CHAPTER 4

DESCRIPTIVE STATISTICS II: DISPERSION

4.1 MEANING OF DISPERSION

The essential purpose of statistical averages discussed in the preceding chapter is to summarize a large mass of data. These averages serve to locate the 'center' of a distribution but they do not reveal how the items or the observations are spread out or scattered on each side of the center. This latter characteristic of a distribution is variously known as the **dispersion**, **'scatter'**, or **'variation'**. It is just as important to measure this property of a distribution as to locate the central values. If the dispersion is small, it indicates high uniformity of the observations in the distribution. Absence of dispersion in the data indicates perfect uniformity. This situation arises when all observations in the distribution are identical. If this were the case, description of any single observation would suffice. But in reality, it rarely happens.

The presence of any degree of variation among the measures, however, necessitates the use of both the concepts 'central tendency' and 'dispersion', for any precise descriptive summary of the data. With summary statistics of these two concepts–a 'measure of central tendency' and a 'measure of variability'–we can describe almost all distributions with a reasonable degree of accuracy.

The three measures of central tendency, mean, median and mode represent the first of two essential types of descriptive statistics. This chapter concerns the second major group, **measures of dispersion or variability.**

A measure of dispersion appears to serve two purposes:

- First, it is one of the most important quantities used to characterize a frequency distribution.
- Second, it affords a basis of comparison between two or more frequency distributions.

The study of dispersion bears its importance from the fact that different distributions may have exactly the same averages, but substantial differences in variability. We illustrate this point by the following example:

Suppose that three students secured the following marks in an examination

Student	Math	Statistics	Physics	English	Average	Range
1	68	30	70	40	52	40
2	49	50	55	54	52	6
3	51	52	52	53	52	2

The three distributions are certainly not identical though their averages are the same. These differences lie in the dispersion of their scores. The student 1 shows largest variation in his secured scores, while the second student shows relatively less variation than the first. As you can see, the third student's scores are even more close to each other compared to the second. Clearly, the performances of the individual students cannot be evaluated simply on the basis of the arithmetic means. It is the primary objective of this chapter to consider the basic techniques by which this important characteristic of a distribution is measured.

The dispersion of a distribution can more clearly be viewed from its graphical presentation. Consider the frequency curves shown in Figure 4.1, which have been drawn from three different distributions.

As we observe, all the three curves have identical measures of location, i.e. the mean, median and the mode have a common value, but they differ clearly in their variability:

- The curve *A* has the **least** variability;
- The curve *B* has **moderate** variability;
- The curve *C* has the **most** variability.

This once again tends to demonstrate that averages alone cannot always tell about the characteristics of a distribution, and hence we must look for other measures that may help to understand the nature of variations in data.

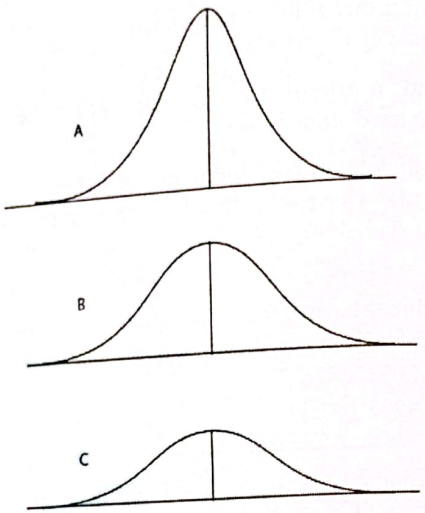


Figure 4.1: Distributions with unequal variability but with identical mean

4.2 MEASURES OF DISPERSION

The measures of dispersion can broadly be classified to fall into one of the two categories: absolute measures and relative measures. The first category of measures includes:

(i) The range
(ii) The quartile deviation
(iii) The mean (or average) deviation
(iv) The variance
(v) The standard deviation

The measures are **absolute** in the sense that they are expressed in the same statistical unit in which the original data are presented, such as dollar, taka, meter, kilogram, etc.

When the two or more data sets are expressed in different units, however, the absolute measures are not comparable, in which case it is necessary to consider some other measures that reduce the absolute deviation in some relative form. These measures are referred to as **relative measures.** The relative measures are usually expressed in the form of **coefficients** and are pure numbers, independent of the unit of measurements. The measures are

(i) Coefficient of range
(ii) Coefficient of quartile deviation
(iii) Coefficient of mean deviation
(iv) Coefficient of variation

4.3 ABSOLUTE MEASURES OF DISPERSION

4.3.1 The Range

The simplest and the crudest measure of dispersion is the **range**. This is defined as the difference between the smallest and the largest values in the distribution. If $x_1, x_2, \ldots\ldots, x_n$ are the values of n observations, then range R of the variable x is given by

$$R(x_1, x_2, \ldots, x_n) = \max(x_1, x_2, \ldots, x_n) - \min(x_1, x_2, \ldots, x_n) \quad \ldots (4.1)$$

In other words, if the x values are arranged in ascending order such that $x_1 < x_2 < \ldots\ldots < x_n$, then

$$R = x_n - x_1 \quad \ldots (4.1a)$$

For a set of observations 90, 110, 20, 51, 210 and 190, say, the smallest value is 20 and the largest value is 210, so that $R = 210 - 20 = 190$. For the age data presented in Table 2.1, $R = 54 - 25 = 29$ years.

For grouped data, the difference between the lower class limit (or boundary) of the lowest class and the higher class limit (or boundary) of the highest class is considered to be the range. Thus the range for the age data presented in Table 2.6 is $54.5 - 24.5 = 30$ years. Obtaining range from grouped distribution is not however recommended for obvious reasons.

Although the range is meaningful, it is of little use because of its marked instability, particularly when the range is based on a small sample. Imagine, if there is one extreme value in a distribution, the range of the values will appear to be large, when in fact, removal of this value may reveal an otherwise compact distribution with extremely low dispersion.

4.3.2 Trimmed Range

Since the range is subject to the undue influence of erratic extreme values, it can be expected that if such values are excluded, the range of remaining items may be a more useful measure. One such measure is the 10 to 90 percentile range, also called trimmed range. It is established by excluding the lowest and the highest 10 percent of the items, and is the difference

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{867}{32} = 27.1$$

Using (4.15) the variance is

$$s^2 = \frac{n\sum f_i x_i^{\,2} - \left(\sum f_i x_i\right)^2}{n(n-1)} = \frac{32(23805) - 867^2}{32 \times 31} = 10.15$$

so that the standard deviation is $s = \sqrt{10.15} = 3.19$

Hence the mean, variance and standard deviation are 27.1, 10.15 and 3.19 respectively.

