4

Mphil zoology
2nd smste
Applied biostatics

Measures of Dispersion, Moments and Skewness

4.1 INTRODUCTION

It is quite possible that two or more sets of data may have the same average (mean, median or mode) but their individual observations may differ considerably from the average. Thus a value of central tendency does not adequately describe the data. We therefore need some additional information concerning with how the data are dispersed about the average. This is done by measuring the *dispersion* by which we mean (the extent to which the observations in a sample or in a population vary about their mean. A quantity that measures this characteristic, is called a measure of *dispersion, scatter or variability)* It is desirable to have the measure of dispersion (i) in the same units as the observations, (ii) zero when all the observations are equal, (iii) independent of origin, (iv) multiplied or divided by the constant when each observation is multiplied or divided by a constant. It is also desirable that it should satisfy the conditions similar to those laid down for an average in previous chapter (see section 3.2).

(i) (ii)

vip (There are two types of measures of dispersion) *absolute* and *relative*.

(i) An *absolute* measure of dispersion is one that measures the dispersion in terms of the same units or in the square of units, as the units of the data. For example, if the units of the data are rupees, metres, kilograms, etc., the units of the measures of dispersion will also be rupees, metres, kilograms, etc. (ii) A *relative* measure of dispersion is one that is expressed in the form of a ratio, co-efficient or percentage and is independent of the units of measurement. It is useful for comparison of data of different nature. A measure of central tendency together with a measure of dispersion gives an adequate description of data.

The main measures of dispersion are the following:

(i) The Range.

(ii) The Semi-Interquartile Range or the Quartile Deviation.

(iii) The Mean Deviation or the Average Deviation.

(iv) The Variance and the Standard Deviation.

4.2 THE RANGE

The *range* R, is defined as the difference between the largest and the smallest observations in a set of data. Symbolically, the range is given by the relation

$$R = x_m - x_0,$$

where x_m stands for the largest observation and x_0 denotes the smallest one. When the data are grouped into a frequency distribution, the range is estimated by finding the difference between the upper boundary of the highest class and the lower boundary of the lowest class. The range cannot be computed if there are any open-end classes in the distribution.

The range is a simple concept, and is easy to compute. It has, however, two serious disadvantages. *First*, it ignores all the information available from the intermediate observations; and *second*, as its value is based only on the two extreme (unusually large or small) observations, it might give a misleading picture of the spread in the data. It is therefore an unsatisfactory measure of dispersion. However, it is appropriately used in statistical quality control charts of manufactured products, daily temperatures, stock prices, etc. This is an absolute measure of dispersion. Its relative measure known as the *co-efficient of dispersion*, is defined by the following relation:

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

This is a pure (i.e. dimensionless) number and is used for the purposes of comparison.

Example 4.1 The marks obtained by 9 students are given below:

45, 32, 37, 46, 39, 36, 41, 48, 36.

Find the range and the co-efficient of dispersion.

Here the highest marks, *i.e.* $x_m = 48$,

and the lowest marks, *i.e.*, $x_0 = 32$.

$\therefore$ $R = x_m - x_0 = 48 - 32 = 16$ marks, and

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

$$= \frac{48 - 32}{48 + 32} = \frac{16}{80} = 0.2$$

4.3 THE SEMI-INTERQUARTILE RANGE OR THE QUARTILE DEVIATION

The *interquartile range* is a measure of dispersion, defined by the difference between the third and first quartiles; and half of this range is called the *semi-interquartile range (S.I.Q.R.)* or *the quartile deviation (Q.D)*. Symbolically, we have

$$I.Q.R = Q_3 - Q_1 \qquad Q.D. = \frac{Q_3 - Q_1}{2},$$

where Q_1 and Q_3 are the first and the third quartiles of the data. The quartile deviation has an attractive feature that the range "Median $\pm$ *Q.D.*" contains approximately 50% of the data. The quartile deviation is superior to range as it is not affected by extremely large or small observations. It is simple to understand and easy to calculate. It has certain disadvantages. It gives no information about the position of observations lying outside the two quartiles; it is not amenable to mathematical treatment and is greatly affected by sampling variability. The quartile deviation is not as widely used as other measures of dispersion. It is, however, used in situations where extreme observations are thought to be unrepresentative.

The quartile deviation is also an absolute measure of dispersion. Its relative measure called the *Co-efficient of Quartile Deviation* or of *Semi-Interquartile Range*, is defined by the relation

$$\text{Co-efficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1},$$

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

Example 4.2 Find the quartile deviation and the co-efficient of quartile deviation for (i) the data in Example 3.11 and (ii) the frequency distribution in Example 3.13.

(i) Using the data of Example 3.11, we find that

$Q_1 = 36$ marks, $Q_3 = 45$ marks, and therefore

$$Q.D. = \frac{45 - 36}{2} = 4.5 \text{ marks}$$

Co-efficient of $Q.D. = \frac{45 - 36}{45 + 36} = \frac{9}{81} = 0.11$

(ii) Values of Q_1 and Q_3 calculated in Example 3.13 are respectively 56 and 74 marks. Therefore

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{74 - 56}{2} = 9 \text{ marks, and}$$

$$\text{Co-efficient of } Q.D. = \frac{74 - 56}{74 + 56} = \frac{18}{130} = 0.14$$

4.4 THE MEAN (OR AVERAGE) DEVIATION

The *mean deviation (M.D.)* of a set of data is defined as the arithmetic mean of the deviations measured either from the mean or from the median, all deviations being counted as positive. The reason to count the deviations as positive, *i.e.* to disregard the algebraic signs (+and−) is to avoid the difficulty arising from the property that the sum of deviations of the observations from their mean is zero. The symbolic definition of the mean deviation from the mean is

$$M.D. = \frac{\sum |x_i - \bar{x}|}{n}, \text{ for sample data,}$$

$$M.D. = \frac{\sum |x_i - \mu|}{N}, \text{ for population data,}$$

where $|x_i - \bar{x}|$ and $|x_i - \mu|$ (pronounced "mod. deviations") indicate the *absolute* deviations of the observations from the mean of a sample and population respectively. It is more appropriate to call it the *mean absolute deviation (M.A.D.)*.

For the data organised into a grouped frequency distribution having k classes with midpoints $x_1, x_2, \ldots, x_k$ and the corresponding frequencies $f_1, f_2 \ldots, f_k$ ($\sum f_i = n$), the mean deviation of the sample is given by

$$M.D. = \frac{\sum f_i |x_i - \bar{x}|}{n}$$

The mean deviation is also defined in terms of absolute deviations from the median in a similar way. Theory tells us that the mean deviation is *least* when the deviations are measured from the median. But in practice, it is generally calculated from the arithmetic mean. The mean deviation gives more information than the *range* or the *quartile deviation* as it is based on all the observed values. It is easily calculated and readily understood. As it is not amenable to mathematical treatment, its usefulness is limited. We introduce artificiality in its

calculation by ignoring the algebraic signs of the deviations and this step is not mathematically defensible. As the mean deviation does not give undue weight to occasional large deviations, so it is used in situations where such deviations are likely to occur. It is unsatisfactory for statistical inference.

Mean deviation is an absolute measure of dispersion. Its relative measure, known as the *co-efficient of mean deviation*, is obtained by dividing the mean deviation by the average used in the calculation of deviations. Thus

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

Example 4.3 Calculate the mean deviation from (i) the mean, (ii) the median, of the following set of examination marks:

45, 32, 37, 46, 39, 36, 41, 48 and 36.

Also calculate the co-efficient of mean deviation.

We first arrange the given marks in an increasing sequence to find the median. The ordered marks are

32, 36, 36, 37, 39, 41, 45, 46, 48.

$\therefore$ Median = Marks obtained by $\left([\frac{n}{2}] + 1\right)th$ student in ordered data as $\frac{n}{2}$ is not an integer.

