu_1'^2 = 4 - 1 = 3$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1^3$$
$$= 10 - 3(4)(1) + 2(1)^2 = 10 - 12 + 2 = 0$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^3$$
$$= 46 - 4(10)(1) + 6(4)(1)^2 - 3(1)^4$$
$$= 46 - 40 + 24 - 3 = 27$$

$$\therefore \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{27} = 0; \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{27}{9} = 3.$$

Since, $\beta_1 = 0$, the distribution has no skewness.

Since, $\beta_2 = 3$, the distribution is mesokurtic.

Example 4 : Find the mean, variance and β_1 and β_2 for the following distribution.

Class interval :	0-10,	10-20,	20-30,	30-40
Frequency :	1,	3,	4,	2

Sol. : **Calculation of mean etc.**

Class	**Frequency**	**Mid Point**	$\frac{m-25}{10}$				
x_i	f_i	m_i	d_i'	$f_i d_i'$	$f_i d_i'^2$	$f_i d_i'^3$	$f_i d_i'^4$
0-10	1	5	−2	−2	4	−8	16
10-20	3	15	−1	−3	3	−3	3
20-30	4	25	0	0	0	0	0
30-40	2	35	1	2	2	2	2
Total	10			−3	9	−9	21

Moments about the assumed mean 25 are given by

$$\mu_1' = \frac{\Sigma f_i d_i'}{N} \cdot h = -\frac{3}{10} \cdot 10 = -3; \qquad \mu_2' = \frac{\Sigma f_i d_i'^2}{N} \cdot h^2 = \frac{9}{10} \cdot 10^2 = 90;$$

$$\mu_3' = \frac{\Sigma f_i d_i'^3}{N} \cdot h^3 = -\frac{9}{10} \cdot 10^3 = -900; \qquad \mu_4' = \frac{\Sigma f_i d_i'^4}{N} \cdot h^4 = \frac{21}{10} \cdot 10^4 = 21000$$

$$\text{Arithmetic mean} = A + \frac{\Sigma f_i d_i'}{N} \cdot h = 25 - \frac{3}{10} \cdot 10 = 22$$

Variance, $\sigma^2 = \mu_2 = \mu_2' - \mu_1'^2 = 90 - 9 = 81$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1^3$$
$$= -900 - 3(90)(3) + 2(27) = -1556$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 21000 - 4(-900)(-3) + 6(90)(-3)^2 - 3(-3)^4$$
$$= 21000 - 10800 + 4860 - 243 = 14817$$

Now, $$\beta_1 = \frac{\mu_3^2}{\mu_2^2} = \frac{(-1556)^2}{(81)^2} = 4{\cdot}555; \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{14817}{81^2} = 2{\cdot}258.$$