Example 4.8: Calculate the variance for the age distribution shown in Table 4.2.

Solution: The accompanying table shows the computational steps with the data in Table 4.2.

Age	f_i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$	$f_i x_i$	$f_i x_i^2$
24.5–29.5	3	27	–12.3	151.29	453.87	81	2187
29.5–34.5	9	32	–7.3	53.29	479.61	288	9216
34.5–39.5	15	37	–2.3	5.29	79.35	555	20535
39.5–44.5	12	42	2.7	7.29	87.48	504	21168
44.5–49.5	7	47	7.7	59.29	415.03	329	15463
49.5–54.5	4	52	12.7	161.29	645.16	208	10816
Total	50	–	–	437.74	2160.5	1965	79385

** $\bar{x} = 39.3$

Using (4.14) the variance is

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{2160.5}{49} = 44.09$$

Employing formula (4.15), we obtain the same value for the variance, as it should have been:

$$s^2 = \frac{50(79385) - (1965)^2}{50(50-1)} = 44.09$$

When n is used as a divisor in place of n–1, we obtain a slightly smaller value for the variance. Check that this value is 43.21:

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n} = \frac{2160.5}{50} = 43.21$$

Summing both sides and dividing throughout by n, we have $\bar{y} = \frac{\bar{x}}{c}$. Using this result, we have

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1} = \frac{\sum(x_i - \bar{x})^2}{c^2(n-1)} = \frac{s_x^2}{c^2}$$

Thus $$s_x^2 = c^2 s_y^2$$

It is easy to verify that when $y_i = cx_i$, we have $s_y{}^2 = c^2 s_x{}^2$, so that

$$s_x^2 = \frac{s_y^2}{c^2}$$

We now state and proof a general theorem combining the effect of changes in origin and the scale of measurement on the variance.

***Theorem* 4.1:** *The variance is independent of origin but dependent on the scale of measurement.*

Proof: Let $x_1, x_2, \ldots x_n$ be a set of n values of a variable x. If these values are transformed to a new set of values $y_1, y_2, \ldots, y_n$ of a variable y, such that

$$y = \frac{x-a}{h} \qquad \ldots \text{(a)}$$

where a and h are two constants and $h>0$, then the theorem asserts that

$$s_x^2 = h^2 s_y^2 \qquad \ldots \text{(b)}$$

To prove this, consider the ith value of the variable y defined in (a) above.

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

giving $$x_i = a + hy_i \qquad \ldots \text{(c)}$$

from which $$\bar{x} = a + h\bar{y} \qquad \ldots \text{(d)}$$

Hence from (c) and (d)

$$x_i - \bar{x} = a + hy_i - (a + h\bar{y}) = h(y_i - \bar{y})$$

Squaring, summing and dividing both sides of the above equation by n–1, we arrive at the following expression:

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

Hence $$s_x^2 = h^2 s_y^2 \text{ (Proved)}$$

As a corollary to the above theorem, we state the following results:

Corollary **4.1:** *If $y=a-bx$, then $s_y=|b|s_x$.*
Corollary **4.2:** *If $y=bx$, then $s_y=|b|s_x$.*
Corollary **4.3:** *If $y=a+x$, then $s_y=s_x$.*

Example 4.9: Verify with Example 4.1 that variance is independent of origin.

Solution: Let us take a=17 and h=1 so that the transformed variable y by virtue of the Theorem (4.1) is of the form $y_i = x_i - a$. We will demonstrate that variance of x is the same as the variance of y. The table below shows the computations.

Child	x_i	y_i *	y^2
1	20	3	9
2	13	–4	16
3	17	0	0
4	17	0	0
5	13	–4	16
6	18	1	1
7	14	–3	9
8	17	0	0
9	16	-1	1
10	15	–2	4
Total	160	–10	56

* $y_i = x_i - 17$

Employing the table values

$$s_y^2 = \frac{n\sum y_i^2 - \left(\sum y_i\right)^2}{n(n-1)} = \frac{10(56) - (-10)^2}{10(10-1)} = 5.11$$

This exactly agrees with our previous value of s_x^2, showing that variance has remained unchanged even if the origin has been shifted to 17.

Example 4.10: Verify Theorem 4.1 with the age distribution in Example 4.2.

Solution: Let us choose a=37 and h=5 (5 being the class size). Then

$$y_i = \frac{x_i - 37}{5}.$$

Here are the detailed computations for the data:

Age	f_i	x_i	y_i	$f_i y_i$	$f_i y_i^2$
24.5–29.5	3	27	–2	–6	12
29.5–34.5	9	32	–1	–9	9
34.5–39.5	15	37	0	0	0
39.5–44.5	12	42	1	12	12
44.5–49.5	7	47	2	14	28
49.5–54.5	4	52	3	12	36
Total	50	–	–	23	97

Hence

$$s_y^2 = \frac{n\sum f_i y_i^2 - \left(\sum f_i y_i\right)^2}{n(n-1)} = \frac{50(97) - (23)^2}{50(49)} = 1.7637$$

Thus $$h^2 s_y^2 = 5^2(1.7637) = 44.09 = s_x^2$$

This numerically demonstrates that variance is independent of origin but dependent on the scale of measurement.

***Theorem* 4.2:** *If $u=x+y$, then $s_u^2 = s_x^2 + s_y^2 + 2\,\mathrm{cov}(x,y)$, where s_u^2 is the variance of u and Cov(x, y) is the covariance between x and y as defined below:*

$$\mathrm{Cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Proof: Since $u = x + y$, $\sum u_i = \sum x_i + \sum y_i$, so that $\bar{u} = \bar{x} + \bar{y}$.

Hence

$$s_u^2 = \frac{\sum (u_i - \bar{u})^2}{n}$$

$$= \frac{\sum [(x_i + y_i) - (\bar{x} + \bar{y})]^2}{n}$$

$$= \frac{\sum [(x_i - \bar{x}) + (y_i - \bar{y})]^2}{n}$$

$$= \frac{\sum (x_i - \bar{x})^2}{n} + \frac{\sum (y_i - \bar{y})^2}{n} + \frac{2\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$= s_x^2 + s_y^2 + 2\mathrm{Cov}(x,y)$$

This proves the theorem.