= Marks obtained by $\left([\frac{9}{2}] + 1\right)th$, *i.e.* 5th student

= 39 marks

and $\bar{x} = \frac{\Sigma x}{n} = \frac{360}{9} = 40$ marks

The necessary calculations are given below:

	x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i$-median$\rvert$
	32	−8	8	7
	36	−4	4	3
	36	−4	4	3
	37	−3	3	2
	39	−1	1	0
	41	1	1	2
	45	5	5	6
	46	6	6	7
	48	8	8	9
Σ	360	0	40	39

$\therefore$ *M.D.* (from mean) $= \dfrac{\sum |x_i - \bar{x}|}{n} = \dfrac{40}{9} = 4.4$ marks

and *M.D.* (from median) $= \dfrac{\sum |x_i - \text{median}|}{n} = \dfrac{39}{9} = 4.3$ marks

Co-efficient of *M.D.* $= \dfrac{M.D.}{\bar{x}}$ or $\dfrac{M.D.}{\text{median}}$

$= \dfrac{4.4}{40}$ or $\dfrac{4.3}{39} = 0.11$ or 0.11

Example 4.4 Calculate the mean deviation of the following frequency distribution showing the weights of apples:

Weight (grams)	65–84	85–104	105–124	125–144	145–164	165–184	185–204
f	9	10	17	10	5	4	5

The calculation of the mean deviation (*M.D.*) from the mean is illustrated below:

Weight	x_i	f_i	$f_i x_i$	$x_i - \bar{x}$	$f_i \lvert x_i - \bar{x} \rvert$
65 – 84	74.5	9	670.5	–48.0	432.0
85 – 104	94.5	10	945.0	–28.0	280.0
105 – 124	114.5	17	1946.5	–8.0	136.0
125 – 144	134.5	10	1345.0	+12.0	120.0
145 – 164	154.5	5	772.5	32.0	160.0
165 – 184	174.5	4	698.0	52.0	208.0
185 – 204	194.5	5	972.5	72.0	360.0
Total	--	60	7350.0	--	1696.0

Here $\bar{x} = \dfrac{\sum f_i x_i}{n} = \dfrac{7350.0}{60} = 122.5$ grams

Hence *M.D.* $= \dfrac{\sum f_i |x_i - \bar{x}|}{n} = \dfrac{1696.0}{60} = 28.27$ grams.

4.5 THE VARIANCE AND STANDARD DEVIATION

The *variance* of a set of observations is defined as the mean of the squares of deviations of all the observations from their mean. When it is calculated from the entire population, the variance is called the

population variance, traditionally denoted by σ^2(σ is the Greek lower-case "sigma"). If, instead, the data from the sample are used to calculate the variance, it is referred to as the *sample variance* and is denoted by S^2 in order to distinguish between the two. The symbolic definition for variance is

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}, \text{ for population data,}$$

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n}, \text{ for sample data,}$$

The variance is also denoted by Var(X). The term *variance* was introduced in 1918 by R.A. Fisher (1890–1962).

It should be noted that the variance is in square of units in which the observations are expressed and the variance is a large number compared to observations themselves. The variance because of its some nice mathematical properties, assumes an extremely important role in statistical theory.

Standard Deviation. The positive square root of the variance is called *standard deviation*. Symbolically,

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}, \text{ for population data,}$$

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}, \text{ for sample data,}$$

The standard deviation is expressed in the same units as the observations themselves and is a measure of the average spread around the mean. Karl Pearson (1857–1936), "founder of the science of Statistics", is credited with the name standard deviation, the most useful measure of dispersion. The *sample variance* in some texts is defined as

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

where n is replaced by $n - 1$ on the basis of the argument that *knowledge of any $n - 1$ deviations automatically determines the remaining deviation as the sum of n deviations must be zero.* This is, in fact, an *unbiased estimator* of the population variance σ^2, the explanation for which is deferred to chapter on *estimation* where we shall learn that sample variance $S^2 = \frac{\sum (x_i - \bar{x})^2}{n}$, for small samples, *underestimates* the population variance σ^2.

When the data are grouped into a frequency distribution having k classes with midpoints $x_1, x_2, \ldots, x_k$ and the corresponding frequencies $f_1, f_2, \ldots, f_k$ ($\Sigma f_i = n$), the sample variance and standard deviation are given by

$$s^2 = \frac{\Sigma f_i (x_i - \bar{x})^2}{n}, \text{ and}$$

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

It should be noted that for a frequency distribution, as the number of observations or the total frequency n is usually large, dividing the sum of squared deviations by $n-1$ is practically equivalent to dividing it by n.

The standard deviation has a definite mathematical meaning, utilizes all the observed values and is amenable to mathematical treatment but is affected by extreme values. The standard deviation is an absolute measure of dispersion. Its relative measure called *coefficient of standard deviation*, is defined as

$$\text{Coefficient of } S.D. = \frac{\text{Standard Deviation}}{\text{Mean}}$$

The quantity $\sqrt{\frac{\Sigma (x_i - a)^2}{n}}$, where a is some arbitrary origin, is called the *root-mean-square-deviation* which becomes the standard deviation when this arbitrary origin coincides with the mean.

To calculate the variance and standard deviation on an *electronic calculator*, the alternative formulas for use are obtained by showing that $\Sigma (x_i - \mu)^2 = \Sigma x_i^2 - (\Sigma x_i)^2/N$.

$$\text{Now} \quad \Sigma (x_i - \mu)^2 = \Sigma (x_i^2 - 2x_i\mu + \mu^2)$$

$$= \Sigma x_i^2 - 2\mu \Sigma x_i + N\mu^2$$

$$= \Sigma x_i^2 - 2N\mu^2 + N\mu^2 \quad \left(\because \mu = \frac{\Sigma x_i}{N}\right)$$

$$= \Sigma x_i^2 - N\mu^2 = \Sigma x_i^2 - \frac{(\Sigma x)^2}{N}$$

Thus the sum of squares of the deviations from the mean is equal to the sum of the squares of all x_i's minus a *correction factor* which is the $(1/N)th$ of the square of the sum of all x_i's.

$$\therefore \quad \sigma^2 = \frac{\Sigma (x_i - \mu)^2}{N} = \frac{\Sigma x_i^2}{N} - \left(\frac{\Sigma x_i}{N}\right)^2$$

i.e. the variance is the mean of the squares minus the square of the mean.
The corresponding formula for sample variance is

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

The alternative formulas for standard deviations are

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x_i}{N}\right)^2}, \text{ and}$$

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

The following alternative formulas for the sample variance and standard deviation of a frequency distribution are obtained in a similar way.

$$s^2 = \frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2, \text{ and}$$

$$s = \sqrt{\frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2}.$$

Example 4.5 A population of N = 10 has the observations 7, 8, 10, 13, 14, 19, 20, 25, 26 and 28. Find its variance and standard deviation.

Calculations appear in the following table:

	x_i	$x_i - \mu$	$(x_i - \mu)^2$	x_i^2
	7	−10	100	49
	8	−9	81	64
	10	−7	49	100
	13	−4	16	169
	14	−3	9	196
	19	+2	4	361
	20	3	9	400
	25	8	64	625
	26	9	81	676
	28	11	121	784
Σ	170	0	534	3424

Now $\mu = \frac{\sum x_i}{N} = \frac{170}{10} = 17.$

Therefore $$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{534}{10} = 53.4,$$

and $$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{53.4} = 7.31$$

Using the alternative method.

$$\sigma^2 = \frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2$$

$$= \frac{3424}{10} - \left(\frac{170}{10}\right)^2 = 342.4 - 289 = 53.4$$

and $$\sigma = \sqrt{\frac{\sum x_i^2}{N} - \left(\frac{\sum x_i}{N}\right)^2} = \sqrt{53.4} = 7.31$$

Example 4.6 Calculate the variance and standard deviation from the following marks obtained by 9 students.

45, 32, 37, 46, 39, 36, 41, 48, 36

The variance S^2 and the standard deviation S for the sample are calculated as below:

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
	45	5	25	2025
	32	−8	64	1024
	37	−3	9	1369
	46	6	36	2116
	39	−1	1	1521
	36	−4	16	1296
	41	1	1	1681
	48	8	64	2304
	36	−4	16	1296
Σ	360	0	232	14632

Here $$\bar{x} = \frac{\sum x_i}{n} = \frac{360}{9} = 40 \text{ marks}$$

Therefore $$S^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{232}{9} = 25.78 \text{ (marks)}^2$$

and $$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{25.78} = 5.08 \text{ marks}$$

Using the alternative method.