When $u=x-y$, you can similarly prove that

$$s_u^2 = s_x^2 + s_y^2 - 2Cov(x,y) \quad \ldots (4.16)$$

If x and y are independent, the covariance term is zero and consequently

$$s_u^2 = V(x \pm y) = s_x^2 + s_y^2 . \quad \ldots (4.16a)$$

You may be inquisitive to know why the variance, and hence the standard deviation are computed in terms of deviations from mean, rather than from any other measures of central tendency. The explanation for this lies in the fact that the variance has the smallest value if it is computed from the mean. That is to say, if we use the standard deviation to measure the extent of error in guessing the value for any item in a distribution, then the magnitude of this error will be at its minimum when we guess the arithmetic mean rather than any other value. We proof this statement in the form a theorem below.

Theorem **4.3:** *The variance is minimized if computed from the arithmetic mean*

Proof: Let us assume that a variance s_a^2 is calculated by subtracting an arbitrary value 'a' from each value, squaring the difference and then averaging. That is

$$s_a^2 = \frac{\Sigma (x_i - a)^2}{n} \quad .., (4.17)$$

The quantity in (4.17) is sometimes referred to as the **mean square deviation**. If a is replaced by $\bar{x}$, the above expression turns out to be the variance as defined in before. This indicates that variance is a special case of the mean square deviation. If s_x^2 is used to denote the variance of x, then

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

We now need to show that

$$\frac{\Sigma (x_i - a)^2}{n} \geq \frac{\Sigma (x_i - \bar{x})^2}{n},$$

Or that $$s_a^2 \geq s_x^2$$

The expression in (4.17) above can be partitioned as follows:

between the remaining two extreme values between which the 80 percent of the items fall. If P_{10}^{90} stands for the 10 to 90 percentile range, we have,

$$P_{10}^{90} = P_{90} - P_{10} \quad \ldots (4.2)$$

where P_{90} and P_{10} are the 90th and 10th percentiles of the distribution.

4.3.3 Inter–quartile Range

A measure similar to the above measures is the **inter-quartile range** (I_{QR}). It is the difference between the third quartile (Q_3) and the first quartile (Q_1) Thus

$$I_{QR} = Q_3 - Q_1 \quad \ldots (4.3)$$

This quantity can be interpreted as the length of the interval that contains the middle 50% of the observations. For example, the age distribution in Example 2.3 (Chapter 2) has an inter-quartile range of $Q_3 - Q_1 = 38.83 - 34.67 = 9.2$. This means that we estimate that the middle 50% of all the ages fall within a range that is 9.2 years long.

For a symmetrical distribution, Q_3 and Q_1 are equidistant from the median ($\tilde{m}$). Then $\tilde{m} \pm I_{QR}$ covers exactly 50% of the observations (see Figure 4.2 below).

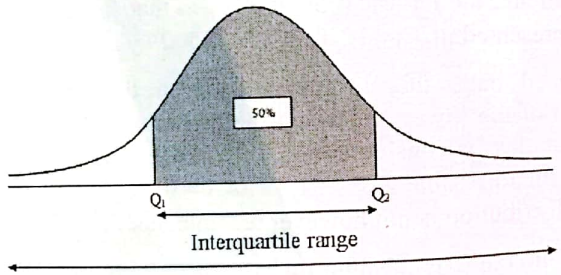


Figure 4.2: Comparison of range and inter-quartile range

The inter-quartile range is frequently reduced to the measure of **semi-inter-quartile range**, also known as the **quartile deviation**, by dividing it by 2. Thus

$$Q_D = \frac{I_{QR}}{2} = \frac{Q_3 - Q_1}{2} \quad \ldots (4.3a)$$

This measure is more meaningful than the range because it is not based on two extreme values.

4.4 Limitations of Range as a Measure of Dispersion

Both the 10 to 90 percentile range and the quartile deviation have serious shortcomings. First of all, they do not take into consideration the values of the items. For example, P_{10}^{90} is not affected by the distribution patterns of those items below P_{10} and above P_{90}. Q_d is not affected by the distribution of all items below Q_1 and above Q_3. Moreover, they remain to be positional measures, failing to provide measurement of scatter of the observations, relative to the typical value. In addition, it does not enter into any of the higher mathematical relationships that are basic to inferential statistics.

4.5 Mean Deviation

For data clustered near the central value, the differences of the individual observations from their typical value will tend to be small. Accordingly, to obtain a measure of the total variation in the data, it is appropriate to find an average of these differences. The resulting average will be called **mean deviation**. It is also known as the **average deviation**.

In practice, the mean deviation is computed as the arithmetic mean of the **absolute values** of the deviations from a **typical value** of a distribution. The typical value may be the arithmetic mean, median, mode or any other arbitrary value. The median is sometimes preferred as a typical value, because the sum of the absolute values of the deviations from the median is smaller than any other value. In practice, however, the arithmetic mean is generally used.

If $x_1, x_2, .., x_n$ form a sample of observations, the formula for computing the average or mean deviation about any arbitrary values 'a' is

$$M_d(a) = \frac{\sum |x_i - a|}{n} \quad \ldots (4.4)$$

where $|\ |$ means that the signs of the deviations whether positive or negative, are ignored.[1] For a grouped frequency distribution with $\sum f_i = n$, the mean deviation about the arbitrary value 'a' is

[1] The absolute value of a number x denoted by $|x|$ is defined as follows: $|x| = x$, if $x \geq 0$ $|x| < c$, if $-c < x < c$, $|x| = -x$, if $x < 0$ $|x| > c$, if $x > c$ or $x < -c$

$$M_d(a) = \frac{\sum f_i |x_i - a|}{n} \quad \ldots (4.5)$$

If we replace 'a' by $\bar{x}$, the resulting mean deviation will be called mean deviation about the mean:

$$M_d(\bar{x}) = \frac{\sum |x_i - \bar{x}|}{n} \quad \ldots (4.5a)$$

where $n = \sum f_i$. For a grouped frequency distribution

$$M_d(\bar{x}) = \frac{\sum f_i |x_i - \bar{x}|}{n} \quad \ldots (4.5b)$$

When the deviations are taken from the median we substitute $\tilde{m}$ for a in (4.5), and the resulting formula for computing mean deviation about the median is

$$M_d(\tilde{m}) = \frac{\sum f_i |x_i - \tilde{m}|}{n} \quad \ldots (4.5c)$$

The following examples demonstrate how the mean deviation is computed.