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

$$= \frac{14632}{9} - \left(\frac{360}{9}\right)^2 = 1625.78 - 1600 = 25.78 \text{ (marks)}^2$$

and $$S = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2} = \sqrt{25.78} = 5.08 \text{ marks}$$

Example 4.7 Calculate the variance and standard deviation from the data of Example 4.4

The necessary calculations may be carried out on an *electronic calculator* as below:

x_i	f_i	$f_i x_i$	$f_i x_i^2$
74.5	9	670.5	49 952.25
94.5	10	945.0	89 302.50
114.5	17	1946.5	222 874.25
134.5	10	1345.0	180 902.50
154.5	5	772.5	119 351.25
174.5	4	698.0	121 801.00
194.5	5	972.5	189 151.25
Σ	60	7350.0	973 335.00

Thus we find $$s^2 = \frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2$$

$$= \frac{973335}{60} - \left(\frac{7350}{60}\right)^2 = 16222.25 - 15006.25$$

$$= 1216 \text{ (grams)}^2$$

and $$s = \sqrt{\frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2} = \sqrt{1216} = 34.87 \text{ grams}$$

4.5.1 Change of Origin and Scale. The computational labour can be reduced by using the same transformation as was used for computing the arithmetic mean.

Let $u_i = \frac{x_i - a}{h}$. Then $x_i = a + hu$ and $\bar{x} = a + h\bar{u}$

Therefore $$\sum (x_i - \bar{x})^2 = \sum [(a + hu_i) - (a + h\bar{u})]^2$$

$$= h^2 \sum (u_i - \bar{u})^2$$

Further $$\Sigma(u_i - \bar{u})^2 = \Sigma u_i^2 - \frac{(\Sigma u_i)^2}{n}$$

Hence $$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = \frac{h^2}{n}\left[\Sigma u_i^2 - \frac{(\Sigma u_i)^2}{n}\right]$$

$$= h^2\left[\frac{\Sigma u_i^2}{n} - \left(\frac{\Sigma u_i}{n}\right)^2\right]$$

and $$S = h\sqrt{\frac{\Sigma u^2}{n} - \left(\frac{\Sigma u}{n}\right)^2}$$

This gives us a *short method* for hand calculations.

When the data are grouped into a frequency distribution, the corresponding *short method* for hand calculations, is

$$s = h\sqrt{\frac{\Sigma f_i u_i^2}{n} - \left(\frac{\Sigma f u_i}{n}\right)^2}$$

where h is the width of the class-interval, f_i is the frequency of the *ith* class and u_i is the deviation of x_i from an assumed mean in terms of class intervals. This method is also known as the *step-deviation method.*

Example 4.8 Find the standard deviation by the short method from the data of Example 4.4.

Let $u_i = \frac{x_i - 114.5}{20}$, where $a = 114.5$, value corresponding to the highest frequency, and $h = 20$, the class-interval. Then $u_i = -2, -1, 0, 1, 2, 3, 4$. Other calculations appear below:

x_i	f_i	u_i	$f_i u_i$	$f_i u_i^2$
74.5	9	−2	−18	36
94.5	10	−1	−10	10
114.5	17	0	−28	0
134.5	10	1	10	10
154.5	5	2	10	20
174.5	4	3	12	36
194.5	5	4	20	80
Σ	60	--	+52 +24	192

Therefore $$s = h \times \sqrt{\frac{\Sigma fu^2}{n} - \left(\frac{\Sigma fu}{n}\right)^2}$$

$$= 20 \times \sqrt{\frac{192}{60} - \left(\frac{24}{60}\right)^2} = 20 \times \sqrt{3.20 - 0.16}$$

$$= 20 \times \sqrt{3.04} = 20 \times (1.7436) = 34.87 \text{ grams.}$$

4.5.2 Interpretation of the Standard Deviation. The standard deviation (σ or s) has not a simple interpretation like the arithmetic mean (μ or $\bar{x}$) that is interpreted as the balancing point for the distribution. The standard deviation is a very important concept that serves as a basic measure of variability. A smaller value of the standard deviation indicates that most of the observations in a data set are close to the mean while a large value implies that the observations are scattered widely about the mean. However, a connection between the standard deviation and fraction of data included in intervals constructed around the mean, was discovered by the Russian mathematician P.L. Chebyshev (1821–1894). This result, generally known as *Chebyshev's rule*, is stated below:

"For any set of data, the interval $\bar{x}-ks$ to $\bar{x}+ks$, where k is any number greater than 1, contains *at least* the fraction $\left(1-\frac{1}{k^2}\right)$ of the data." For example, the intervals $\bar{x}\pm 2s$ and $\bar{x}\pm 3s$ will contain respectively *at least* the fractions $\left(1-\frac{1}{2^2}\right)$, *i.e.* $\frac{3}{4}$ and $\left(1-\frac{1}{3^2}\right)$ *i.e.* $\frac{8}{9}$ of the data.

This rule is applied to any distribution (Population or Sample) and guarantees the inclusion of a minimum fraction of the data in the constructed interval whereas the actual fraction of the data included (especially in bell-shaped distributions) will exceed $\left(1-\frac{1}{k^2}\right)$.

4.5.3 Co-efficient of Variation. The variability of two or more than two sets of data cannot be compared unless we have a relative measure of dispersion. For this purpose, Karl Pearson (1857–1936) introduced a relative measure of variation, known as the *co-efficient of variation*, abbreviated C.V. which expresses the standard deviation as a percentage of the arithmetic mean of a data set. Symbolically, it is defined as

$$\text{C.V.} = \frac{S}{\bar{x}} \times 100, \text{ for sample data,}$$

$$= \frac{\sigma}{\mu} \times 100, \text{ for population data.}$$

As the coefficient of variation is a pure number without units, it is therefore used to compare the variation in two or more data sets or distributions that are measured in different units, e.g. one may be measured in hours and the other in kilograms or rupees. A large value of C.V. indicates that the variability is great and a small value of C.V. indicates less variability.

The coefficient of variation is also used to compare the performance of two candidates or of two players given their scores in various papers or games, the smaller the coefficient of variation the more consistent is the performance of the candidates or players. Thus it is used as a criterion for the consistent performance of the candidates or the players. It should be noted that this co-efficient is unreliable when the arithmetic mean is very small.

Example 4.9 Using the co-efficient of variation, determine whether or not there is greater variation among the prices of certain similar commodities given, than among the life in hours under test.

Price in Rupees: 8, 13, 18, 23, 30
Life in hours: 130, 150, 180, 250, 345

We have to compute the mean and the standard deviation for each set so that the corresponding coefficient of variation can be obtained. The necessary arithmetic is shown below:

Price in Rupees (X)		Life in hours (Y)	
X	X^2	Y	Y^2
8	64	130	16900
13	169	150	22500
18	324	180	32400
23	529	250	62500
30	900	345	119025
92	1986	1055	253325

Price of Commodities

$$\bar{X} = \text{Rs. } \frac{92}{5} = \text{Rs. } 18.4$$

$$S_X = \sqrt{\frac{1986}{5} - \left(\frac{92}{5}\right)^2}$$

$$= \sqrt{397.2 - 338.56}$$

$$= \sqrt{58.44} = \text{Rs. } 7.66$$

$$\therefore \text{C.V.} = \frac{7.66}{18.4} \times 100 = 41.63\%$$

Life in Hours

$$\bar{Y} = \frac{1055}{5} = 211 \text{ hours}$$

$$S_Y = \sqrt{\frac{253325}{5} - \left(\frac{1055}{5}\right)^2}$$

$$= \sqrt{50665 - 44521}$$

$$= \sqrt{6144} = 78.38 \text{ hours}$$

$$\therefore \text{C.V.} = \frac{78.38}{211} \times 100 = 37.15\%$$

We see that the co-efficient of variation for the prices of commodities (X) is larger than that for the life in hours (Y). Hence the prices of certain similar commodities are showing greater variation than that among the life in hours under test.

Example 4.10 Goals scored by two teams A and B in a football season were as follows:

No. of goals scored in a match (x_i)	Number of matches	
	A	B
0	27	17
1	9	9
2	8	6
3	5	5
4	4	3

By calculating the co-efficient of variation in each case, find which team may be considered more consistent. (P.U., B.Com.)