Example 4.1: Ten persons of varying ages were weighed and the following weights in kg were recorded:

110, 125, 125, 147, 117, 125, 136, 157, 124, 110.

Compute mean deviation about the mean, median and an arbitrary value 120.

Solution: To compute the mean deviation about mean for the given data, the following steps are involved:

a) Compute the arithmetic mean. This is 127.6 in the present instance.
b) Obtain the absolute deviation of each value in column (2) of Table 4.1 from the computed mean. These deviations are shown in column (3).
c) Obtain the sum of column (3) and divide the resulting sum by the total number of observations (n=10).
d) The result obtained in (c) above is the mean deviation about the mean.

Repeat the procedure outlined above to compute the mean deviations about the median (which is 125 for the data set) and the arbitrary value 120, i.e. a=120. The corresponding deviations are shown in last two columns of Table 4.1.

Table 4.1: Computation of mean deviations

Serial no.	Weight (x_i)	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \tilde{m}\rvert$	$\lvert x_i - a\rvert$
(1)	(2)	(3)	(4)	(5)
1	110	17.6	15.0	10.0
2	125	2.6	0	5.0
3	125	2.6	0	5.0
4	147	19.4	22.0	27.0
5	117	10.6	8.0	3.0
6	125	2.6	0	5.0
7	136	8.4	11.0	16.0
8	157	29.4	32.0	37.0
9	124	3.6	1.0	4.0
10	110	17.6	15.0	10.0
Total	1276	114.4	104.0	122.0

e mean deviations about the mean, median and an arbitrary value 120 e respectively

$$M_d(\bar{x}) = \frac{\sum |x_i - \bar{x}|}{n} = \frac{\sum |x_i - 127.6|}{n} = \frac{114.4}{10} = 11.44$$

$$M_d(\tilde{m}) = \frac{\sum |x_i - \tilde{m}|}{n} = \frac{\sum |x_i - 125|}{n} = \frac{104.0}{10} = 10.40$$

d

$$M_d(a) = \frac{\sum |x_i - a|}{n} = \frac{\sum |x_i - 120|}{n} = \frac{122.0}{10} = 12.20$$

ote that among the three mean deviations, mean deviation about the median is e smallest.

xample 4.2: Compute the mean deviations about the mean, median and arbitrary value 42 for the frequency distribution in Table 3.2.

olution: To compute the mean deviation about the mean, follow the steps elow:

a) Calculate the arithmetic mean $\bar{x}$. This is 39.3
b) Take the absolute deviation of each mid-point from $\bar{x}$=39.3 and multiply the deviation by the corresponding frequency to obtain $f_i|x_i - \bar{x}|$.
c) Sum these deviations and divide the resulting sum by the total frequency.

e resulting value in step (c) is the mean deviation about the arithmetic ean.

We reproduce Table 3.2 below and other necessary columns required for the computation.

Table 4.2: Computation of mean deviation about mean

Age	f_i	x_i	$f_i x_i$	$\lvert x_i - \bar{x}\rvert$	$f_i\lvert x_i - \bar{x}\rvert$
24.5–29.5	3	27	81	12.3	36.9
29.5–34.5	9	32	288	7.3	65.7
34.5–39.5	15	37	555	2.3	34.5
39.5–44.5	12	42	504	2.7	32.4
44.5–49.5	7	47	329	7.7	53.9
49.5–54.5	4	52	208	12.7	50.8
Total	50	–	1965	–	274.2

The computed value of the mean $\bar{x}$ and $M_d(\bar{x})$ appear below:

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{1965}{50} = 39.3$$

$$M_d(\bar{x}) = \frac{\sum f_i |x_i - \bar{x}|}{n} = \frac{274.2}{50} = 5.48$$

To calculate mean deviation from median, compute median of the distribution in usual manner and replace the mean by the median. The other steps remain the same. Verify that the median of this distribution is 38.83 and the mean deviation about this value is:

$$M_d(\tilde{m}) = \frac{\sum f_i |x_i - \tilde{m}|}{n} = \frac{\sum f_i |x_i - 38.3|}{50} = \frac{272.32}{50} = 5.45$$

Similarly, we can show that mean deviation about a=42

$$M_d(a) = \frac{\sum f_i |x_i - a|}{n} = \frac{\sum f_i |x_i - 42|}{50} = \frac{285}{50} = 5.7$$

Compare $M_d(\bar{x}), M_d(\tilde{m})$ and $M_d(a)$ and see that mean deviation about median is the smallest of all.

4.3.6 An Alternative Formula for Computing $M_d(\tilde{m})$

The computation of the mean deviation is relatively cumbersome. The calculation, however, can be made simpler using the formula:

$$M_d(\tilde{m}) = \frac{S_1 - S_2}{n} \quad \ldots (4.6)$$

ere
S_1=sum of the observations in the distribution smaller than $\tilde{m}$
S_2=sum of the observations in the distribution larger than $\tilde{m}$

check this, suppose $x_1, x_2, \ldots\ldots, x_n$ is a set of n observations for wh median is $\tilde{m}$. Then by definition

$$M_d(\tilde{m}) = \frac{\sum |x_i - \tilde{m}|}{n}$$

$$= \frac{1}{n}\left[\sum_{x_i > \tilde{m}} |x_i - \tilde{m}| + \sum_{x_i < \tilde{m}} |x_i - \tilde{m}|\right]$$

$$= \frac{1}{n}\left[\sum_{x_i > \tilde{m}} x_i - \sum_{x_i > \tilde{m}} \tilde{m} + \sum_{x_i < \tilde{m}} \tilde{m} - \sum_{x_i < \tilde{m}} x_i\right]$$

ce the median divides the entire distribution into two equal parts,

$$\sum_{x_i > \tilde{m}} \tilde{m} = \sum_{x_i < \tilde{m}} \tilde{m}$$

ence

$$M_d(\tilde{m}) = \frac{1}{n}\left[\sum_{x_i > \tilde{m}} x_i - \sum_{x_i < \tilde{m}} x_i\right] = \frac{1}{n}(S_1 - S_2)$$

r the grouped frequency distribution, the short-cut formula for mputing mean deviation from the median assumes the form

$$M_d(\tilde{m}) = \frac{\sum_{x_i > \tilde{m}} f_i x_i - \sum_{x_i < \tilde{m}} f_i x_i + \tilde{m}\left(\sum_{x_i < \tilde{m}} f_i - \sum_{x_i > \tilde{m}} f_i\right)}{n} \quad \ldots (4.7)$$

ample 4.3: Calculate the mean deviation about median for the servations 4, 6, 7, 10, 12 using (4.6).