The necessary arithmetic is shown below:

No. of goals	Team A			Team B		
(x_i)	f_i	f_ix_i	$f_ix_i^2$	f_j	f_jx_i	$f_jx_i^2$
0	27	0	0	17	0	0
1	9	9	9	9	9	9
2	8	16	32	6	12	24
3	5	15	45	5	15	45
4	4	16	64	3	12	48
Total	53	56	150	40	48	126

Team A:

$$\text{Mean} = \frac{\Sigma f_ix_i}{n} = \frac{56}{53} = 1.06, \text{ and}$$

$$s = \sqrt{\frac{\Sigma f_ix_i^2}{n} - \left(\frac{\Sigma f_ix_i}{n}\right)^2}$$

$$= \sqrt{\frac{150}{53} - \left(\frac{56}{53}\right)^2} = \sqrt{1.7138} = 1.308$$

$$\therefore \quad \text{C.V.} = \frac{s}{\bar{x}} \times 100 = \frac{1.308}{1.06} \times 100 = 123.4\%$$

Team B:

$$\text{Mean} = \frac{\sum f_i x_i}{n} = \frac{48}{40} = 1.20, \text{ and}$$

$$s = \sqrt{\frac{\sum f_i x_i^2}{n} - \left(\frac{\sum f_i x_i}{n}\right)^2}$$

$$= \sqrt{\frac{126}{40} - \left(\frac{48}{40}\right)^2} = \sqrt{1.71} = 1.308$$

$$\text{Thus} \quad \text{C.V.} = \frac{s}{\bar{x}} \times 100 = \frac{1.308}{1.20} \times 100 = 109.0\%$$

We see that the co-efficient of variation for the team B is smaller than that for the team A. Hence team B is more consistent than team A.

4.5.4 Properties of Variance and Standard Deviation. The variance and standard deviation have the following useful and interesting properties:

(i) The variance of a constant is equal to zero. If a is any constant, then

$$\text{Var}(a) = \frac{1}{N}\sum [a - a]^2 \quad (\because \text{mean of a constant is constant itself})$$

$$= 0$$

(ii) The variance is independent of the origin, *i.e.* it remains unchanged when a constant is added to or subtracted from each observation of the variable X. Symbolically,

$$\text{Var}(X + a) = \text{Var}(X)$$

$$\text{Now} \quad \text{Var}(X + a) = \frac{1}{N}\sum [(x_i + a) - (\mu + a)]^2 \quad \left(\because \frac{\sum (x_i + a)}{N} = \mu + a\right)$$

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

Hence Var(X) is *invariant* to change of the origin.

(iii) The variance is multiplied or divided by the square of the constant, when each observation of the variable X is either multiplied or divided by a constant.

$$\text{Var}(aX) = \frac{1}{N}\sum (ax_i - a\mu)^2$$

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

This may also be interpreted as that the variance increases by a^2 when the *scale* of X is changed by a.

(iv) The variance of the sum or difference of two *independent* variables is equal to the sum of their respective variances.

If X and Y are two *independent* variables, then

$$\text{Var}(X \pm Y) = \frac{1}{N}\sum [(x_i \pm y_i) - (\mu_{x+y})]^2$$

$$= \frac{1}{N}\sum [(x_i - \mu_x) \pm (y_i - \mu_y)]^2$$

$$= \frac{1}{N}\sum (x_i-\mu_x)^2 + \frac{1}{N}\sum(y_i-\mu_y)^2 \pm \frac{2}{N}\sum(x_i-\mu_x)(y_i-\mu_y)$$

The quantity $\frac{1}{N}\sum (x_i - \mu_x)(y_i - \mu_y)$ is called the *covariance* and is denoted by Cov (X, Y). We shall show at some later stage that the covariance of two *independent* variables is equal to zero. Thus we are left with

$$\text{Var}(X \pm Y) = \frac{1}{N}\sum (x_i - \mu_x)^2 + \frac{1}{N}\sum (y_i - \mu_y)^2$$

$$= \text{Var}(X) + \text{Var}(Y).$$

(v) If k subgroups of data consisting of $N_1, N_2, \ldots, N_k$ $(\sum N_i = N)$ observations have respective means $\mu_1, \mu_2, \ldots, \mu_k$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$, then the variance σ^2 of the combined observations is given by

$$\sigma^2 = \frac{1}{N}\sum N_i (\sigma_i^2 + D_i^2), \quad i = 1, 2, \ldots, k$$

where $D_i = \mu_i - \mu$ and μ is the mean for all the data.

Let for the *ith* subgroup with mean μ_i, μ, the general mean, be considered as an arbitrary origin. Then the sum of squares of deviations of the observations in the *ith* subgroup from μ is given by

$$\sum_1^{N_i} (x_i - \mu)^2 = \sum_1^{N_i} (x_i - \mu_i + \mu_i - \mu)^2$$

$$= \sum_1^{N_i} (x_i-\mu_i)^2 + N_i(\mu_i-\mu)^2, \quad (\because \text{product term vanishes})$$

$$= N_i \sigma_i^2 + N_i D_i^2$$

$$= N_i (\sigma_i^2 + D_i^2).$$

But the variance σ^2 of the combined observations is the mean of the *sum of the deviations of all observations in k subgroups from the general mean* μ. Hence summing over k-subgroups, we get

$$N\sigma^2 = \sum N_i (\sigma_i^2 + D_i^2)$$

It is relevant to note that all these properties are valid for standard deviation (*S.D*), which is the positive square root of variance. In other words,

(i) $S.D.\ (a) = 0.$

(ii) $S.D.\ (X + a) = S.D.\ (X)$

(iii) $S.D.\ (aX) = |a|\ S.D.\ (X)$, as *S.D.* cannot be negative.

(iv) $S.D.\ (X \pm Y) = \sqrt{\text{Var}\ (X) + \text{Var}\ (Y)}$

(v) $\sigma = \sqrt{\frac{1}{N}\sum N_i (\sigma_i^2 + D_i^2)}$

For sample data, the corresponding results may be obtained in the same way.

Example 4.11 Let $\bar{x}_1$ and S_1^2 be the mean and variance respectively of n_1 observations, $\bar{x}_2$ and S_2^2 be the mean and variance respecitvely of n_2 observations. Then, if the variance of all $n_1 + n_2$ observations is S^2, prove that

$$S^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2 \text{(P.U, B.A./B.Sc. 1986)}$$

Let $\bar{x}$ denote the general mean and be regarded as an arbitrary origin for the set of n_1 observations and the set of n_2 observations. Then the variance of all $n_1 + n_2$ observations, by definition, is given by

$$S^2 = \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} (x_i - \bar{x})^2$$

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

$$= \frac{1}{n_1 + n_2} \left[\sum_{i=1}^{n_1} (x_i - \bar{x}_1 + \bar{x}_1 - \bar{x})^2 + \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2 + \bar{x}_2 - \bar{x})^2 \right]$$

$$= \frac{1}{n_1 + n_2} [n_1 \{S_1^2 + (\bar{x}_1 - \bar{x})^2\} + n_2 \{S_2^2 + (\bar{x}_2 - \bar{x})^2\}]$$

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

Since $\bar{x}$ is the mean of all $n_1 + n_2$ observations, *i.e.* $\bar{x} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$,

therefore $\bar{x}_1 - \bar{x} = \bar{x}_1 - \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \dfrac{n_2(\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}$, and

$$\bar{x}_2 - \bar{x} = \bar{x}_2 - \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{-n_1(\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}.$$

Hence substituting and simplifying, we get

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

4.5.5 Standardized Variables. A variable is defined to be *Standardized* or in *standard units* if it is expressed in terms of deviations from its mean and divided by its standard deviation. It is denoted by Z. In symbols, this means that

$$z_i = \frac{x_i - \mu}{\sigma}, \quad \text{for population data,}$$

$$z_i = \frac{x_i - \bar{x}}{S}, \quad \text{for sample data.}$$

This is a very important concept in advanced statistics as the mean of a standardized variable is equal to zero and its variance is equal to one. Thus

$$\bar{Z} = \frac{1}{N}\Sigma\left(\frac{x_i - \mu}{\sigma}\right) = \frac{1}{\sigma}\frac{\Sigma(x_i - \mu)}{N} = 0;$$

and
$$\text{Var}(Z) = \frac{1}{N}\Sigma\left[\frac{(x_i - \mu)}{\sigma} - 0\right]^2$$

$$= \frac{1}{\sigma^2}\frac{\Sigma(x_i - \mu)^2}{N} = \frac{1}{\sigma^2}\cdot\sigma^2 = 1$$

The Z-values, being independent of the units of measurement, provide a basis for comparison between individual values, even though they belong to different distributions. That is why they are often used in psychological and educational testing, where they are known as *standard scores*. The negative numbers are avoided by multiplying the Z values by 10, an arbitrary *S.D.*, and adding 50, an arbitrary mean, to them. The values so obtained are called the *standard Z scores*. Thus a standard Z score is given by the relation

$$Z = 50 + 10\left(\frac{x - \bar{x}}{S}\right)$$

4.6 TRIMMED AND WINSORIZED MEASURES

Data sets often contain *extreme* (unusually large or small) observations which may be very different from the main body of the data set and may seem to be incorrect. Such extreme observations are generally called *Wild observations* or *Outliers*. These outliers can cause problems. In the presence of outliers, the mean and the standard deviation, being affected by the extreme observations, are therefore misleading measures of central tendency and variability. The appropriate measures then may be the Median and the Interquartile Range, which are much less sensitive to extreme or wild observations. However, it is important to examine a data set for outliers and if present, should be excluded.

For this purpose, we *either* remove a certain percentage of the smallest and largest observations to get the so-called *Trimmed data* set *or* replace trimmed values by those next in magnitude to obtain what is known as *Winsorized data* set (proposed by C.P. Winsor). The mean and the standard deviation of such data sets are known as the *Trimmed Mean* and the *Trimmed Standard Deviation*, and the *Winsorized Mean* and the *Winsorized Standard Deviation* respectively.

Generally, the *Trimmed* mean is obtained from the data set after having removed all observations below the first quartile and all observations above the third quartile. The *Winsorized* mean is calculated from the modified data set obtained by replacing each observation below the first quartile with the value of the first quartile and each observation above the third quartile with the value of the third quartile. The *Trimmed* standard deviation and the *Winsorized* standard deviation are computed from the *trimmed* data set and *Winsorized* data set as usual. The *trimmed* and *Winsorized* measures have gained importance in recent years as they are not disturbed by the presence of a few wild observations and have been found almost as good as the corresponding measures in symmetric distributions with no unusual observations.

Example 4.12 Calculate the trimmed and Winsorized means and standard deviations for the data given in Example 3.11.

The data ordered from smallest to largest and the two quartiles were found to be

32, 36, 36, 37, 39, 41, 45, 46, 48

↑ ↑

Q_1 Q_3

To find the trimmed mean and the trimmed standard deviation, we remove the two observations 32 and 36 below the first quartile and the two observations 46 and 48 above the third quartile. Thus we have five observations 36, 37, 39, 41, 45 as trimmed data set.

$$\therefore \quad \text{Trimmed mean} = \frac{36 + 37 + 39 + 41 + 45}{5} = \frac{198}{5} = 39.6, \text{ and}$$

$$\text{Trimmed } S.D. = \sqrt{\frac{(36)^2 + \ldots + (45)^2}{5} - \left(\frac{198}{5}\right)^2}$$

$$= \sqrt{1598.4 - 1568.16} = \sqrt{30.24} = 5.5$$

To find the Winsorized mean and standard deviation, we replace the two values 32, 36 below the first quartile with 36, and the two values 46, 48 above Q_3 with 45 to get the Winsorized data set as 36, 36, 36, 37, 39, 41, 45, 45 and 45. Thus

$$\text{the Winsorized mean} = \frac{\sum X_i}{n} = \frac{360}{9} = 40, \text{ and}$$

$$\text{the Winsorized } S.D. = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{14534}{9} - \left(\frac{360}{9}\right)^2}$$

$$= \sqrt{1614.89 - 1600} = \sqrt{14.89} = 3.86$$

4.7 MOMENTS

A *moment* designates the power to which deviations are raised before averaging them, *e.g.* the quantity $\frac{1}{N}\sum(x_i - \mu)$ is called the first population moment and is denoted by μ_1. Similarly, the quantity $\frac{1}{N}\sum(x_i - \mu)^2$ is called the second population moment and is denoted by μ_2. The corresponding sample moments are denoted by m_1 and m_2. In general, the *rth moment about the mean* is the arithmetic mean of the *rth* power of the deviations of the observations from the mean. In symbols, this means that

$$\mu_r = \frac{1}{N}\sum (x_i - \mu)^r, \text{ for population data.}$$

$$m_r = \frac{1}{n}\sum (x_i - \bar{x})^r, \text{ for sample data.}$$

These moments are also called the *central moments* or the *mean moments* and are used to describe a set of data.

In a similar way, *moments about an arbitrary origin*, say a, are defined by the relation

$$\mu'_r = \frac{1}{N}\sum (x_i - a)^r, \text{ for population data.}$$

$$m'_r = \frac{1}{n}\sum (x_i - a)^r, \text{ for sample data.}$$

Now, if we put $r = 0$, we see that

$$\mu_0 = \mu'_0 = 1, \text{ and } m_0 = m'_0 = 1$$

For $r = 1$, we have

$$\mu_1 = \frac{1}{N}\sum (x_i - \mu) = \frac{\sum x_i}{N} - \mu = \mu - \mu = 0, \text{ and}$$

$$\mu_1' = \frac{1}{N}\sum (x_i - a) = \frac{\sum x_i}{N} - a = \mu - a.$$

The corresponding sample results are $m_1 = 0$ and $m_1' = \bar{x} - a$.

Putting $r = 2$ in the relation for mean moments, we see that

$$\mu_2 = \frac{1}{N}\sum (x_i - \mu)^2 = \sigma^2, \text{ which is the population variance,}$$

and $$m_2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = S^2, \text{ which is the sample variance,}$$

When $a = 0$, the moment $m'_r = \frac{1}{n}\sum x_i^r$ is called the *rth moment about zero.*

The moments about the mean or about the arbitrary origin are also called the *power moments*.

When the sample data are grouped into a frequency distribution having k classes with midpoints $x_1, x_2, \ldots, x_k$ and the corresponding frequencies $f_1, f_2, \ldots, f_k$ ($\sum f_i = n$), the *rth* sample moments are given by

$$m_r = \frac{1}{n}\sum f_i (x_i - \bar{x})^r, \text{ and}$$

$$m'_r = \frac{1}{n}\sum f_i (x_i - a)^r.$$

4.7.1 Moments about the Mean in terms of Moments about any arbitrary origin, say a, and conversely. It is easier to calculate the moments in the first instance, about an arbitrary origin. They are then transformed to the mean-moments. This is done by using the relationships obtained as follows.

By definition, the *rth* sample moment about the mean is given by

$$m_r = \frac{1}{n}\sum f_i(x_i - \bar{x})^r$$

The quantity within brackets may be written as

$$(x_i - \bar{x}) = (x_i - a + a - \bar{x}) = (x_i - a) - (\bar{x} - a)$$

$$= D_i - m_1' \text{ where } D_i = (x_i - a) \text{ and } m_1' = (\bar{x} - a)$$

Thus, we have $m_r = \frac{1}{n}\sum f_i(D_i - m_1')^r$

By means of Binomial expansion, we have

$$m_r = \frac{1}{n}\sum f_i \left[D_i^r - \binom{r}{1} D_i^{r-1} m'_1 + \binom{r}{2} D_i^{r-2}(m'_1)^2 + \ldots + (-1)^r (m'_1)^r\right]$$

where $\binom{r}{j} = \frac{r!}{j!\,(r-j)!}$ and $r! = r(r-1)(r-2)\ldots 3 \times 2 \times 1$.

i.e. $$m_r = \frac{1}{n}\sum f_i D_i^r - \binom{r}{1}\frac{1}{n}\sum f_i D_i^{r-1} m'_1 + \binom{r}{2}\frac{1}{n}\sum f_i D_i^{r-2}(m'_1)^2 +$$

$$\ldots + (-1)^r (m'_1)^r \frac{1}{n}\sum f_i$$

$$= m'_r - \binom{r}{1} m'_{r-1} m'_1 + \binom{r}{2} m'_{r-2} (m'_1)^2 + \ldots + (-1)^r (m'_1)^r$$

Putting $r = 1, 2, 3$ and 4, we get

$$m_1 = m'_1 - m'_1 = 0;$$

$$m_2 = m'_2 - \binom{2}{1} m'_1 . m'_1 + \binom{2}{2} (m'_1)^2 . m'_0$$

$$= m'_2 - (m'_1)^2;$$

$$m_3 = m'_3 - 3\, m'_2\, m'_1 + 2\,(m'_1)^3, \text{ and}$$

$$m_4 = m'_4 - 4\, m'_3\, m'_1 + 6 m'_2\,(m'_1)^2 - 3\,(m'_1)^4$$

The corresponding results for population data are:

$$\mu_1 = 0;$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2;$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\,(\mu'_1)^3; \text{ and}$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\,(\mu'_1)^2 - 3\,(\mu'_1)^4.$$

It should be noted that in each of these relations, the sum of the coefficients of various terms on the right hand side equals zero and each term on the right is of the same dimension as the term on the left.