lution: The median of the given observations is 7 and the mean viation from the median is 2.4:

$$M_d(\tilde{m}) = \frac{|4-7| + |6-7| + \ldots + |12-7|}{5} = 2.4.$$

e sum of the observations above the median value 7 is 22:

$$S_2=10+12=22$$

And the sum of the observations below the median value is 10:

$$S_1=4+6=10.$$

Hence using (4.6)

$$M_d(\tilde{m})=\frac{S_2-S_1}{n}=\frac{22-10}{5}$$

This agrees with the value obtained by usual method.

Example 4.4: Use the data in Table 4.2 to compute the mean deviation about the median value of the distribution using (4.7).

Solution: The table referred to above is re-produced below with x_i (the class mid-point), f_i and f_ix_i values as follows to facilitate the computations:

Mid-points (x_i)	Frequency (f_i)	Product (f_ix_i)
27	3	81
32	9	288
37	15	555
42	12	504
47	7	329
52	4	208
Total	50	1965

The median of this distribution is 38.83, which lies between 37 and 42. Hence

$$\sum_{x_i>\tilde{m}} f_ix_i=504+329+208=1041,\quad \sum_{x_i<\tilde{m}} f_ix_i=81+288+555=924.$$

$$\sum_{x_i>\tilde{m}} f_i=12+7+4=23,\quad \sum_{x_i<\tilde{m}} f_i=3+9+15=27$$

Substituting these values in (4.7)

$$M_d(\tilde{m})=\frac{(1041-924)+38.83(27-23)}{50}=5.45$$

which exactly agrees with the one we obtained earlier in Example 4.2 using usual method.

4.3.7 Variance and Standard Deviation

Instead of ignoring the signs of deviations from the mean as in the computation of an average deviation, they may each be squared and then

results are added[1]. The sum of squares can be regarded as a measure of total dispersion of the distribution. By dividing the sum by n (the total number of observations), we obtain the average of the squares of deviations, a measure, called **variance**, of the distribution. If the observations are all from a population, the resulting variance is referred to as the **population variance**. As a formula, the variance of population observations $x_1, x_2, ..., x_N$, commonly designated σ^2 is

$$\sigma^2=\frac{\sum(x_i-\mu)^2}{N} \quad \ldots.(4.8)$$

where μ is the mean of all the observations in the population and N is the total number of observations in the population. Because of the operation of squaring, the variance is expressed in square units (e.g.km^2, $taka^2$, etc.), and not (e.g. *km*, *taka*, etc.), of the original unit. It is therefore necessary to extract the positive square root to restore the original unit. The measure of dispersion thus obtained is called the **population standard** deviation and is usually denoted by σ. Thus

$$\sigma=\sqrt{\frac{\sum(x_i-\mu)^2}{N}} \quad \ldots (4.9)$$

Thus, by definition, the standard deviation is the positive square root of the mean-square deviations of the observations from their arithmetic mean.

In many statistical applications, we deal with a **sample** rather than a population. Thus, while a set of population observations yields a population variance, a set of sample observations will yield a **sample variance**. Thus, if x_1, x_2, ...,x_n is a set of sample observations of size n, then the sample variance, denoted by s^2, is expressed as

$$s^2=\frac{\sum(x_i-\bar{x})^2}{n} \quad \ldots (4.10)$$

where $\bar{x}$ is the mean of all the sample observations.

The standard deviation of a sample mean is often called the standard error of a mean or simply **standard error**. In other words, the term 'standard

[1] By the squaring operation, the deviations from the mean will sum not to 0, but to a positive number, and each deviation will contribute to the sum of squares regardless of sign.

error' applies to means, unless otherwise specified. The standard error of the mean, denoted by $s_{\bar{x}}$, is computed as

$$s_{\bar{x}} = \sqrt{\frac{\text{Sample variance}}{\text{Sample size}}} = \frac{s_x}{\sqrt{n}} \quad \ldots (4.11)$$

Thus, a standard error can be calculated if an s^2 or s is available, more than one $\bar{x}$ is not required.

The concept of standard error is best understood with reference to a sampling distribution, an analogous counterpart of a frequency distribution. Just as the standard deviation applies to a frequency distribution, a standard error is applied to a sampling distribution. This implies that standard error is the standard deviation of a sampling distribution.

When we compute a measure of variability for the sample, we often are interested in using the sample variance s^2 as an estimate of the population variance σ^2. At this point, it might seem that the average of the squared deviations in the sample would provide a good estimate of the population variance. However, statisticians have found that the average squared deviation for the sample has the undesirable feature that it tends to underestimate the population variance σ^2. Because of this tendency toward underestimation, we say that it provides a biased estimate. This means that such an estimate shows a systematic tendency to be less than σ^2, the population variance.

Fortunately, it can be shown that if the sum of the squared deviations in the sample is divided by $n-1$, and not by n, then the resulting sample variance will provide an unbiased estimate of the population variance[a]. For this reason, the **sample variance** is defined as follows:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \quad \ldots (4.12)$$

Such an estimate will show no systematic tendency to be either greater than or less than the population variance σ^2. The division by $n-1$ instead of n makes the average squared deviation consistent with many similar measures used in statistical measures.

[a] By unbiased estimate, we mean that average of all possible sample variances will be equal to the population variance. Symbolically, $E(s^2)=\sigma^2$.