Conversely, the *rth* sample moment about an arbitrary origin, a, is given by

$$m'_r = \frac{1}{n}\sum f_i (x_i - a)^r$$

$$= \frac{1}{n}\sum f_i (x_i - \bar{x} + \bar{x} - a)^r$$

$$= \frac{1}{n}\sum f_i (d_i + m'_1)^r, \text{ where } d_i = x_i - \bar{x} \text{ and } m'_1 = \bar{x} - a$$

$$= \frac{1}{n}\sum f_i d_i^r + \binom{r}{1} m'_1 \frac{1}{n}\sum f_i d_i^{r-1} + \binom{r}{2} (m'_1)^2 \times$$

$$\frac{1}{n}\sum f_i d_i^{r-2} + \ldots + (m'_1)^r \frac{1}{n}\sum f_i$$

$$= m_r + \binom{r}{1} m_{r-1} (m'_1) + \binom{r}{2} m_{r-2} (m'_1)^2 + \ldots + (m'_1)^r$$

Putting $r = 2$, 3 and 4, we get

$$m'_2 = m_2 + (m'_1)^2$$

$$m'_3 = m_3 + 3 m_2 (m'_1) + (m'_1)^3$$

$$m'_4 = m_4 + 4 m_3 (m'_1) + 6 m_2 (m'_1)^2 + (m'_1)^4.$$

For a population of size N, the corresponding relations are

$$\mu'_2 = \mu_2 + \mu'_1{}^2.$$

$$\mu'_3 = \mu_3 + 3\mu'_1 \mu_2 + \mu'_1{}^3;$$

$$\mu'_4 = \mu_4 + 4\mu'_1 \mu_3 + 6\mu'_1{}^2 \mu_2 + \mu'_1{}^4.$$

4.7.2 Sheppard's Corrections. In the calculation of moments from a grouped frequency distribution, certain errors are introduced by the assumption that the frequencies associated with a class are located at the midpoint of the class interval. These errors therefore need corrections. It has been shown by W.F. Sheppard that, if the frequency distribution (i) is continuous and (ii) tails off to zero at each end, the corrected moments are as given below:

$$m_2 \text{ (corrected)} = m_2 \text{ (uncorrected)} - \frac{h^2}{12};$$

$$m_3 \text{ (corrected)} = m_3 \text{ (uncorrected)};$$

$$m_4 \text{ (corrected)} = m_4 \text{ (uncorrected)} - \frac{h^2}{2} \cdot m_2 \text{ (uncorrected)} + \frac{7}{240} \cdot h^4;$$

where h denotes the uniform class-interval and m's are the moments about the mean of the grouped frequency distribution. The important point to note here is that these corrections are not applicable to highly skewed distributions and distributions having unequal class-intervals.

4.7.3 Moment-Ratios. There are certain ratios in which both the numerators and the denominators are moments. The most common of these moment-ratios are β_1 and β_2 defined by the relations, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\beta_2 = \frac{\mu_4}{\mu_2^2}$. They are independent of origin and units of measurement, *i.e.* they are pure numbers. Actually, β_1 is the square of the third population moment expressed in *standard* units and β_2 is the fourth standardized moment for a population, where a standardized variable has been defined as

$$Z = (x - \mu)/\sigma.$$

For symmetrical distributions, β_1 is equal to zero. It is, therefore, used as a measure of skewness. β_2 is used to explain the shape of the curve and is a measure of peakedness. For the normal distribution to be discussed later, $\beta_2 = 3$.

The moment-ratios (or the standardized moments) for sample data are similarly defined as

$$b_1 = \frac{(m_3)^2}{(m_2)^3} \text{ and } b_2 = \frac{m_4}{(m_2)^2}$$

Example 4.13 Calculate the first four moments about the mean for the following set of examination marks: 45,32,37,46,39,36,41,48 & 36.

For convenience, the observed values are written in an increasing sequence. The necessary calculations appear in the table below:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
32	−8	64	−512	4096
36	−4	16	−64	256
36	−4	16	−64	256
37	−3	9	−27	81
39	−1	1	−1	1
41	1	1	1	1
45	5	25	125	625
46	6	36	216	1296
48	8	64	512	4096
360	0	232	186	10708

Now $\bar{x} = \frac{\sum x_i}{n} = \frac{360}{9} = 40$ marks

Therefore $m_1 = \frac{\sum (x_i - \bar{x})}{n} = 0$

$$m_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{232}{9} = 25.78 \text{ (marks)}^2$$

$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n} = \frac{186}{9} = 20.67 \text{ (marks)}^3$$

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{n} = \frac{10708}{9} = 1189.78 \text{ (marks)}^4$$

Example 4.14 Compute the first four moments for the following distribution of wages after applying Sheppard's corrections.

Weekly Earnings (Rupees)	5	6	7	8	9	10	11	12	13	14	15
No. of men	1	2	5	10	20	51	22	11	5	3	1

(P.U., B.A./B.Sc. (Part I), 1962)

We first calculate moments about an arbitrary origin. The necessary calculations are shown below. The moments about $x = 10$ are obtained by dividing the column sums by n.

Earnings in Rs. (x_i)	Men f_i	D_i $(x_i - 10)$	f_iD_i	$f_iD_i^2$	$f_iD_i^3$	$f_iD_i^4$
5	1	−5	−5	25	−125	625
6	2	−4	−8	32	−128	512
7	5	−3	−15	45	−135	405
8	10	−2	−20	40	−80	160
9	20	−1	−20	20	−20	20
10	51	0	−68	0	−488	0
11	22	1	22	22	22	22
12	11	2	22	44	88	176
13	5	3	15	45	135	405
14	3	4	12	48	192	768
15	1	5	5	25	125	625
Sums	131	--	+76 +8	346	+562 +74	3718
Sums ÷ n	1	--	0.06 $= m'_1$	2.64 $= m'_2$	0.56 $= m'_3$	28.38 $= m'_4$

Moments about the mean are:

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2 = 2.64 - (0.06)^2 = 2.64;$$

$$m_3 = m'_3 - 3m'_2m'_1 + 2(m'_1)^3$$

$$= 0.56 - 3(2.64)\,(0.06) + 2(0.06)^3 = 0.08;$$

$$m_4 = m'_4 - 4m'_3m'_1 + 6m'_2(m'_1)^2 - 3(m'_1)^4$$

$$= 28.38 - 4(0.56)\,(0.06) + 6(2.64)\,(0.06)^2 - 3(0.06)^4 = 28.30$$

Applying Sheppard's corrections, we have

$$m_2 \text{ (corrected)} = m_2 \text{ (uncorrected)} - \frac{h^2}{12} = 2.64 - 0.08 = 2.56,$$

$$m_3 \text{ (corrected)} = m_3 \text{ (uncorrected)} = 0.08,$$

$$m_4 \text{ (corrected)} = m_4 \text{ (uncorrected)} - \frac{h^2}{2}.m_2\text{(uncorrected)} + \frac{7h^4}{240}$$

$$= 28.30 - 1.32 + 0.03 = 27.01$$

4.7.4 Change of Origin and Scale. Let a and h denote the arbitrary origin and the class-interval. Then we define a new variable u as

$$u_i = \frac{x_i - a}{h}$$

so that $x_i - a = hu_i$; $\bar{x} - a = h\bar{u}$ and hence $x_i - \bar{x} = h(u_i - \bar{u})$.