4.8 Computing Variance for Ungrouped Data

The variance and hence the standard deviation are simple to compute for ungrouped data. Suppose a data set consists of n values $x_1, x_2, \ldots, x_n$. As a first step, compute the arithmetic mean $\bar{x}$ for this data set. Then subtract this mean from each of the values of x and obtain a set of deviations $(x_1-\bar{x}), (x_2-\bar{x}), \ldots, (x_n-\bar{x})$. Then square these deviations, sum them and divide the resulting sum by n. This gives you the variance of the given values $x_1, x_2, \ldots x_n$. Let us illustrate the computation of variance from raw data by an example.

Example 4.5: Compute the variance and standard deviation from the data on weight of ten children in Example 4.1.

Solution: The data were as follows: 20, 13, 17, 17, 13, 18, 14, 17, 16, and 15. The mean of this set is 16. Following the steps outlined above, the accompanying table is constructed to illustrate the computation of variance.

Table 4.3: Computation of variance and standard deviation

Child	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x^2
1	20	4	16	400
2	13	–3	9	169
3	17	1	1	289
4	17	1	1	289
5	13	–3	9	169
6	18	2	4	324
7	14	–2	4	196
8	17	1	1	289
9	16	0	0	256
10	15	–1	1	225
Total	160	0	46	2606

The variance is thus

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{46}{9} = 5.11\,\text{kg}^2$$

Taking square root of the variance, we obtain the standard deviation:

$$s = \sqrt{5.11\,\text{kg}^2} = 2.26\,\text{kg}$$

The process outlined above, however, is rather laborious, because the arithmetic mean needs to be subtracted from each and every observation. It

is specially time consuming if the mean is any number with several digits or decimal places. This problem may be avoided by using an alternative form of the formula as derived below.

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

Comparing this with (4.12)

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

so that

$$s^2=\frac{n\sum x_i^2-\left(\sum x_i\right)^2}{n(n-1)} \qquad \ldots (4.13)$$

The formula (4.13) makes it unnecessary to subtract the mean from each observation. Table 4.3 can now be used to compute the variance:

$$s^2=\frac{10\times 2606-160^2}{10\times 9}=5.11, \text{ as before.}$$

The quantity $\sum(x_i-\bar{x})^2$ is often known as the **corrected sum of squares** or simply **sum of squares of** x while the quantities $\sum x_i^2$ is referred to as the **raw sum of squares.**

4.3.9 Computing Variance for Frequency Distribution

The formula for computation of variance and standard deviation of a frequency distribution should be modified to take into account the values of x and their corresponding frequencies. Thus if the variable values $x_1, x_2, \ldots, x_k$ each occur with frequencies $f_1, f_2, \ldots, f_k$ respectively, then

$$s^2=\frac{\sum f_i(x_i-\bar{x})^2}{n-1} \qquad \ldots (4.14)$$

For grouped data x_i will be the mid-value of the i-th class.

The formula for the computation of the variance presented above can be rewritten in a compact form as follows:

$$s^2=\frac{1}{n-1}\left[\sum f_i x_i^2-\frac{\left(\sum f_i x_i\right)^2}{n}\right]$$

$$=\frac{n\sum f_i x_i^2-\left(\sum f_i x_i\right)^2}{n(n-1)} \qquad \ldots (4.15)$$

many textbooks, the divisor n is used in place of $n-1$. The discrepancy the value of s^2 resulting from the use of n instead of n-1 is not however bstantial when n is large.

xample 4.6: Compute the variance and standard deviation for the llowing frequency distribution:

x:	3	5	7	8	9
f:	2	3	2	2	1

olution: The following table illustrates the computation of variance from e above distribution.

x_i	f_i	f_ix_i	$f_ix_i^2$
3	2	6	18
5	3	15	75
7	2	14	98
8	2	16	128
9	1	9	81
Total	10	60	400

$$s^2=\frac{n\sum f_ix_i^2-\left(\sum f_ix_i\right)^2}{n(n-1)}=\frac{10(400)-(60)^2}{10(10-1)}=4.44$$

xample 4.7: The lengths of 32 leaves were measured correct to the earest mm. Find the mean, variance and hence the standard deviation of e lengths.

Length:	20-22	23-25	26-28	29-31	32-34
Frequency:	3	6	12	9	2

olution: In order to compute the required measures, we construct the llowing table:

Length	x_i	x_i^2	f_i	f_ix_i	$f_ix_i^2$
20–22	21	441	3	63	1323
23–25	24	576	6	144	3456
26–28	27	729	12	324	8748
29–31	30	900	9	270	8100
32–34	33	1089	2	66	2178
Total	-	-	32	867	23805

The mean length is

$$s_a^2 = \frac{\sum[(x_i - \bar{x}) + (\bar{x} - a)]^2}{n}$$

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

$$= s_x^2 + (\bar{x} - a)^2 + 0$$

Hence $$s_a^2 \geq s_x^2$$

This proves that the variance and hence the standard deviation about the arithmetic mean are always smaller than when the variance (or standard deviation) is computed about any value other than the arithmetic mean.

Example 4.11: Given the values 2, 5, 8. Verify Theorem 4.3.

Solution: Here $\bar{x}=5$, so that

$$s_x^2 = \frac{(2-5)^2 + (5-5)^2 + (8-5)^2}{3} = 6.$$

For an arbitrary value $a=8$,

$$s_a^2 = \frac{(2-8)^2 + (5-8)^2 + (8-8)^2}{3} = 15.$$

This demonstrates that $s_a^2 \geq s_x^2$.

Example 4.12: From a certain frequency distribution consisting of 18 observations, the mean and the standard deviation were computed to be 7 and 4 respectively. But on comparing the original data, it was found that an observation 12 was misreported as 21 in the computation. Compute the correct mean and correct standard deviation.