Substituting these values in the *rth* sample moments, we get

$$m'_r = \frac{1}{n}\sum f_i\,(x_i - a)^r = h^r \cdot \frac{1}{n}\sum f_i u_i^r;$$

and

$$m_r = \frac{1}{n}\sum f_i\,(x_i - \bar{x})^r = h^r \cdot \frac{1}{n}\sum f_i\,(u_i - \bar{u})^r.$$

This shows that the *rth* moments of the variable X are h^r times the corresponding moments of the variable u, and are independent of the origin 'a'. In other words, the *moments are not affected by a change of origin but are affected by a change of scale.*

4.7.5 Charlier Check. We have seen that the computation of the moments depends upon the sum of the products of the frequencies by the corresponding values of the variable. It is, therefore, desirable to check these computations so that arithmetic mistakes, if any, are avoided. For this purpose, L.V. Charlier, the Norwegian statistician, introduced a check known as *Charlier check*. This check actually consists

in shifting the assumed origin in the *coded* form by one interval. The relations used for this purpose are given below:

$$\Sigma f(u+1) = \Sigma fu + n$$

$$\Sigma f(u+1)^2 = \Sigma fu^2 + 2\Sigma fu + n$$

$$\Sigma f(u+1)^3 = \Sigma fu^3 + 3\Sigma fu^2 + 3\Sigma fu + n$$

$$\Sigma f(u+1)^4 = \Sigma fu^4 + 4\Sigma fu^3 + 6\Sigma fu^2 + 4\Sigma fu + n$$

Example 4.15 Calculate the first four moments about the mean from the data of Example 4.4.

The necessary calculations by taking $u_i = \dfrac{x_i - 114.5}{20}$, are set out in the following table. The last column is used for *Charlier's check* and the column sums are divided by n to get m'_r.

Data		Computations					
x_i	f_i	u	fu	fu^2	fu^3	fu^4	$f(u+1)^4$
74.5	9	−2	−18	36	−72	144	9
94.5	10	−1	−10	10	−10	10	0
114.5	17	0	0	0	0	0	17
134.5	10	1	10	10	10	10	160
154.5	5	2	10	20	40	80	405
174.5	4	3	12	36	108	324	1024
194.5	5	4	20	80	320	1280	3125
Sum	60	--	24	192	396	1848	4740
Sums ÷n	1	--	0.4 $= m'_1$	3.2 $= m'_2$	6.6 $= m'_3$	30.8 $= m'_4$	For Charlier's check

Charlier's check.

$$\Sigma f(u+1)^4 = \Sigma fu^4 + 4\Sigma fu^3 + 6\Sigma fu^2 + 4\Sigma fu + n$$

$$= 1848 + 4(396) + 6(192) + 4(24) + 60$$

$$= 1848 + 1584 + 1152 + 96 + 60 = 4740,$$ which is the sum in the last column.

Hence the moments about the mean and in *class-interval units* are obtained as below:

$$m_1 = 0$$

$$m_2 = m'_2 - (m'_1)^2$$
$$= 3.2 - (0.4)^2 = 3.04$$
$$m_3 = m'_3 - 3m'_2m'_1 + 2(m'_1)^3$$
$$= 6.6 - 3(3.2)(0.4) + 2(0.4)^3 = 2.89$$
$$m_4 = m'_4 - 4m'_3m'_1 + 6m'_2(m'_1)^2 - 3(m'_1)^4$$
$$= 30.8 - 4(6.6)(0.4) + 6(3.2)(0.4)^2 - 3(0.4)^4 = 23.24$$

To get the moments about the mean in *ordinary units*, we multiply m_2 by h^2, *i.e.* 400, m_3 by $(20)^3$ and m_4 by $(20)^4$. Thus $m_2 = 1216$, $m_3 = 23120$ and $m_4 = 3718400$.

Example 4.16 The first three moments of a distribution about the value 2 of the variable are 1, 16 and −40. Show that the mean is 3, the variance 15 and $m_3 = -86$. Also show that the first three moments about $x = 0$ are 3, 24 and 76.

Here we are given $m'_1 = \frac{1}{n}\sum f(x-2) = 1$ ... (1)

$$m'_2 = \frac{1}{n}\sum f(x-2)^2 = 16 \quad \text{... (2)}$$

$$m'_3 = \frac{1}{n}\sum f(x-2)^3 = -40 \quad \text{... (3)}$$

We also know that $m'_1 = \bar{x} - a$, so that

$$\bar{x} = m'_1 + a = 1 + 2 = 3 \qquad (\because a = 2)$$

and variance, $S^2 = m_2$ (second moment about mean)

$$= m'_2 - (m'_1)^2 = 16 - 1 = 15, \text{ and}$$
$$m_3 = m'_3 - 3m'_2m'_1 + 2(m'_1)^3$$
$$= -40 - 3(16)(1) + 2(1)^3 = -86.$$

To find the moments about $x = 0$, we need the values of $\frac{1}{n}\sum fx$, $\frac{1}{n}\sum fx^2$ and $\frac{1}{n}\sum fx^3$, which are obtained from relations (1), (2) and (3).

From (1), $\frac{\sum fx}{n} = 3$

From (2), $\frac{1}{n}\sum f(x-2)^2 = 16$

or $$\frac{1}{n}\sum f(x^2 - 4x + 4) = 16$$

or $$\frac{1}{n}\sum fx^2 - 4\frac{\sum fx}{n} + 4 = 16$$

or $$\frac{1}{n}\sum fx^2 = 16 - 4 + 4(3) = 24$$

(3) on expansion can be written as

$$\frac{1}{n}\sum fx^3 - 6\frac{1}{n}\sum fx^2 + 12\frac{\sum fx}{n} - 8 = -40$$

or $$\frac{1}{n}\sum fx^3 = -40 + 8 - 12(3) + 6(24) = 76$$

Hence the moments about $x = 0$ are

$$m'_1 = \frac{\sum fx}{n} = 3,$$

$$m'_2 = \frac{\sum fx^2}{n} = 24, \text{ and}$$

$$m'_3 = \frac{\sum fx^3}{n} = 76.$$

Example 4.17 Show that for discrete distributions, $\beta_2 > 1$.

By definition, $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$. (P.U., D. St. 1962)

Now β_2 will be greater than one if the numerator is greater than the denominator, i.e. if $\mu_4 > \mu_2^2$, or if $\mu_4 - \mu_2^2 > 0$.

Now $$\mu_4 - \mu_2^2 = \frac{1}{N}\sum f(x-\mu)^4 - \sigma^4 \quad (\because \mu_2 = \sigma^2)$$

$$= \frac{1}{N}\sum f(x-\mu)^4 + \sigma^4 - 2\sigma^4$$

$$= \frac{1}{N}\sum f(x-\mu)^4 + \frac{\sigma^4\sum f}{N} - 2\sigma^2 \cdot \frac{\sum f(x-\mu)^2}{N},$$

$$\left[\sigma^2 = \frac{\sum f(x-\mu)^2}{N}\right]$$

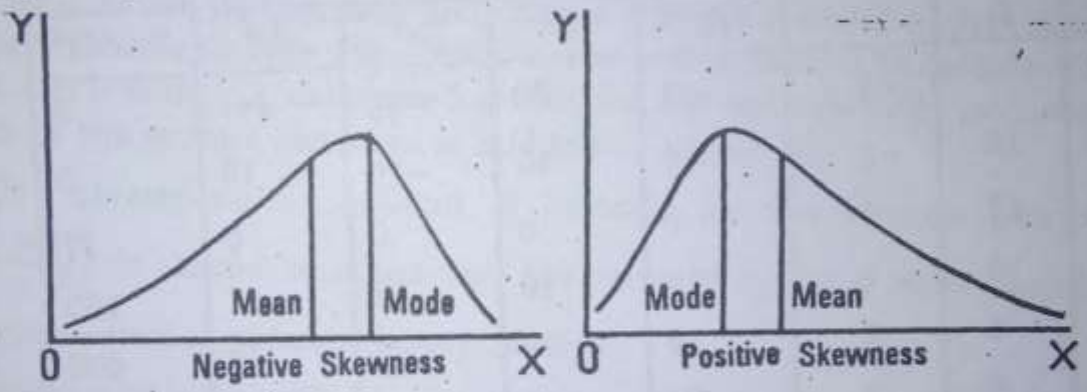
$$= \frac{1}{N}\sum f[(x-\mu)^4 + \sigma^4 - 2\sigma^2(x-\mu)^2] \quad (N\sigma^4 = \sigma^4\textstyle\sum f)$$

$$= \frac{1}{N}\sum f[(x-\mu)^2 - \sigma^2]^2 \text{ which is essentially positive.}$$

Hence $\beta_2 \geq 1$ because $\mu_4 - \mu_2^2$ is always positive.