Solution: With the incorrect mean $\bar{x} = 7$, we have $\sum x_i = n\bar{x}$ =18×7=126 while the correct sum will be $\sum x_i' = 126-21+12=117$. Hence the correct mean $(\bar{x}_c)$ will be

$$\bar{x}_c = \frac{\sum x_i'}{n} = \frac{117}{18} = 6.5$$

Based on the incorrect standard deviation 4

$$4^2 = \frac{\sum x_i^2}{18} - 7^2$$

which yields $$\sum x_i^2 = 1170$$

But this is not the correct sum of squares. The correct sum of squares will be

$$\sum x_i'^2 = 1170 - (21)^2 + (12)^2 = 873$$

Hence the correct standard deviation will be

$$s_c^2 = \frac{\sum x_i'^2}{n} - \bar{x}_c^2 = \frac{873}{18} - 6.5^2 = 6.25$$

Hence the correct standard deviation is

$$s_c = \sqrt{6.25} = 2.5$$

4.3.11 Uses of Standard Deviation

A thorough understanding of the use of standard deviation is difficult for us at this stage, unless we acquire some knowledge on some theoretical distributions in statistics. Nevertheless, we shall try to introduce the idea of its use through a few simple illustrative examples. The standard deviation of a population (σ) is a measure of the dispersion in the population, while the standard deviation of sample observations (s) is a measure of the dispersion in the distribution constructed from the sample. In both the cases, the standard deviation (like the mean deviation) represents the average variability in a distribution. The greater this variability around the mean of a distribution, the larger the standard deviation. Thus s=4.5, for example, indicates greater variability than s=2.5.

The use of standard deviation can be best understood with reference to a normal distribution[1]. The normal distribution is completely defined by its mean (μ) and standard deviation (σ). An important characteristic feature of a normal distribution (more precisely to say, of a normally distributed variable) is that

[1] *A normal distribution is a bell-shaped distribution, symmetric in shape, with equal mean, median and mode. As we will see, a normal distribution has wide applications in statistics (for more details, see Chapter X).*

- 68.27 percent of all observations are expected to lie within one standard deviation of the mean, i.e. in the interval $\mu \pm \sigma$.
- 95.45 percent of all observations are expected to lie within two standard deviations of the mean i.e. in the interval $\mu \pm 2\sigma$.
- 99.73 percent of all observations are expected to lie within three standard deviations of the mean i.e. in the interval $\mu \pm 3\sigma$.

Not only that the above feature is true for a normal distribution, for most distributions that we deal with, have this appealing feature (see Figure 4.3).

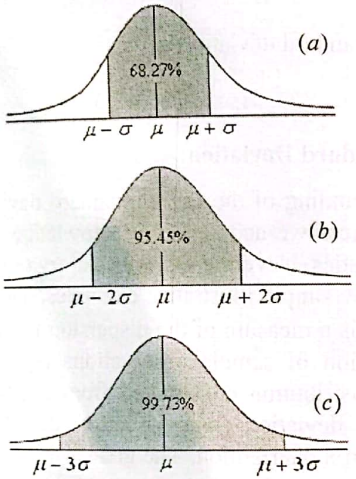


Figure 4.3(a-c): Percent of total area under the curve between various parts

Example 4.13: Suppose a group of 1000 women have a mean height of 158 cm with a standard deviation of 3 cm. We assume that the heights of these women have an approximately normal distribution. Using the above empirical rules, we can make the following assertions:

- 683 women have height between 158±1(3) cm, i.e. between 155 cm and 161 cm.
- 955 women have height between 158±2(3) cm, i.e. between 152 cm and 164 cm.
- 997 women have height between 158±3(3) cm, i.e. between 149 cm and 167 cm.

It is important to note that the rule is applicable regardless of the value of the mean and standard deviation so long as the distribution is normal.

Example 4.14: The accompanying data refer to the payment delay times of 75 telephone users as displayed in Example 2.17 in Chapter 2.

22	29	16	15	18	17	12	13	17	16	15	19	17
10	21	15	14	17	18	12	20	14	16	15	16	20
22	14	25	19	23	15	19	18	23	22	16	16	19
13	18	24	24	26	13	18	17	15	24	15	17	14
18	17	21	16	21	25	19	20	27	16	17	16	21

(a) Display the data by a stem and leaf plot and comment on its skewness of the distribution indicated by the data.

(b) Compute mean and standard deviation of the data set and hence show how well the empirical rule based on the mean and standard deviation aggress with nature of the distribution as indicated by the stem and leaf plot.

Stem	Leaf
10	0
11	
12	00
13	000
14	0000
15	0000000
16	000000000
17	00000000
18	000000
19	00000
20	000
21	0000
22	000
23	00
24	000
25	00
26	0
27	0
28	
29	0

Solution: As displayed by the stem and leaf plot, the distribution is moderately skewed to the right. What does our empirical rule say? For this, we calculate mean and standard deviation of the data set. These are 18.1077 and 3.9612. With these values, we construct the following intervals:

(a) $\bar{x} \pm s = (14.1,\ 22.1)$, (b) $\bar{x} \pm 2s = (10.2,\ 26.0)$, (c) $\bar{x} \pm 3s = (6.2,\ 30.0)$

Through actual counting we find that the interval shown in (a) i.e. 14.1–22.1 contains 45 (69.2%) observations, while the empirical rule suggests 68.3% to be included in this interval. The interval shown in (b) i.e. 10.2–26.0 contains 62 (95.4%) observations, while the empirical rule suggests 95.5% to be included in this interval. The interval shown in (c) i.e. 6.2–30.0 contains 100% observations, while the empirical rule suggests 99.7% to be included in this interval. Our conclusion is that the empirical rules work reasonably well in this particular instance, since the distribution is moderately skewed.

Another important use of the standard deviation is in measuring the difference of a given observation from the mean. If an observation has a value x, the difference of this value from the population mean μ, i.e. $x-\mu$ can be expressed as a given number of standard deviation (σ) denoted by z. The z value, which is known as the **standardized normal variate**, is given by

$$z=\frac{x-\mu}{\sigma} \qquad \ldots (4.18)$$

A z-value indicates the relative location of a value within a population or a sample. A positive z-value says that x is above (greater than) the mean, while a negative z-value says that x is below (less than) the mean. For instance, a z-value equal to 2.5 says that x is 2.5 standard deviation above the mean. Values in 2 different populations or samples having the same z-value are the same number of standard deviations from their respective means and therefore, have the same relative locations.

Very often in statistical studies, we are interested in specifying the percentage or proportion of items in a data set that lie within some specified interval when only the mean and standard deviation of the data set are known. The Russian mathematician **Tchebysheff** discovered that the fraction or proportion of the data set lying between any two values symmetric about the mean is related to the standard deviation. This rule applies to all distributions, skewed or otherwise. The rule is expressed by the following inequality:

$$P\left[|x_i-\bar{x}|\le ks\right]\ge 1-\frac{1}{k^2} \qquad \ldots (4.19)$$

For $k=2$, the theorem states that at least $1-1/2^2=75\%$ of the observations must lie within two standard deviations from the mean. That is, 75% or more of the observations of any distribution lie in the interval $\bar{x}\pm 2s$.