4.8 SKEWNESS

A distribution in which the values equidistant from the mean have equal frequencies is defined to be *symmetrical* and any departure from symmetry is called *skewness*. It is important to note that in a perfectly symmetrical distribution, the mean, median and mode coincide and that the two tails of the distribution are equal in length from the mean. These values are pulled apart when the distribution departs from symmetry and consequently one tail becomes longer than the other. If the right tail is longer than the left tail, the distribution is said to have *positive skewness*. If the left tail of the distribution is longer than its right tail, it is said to be *negatively skewed* or to have *negative skewness*. In a positively skewed distribution, the mean is greater than the median and the median is greater than the mode, *i.e.* mean > median > mode and in a negatively skewed distribution, mode > median > mean.



The *difference* between the measures of location, being an indication of the amount of skewness or asymmetry, is used as a measure of skewness. A measure of skewness is defined in such a way that (i) the measure should be zero when the distribution is symmetric and (ii) the measure should be a pure number, *i.e.* independent of origin and units of measurements.

Accordingly, to measure the degree of skewness of a distribution or curve, Karl Pearson (1857-1936) introduced a coefficient of skewness denoted by Sk and defined by

$$Sk = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

We know that mode is sometimes ill-defined and is difficult to locate by simple methods. It is, therefore, replaced by its equivalent from empirical relation holding good in moderately skewed distributions. The *Pearsonian co-efficient of skewness* then becomes

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

This coefficient usually varies between −3(negative skewness) and +3 (positive skewness) and the sign indicates the direction of skewness. The formula satisfies both the requirements considered essential for a measure of skewness.

Arthur Lyon Bowley (1869-1957), a British statistician, has also proposed a measure of skewness that is based on the median and the two quartiles. In a symmetrical distribution, the two quartiles are equidistant from the median but in an asymmetrical distribution, this will not be the case. The *Bowley's co-efficient of skewness* is

$$Sk = \frac{Q_1 + Q_3 - 2\,\text{Median}}{Q_3 - Q_1}$$

Its value lies between 0 and ±1.

Another measure of skewness that is often used, is the third moment expressed in standard units (or the moment ratio) and thus is given by

$$Sk = \frac{\mu_3}{\sigma^3}, \text{ for population data,}$$

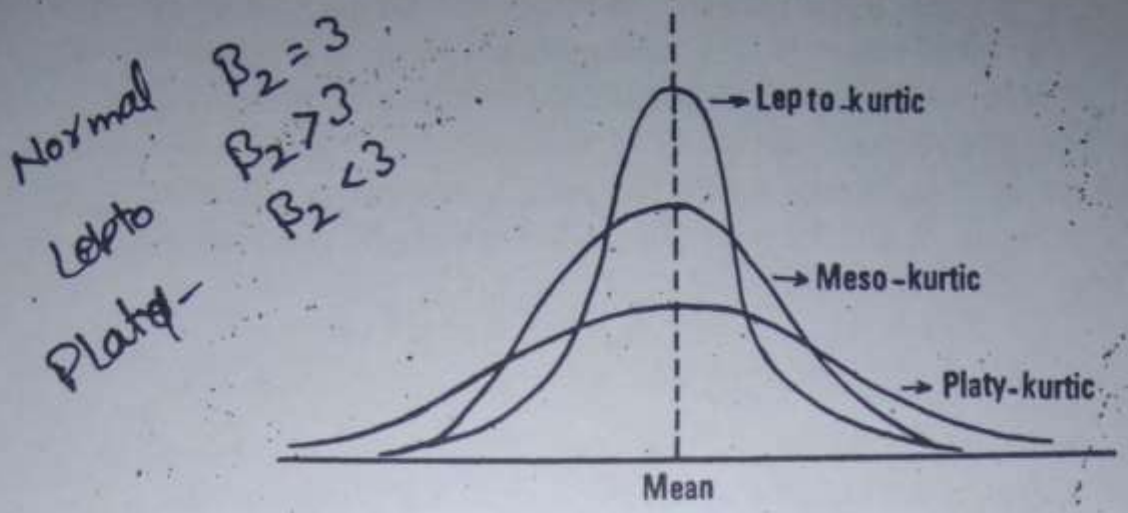
$$= \frac{m_3}{s^3}, \text{ for sample data,}$$

This coefficient for most distributions, will be between −2 and +2. Some statisticians denote it by α_3 or $\sqrt{\beta_1}$. If the coefficient is greater than zero, the distribution or curve is positively skewed. If $Sk < 0$, there is negative shewness. For symmetrical distributions or curves, the coefficient is zero.

4.9 KURTOSIS

Karl Pearson (1857–1936) introduced the term *Kurtosis* (literally the amount of hump) for the degree of *peakedness* or *flatness* of a unimodal frequency curve. When the values of a variable are closely bunched round the mode in such a way that the peak of the curve becomes relatively high, we say that the curve is *leptokurtic*. If, on the other hand, the curve is flat-topped, we say that the curve is *platykurtic*. Since the *normal curve* (to be described later) is neither very peaked nor very flat-topped, it is taken as a basis for comparison. The normal curve itself is called *mesokurtic*.

Kurtosis is usually measured by the fourth standardized moment or the moment-ratio $\beta_2 = (\mu_4/\mu_2^2)$ whose value for a normal distribution is



equal to 3. When β_2 is greater than 3, the curve is more sharply peaked and has wider tails than the normal curve and is said to be *leptokurtic*. When it is less than 3, the curve has a flatter top and relatively narrower tails than the normal curve and is said to be *platykurtic*.

The corresponding measure of kurtosis for the sample data is $b_2 \left(= \frac{m_4}{m_2^2}\right)$. It should be noted that the value of b_2 for a large sample from the *normal population* is very nearly 3.

Another measure of kurtosis not widely used, is given by

$$K = \frac{Q.D.}{P_{90} - P_{10}},$$

where $Q.D.$ is the semi-interquartile range and P's are the *percentiles*. This is known as the *Percentile co-efficient of kurtosis*. It has been shown that K for a normal distribution is 0.263 and that it lies between 0 and 0.50.

4.10 DESCRIBING A FREQUENCY DISTRIBUTION

To describe the major characteristics of a 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 standard deviation 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 the mean is the variance and its positive square root is the standard deviation. The third mean 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 distributions.

EXERCISES

4.1 Explain clearly the meaning of the term Dispersion. What are the most usual methods of measuring dispersion? Indicate the advantages and disadvantages of these methods.

(P.U., B.Com. 1960; B.A. (Hons.), 1960; B.A. (Part I), 1961)

4.2 Discuss the different measures of dispersion. Describe the method of computation of any two of them with suitable examples. (P.U., M.A., Econ. 1969)

4.3 Describe carefully how Mean Deviation, Standard Deviation and Quartile Deviation of any given distribution are obtained. In what problems, should each be used? (P.U., B.A. (Part I), 1962-S)

4.4 (a) What is Range and how is it calculated? What are its uses?

(b) Define Quartile Deviation. Find the quartile deviation from the following data (i) graphically, (ii) using an appropriate formula.

Income per week (Rs.)	41–50	51–60	61–70	71–80	81–90	91–100	Total
No. of Earners	30	36	43	104	73	14	300

(P.U., B.A./B.Sc. 1960)

4.5 The members of a sports club, 60 male adults, had their weights recorded, in pounds. The weights are given below:

171	160	144	132	154	160	160	158	148	160	131	153
131	165	139	163	149	149	140	149	150	161	136	144
165	174	153	149	157	169	147	156	149	171	149	154
153	149	147	154	145	158	160	152	156	138	167	142
165	155	140	155	158	147	149	169	148	174	150	144

Construct a cumulative frequency table for these weights, using classes of width 5 lb, starting at 129.5 lb. Hence draw a cumulative frequency graph, and use this to find the median and semi-interquartile range.