The use of standard deviation is manifold. It employs the mathematically acceptable procedure of clearing the signs, and because of this reason, it has wider application than the mean deviation. As a result, the standard deviation has become the initial step for obtaining certain other statistical measures, especially in the context of statistical decision making.

4.4 RELATIVE MEASURES OF DISPERSION

4.4.1 Coefficient of Variation

The **coefficient of variation** (CV) is one of the important measures of dispersion that attempts to measure the variability in data relative to the mean. When mean values of two or more data sets vary considerably, we do not get an accurate picture of the relative variability in the sets just by comparing the standard deviations. Coefficient of variation tends to overcome this difficulty. This is a measure that represents the spread of the distribution relative to the mean of the same distribution.

A coefficient of variation is computed as a ratio of the standard deviation of the distribution to the mean of the same distribution. Expressing in percentage form, the symbolic representation of the coefficient is:

$$CV=\frac{s_x}{\bar{x}}\times 100 \qquad \ldots (4.20)$$

Clearly, if the mean of a data set is zero, CV cannot be computed. The measure is a pure number and independent of units.

A value of 33 percent, for example, for CV implies that the standard deviation of the sample value is 33 percent of the mean of the same distribution. As an illustration of the use of CV as descriptive statistics, let us look at the following examples:

Example 4.15: Suppose that we wish to obtain some insight into whether height is more variable than the weight in the same population. For this purpose, we have the following data obtained from 150 children in a community.

	Height	Weight
Mean	40 inch	10 kg
SD	5 inch	2 kg
CV	12.5%	20.0%

Examination of the respective standard deviations does not tell us in any meaningful way which characteristic has more variability than the other, because they are measured in different units. If we now compute coefficient of variation, the results become comparable, because coefficient of variation is a unit-free quantity. Thus, since the coefficient of variation for weight is greater than that of the height, we conclude that weight has more variability than height in the population.

Even if two variables in the same population are measured in the same unit, the standard deviation may fail to provide a correct picture of their relative variability. This is illustrated by an example below.

Example 4.16: Consider that the blood pressures of a group of patients were measured at two levels: systolic and diastolic, both being measured in the same unit. The results were as follows:

	Systolic	Diastolic
Mean	130 mm Hg	60 mm Hg
SD	15 mm Hg	8 mm Hg
CV	11.5%	13.3%

As implied by the standard deviations, systolic pressure is more variable (sd=15 mm Hg) than the diastolic pressure (sd=8 mm Hg). However in relative terms, as measured by the CV, the diastolic pressure has the greater variability. This shows that the relative variability is of more concern than absolute variation – hence the importance of the coefficient of variation.

The discussions and examples above tend to demonstrate that coefficient of variation is a very useful measure when:

1. The data are in different units
2. The data are in the same units but the means are far apart
3. When the data sets involve all or nearly all positive values.

Example 4.17: The average weekly wage in a factory had increased from Tk.8000 to Tk.12000 as result of negotiation between the employees and the employer. Alongside, the standard deviation had decreased from Tk.150 to Tk.100. Can we conclude that after negotiation, the wage has become higher and more uniform?

Solution: As the standard deviation after the settlement shows a lower value than before, one might tend to conclude that disparity in wage has been considerably reduced. But the average wage differs considerably before and after the settlement. It is therefore not safe to base our decision only on the basis of standard deviation. Coefficient of variation seems to be the best tool in this instance. Thus

$$\text{CV (before settlement)} = \frac{100}{8000} \times 100 = 125\%$$

$$\text{CV (after settlement)} = \frac{150}{12000} \times 100 = 125\%$$

The variability and hence the disparity in the distribution of wages remained as before as shown by the CV, although the average wage has shown an increase from 8000 to 12000.

***Theorem* 4.4:** *Coefficient of variation is independent of scale but not of the origin.*

Proof: Let x be our original variable taking on values $x_1, x_2, \ldots, x_n$. We change this variable to y taking on values $y_1, y_2, \ldots, y_n$ such that $y=x-a$. The implication of this change is that $\bar{y} = \bar{x} - a$. Since the variance is independent of origin, $s_y = s_x$ so that

$$\text{CV}(y) = \frac{s_y}{\bar{y}} = \frac{s_x}{\bar{x} - a} \neq \text{CV}(x)$$

Thus CV is not independent of the origin.

Let us make a change in the value of x by dividing each value by a scale factor h such that $y=x/h$. The mean and standard deviation of y are respectively $\bar{y} = \bar{x}/h$ and $s_y = s_x/h$, so that

$$\text{CV}(y) = \frac{s_y}{\bar{y}} = \frac{s_x}{h} / \frac{\bar{x}}{h} = \frac{s_x}{\bar{x}} = \text{CV}(x)$$

which does not involve 'h', the scale factor. This proves that CV is independent of scale.

4.4.2 Coefficient of Range

The **coefficient of range** is a relative measure corresponding to range and is obtained by the following formula:

$$C_R = \frac{L - S}{L + S} \times 100 \qquad \ldots (4.21)$$

where L and S are respectively the largest and the smallest observations in the data set. The coefficient of range is rarely used as a measure of dispersion because of its inherent difficulties in interpretation.

4.4.3 Coefficient of Mean Deviation

The third relative measure is the **coefficient of mean deviation** (C_{MD}). As the mean deviation can be computed from mean, median,, mode or from any arbitrary value, a general formula for computing coefficient of mean deviation may be put as follows:

$$C_{MD}(a)=\frac{M_d(a)}{a}\times 100 \qquad \ldots (4.22)$$

where 'a' may be the mean, median, mode or any other arbitrary value.

4.4.4 Coefficient of Quartile Deviation

The coefficient of quartile deviation (C_{QD}) is computed from the first and the third quartiles using the following formula:

$$C_{QD}=\frac{Q_3-Q_1}{Q_3+Q_1}\times 100 \qquad \ldots (4.23)$$

It is worth to mention that most of the absolute measures, except CV, are of little significance because of their limited practical utility.

4.5 EMPIRICAL RELATIONS AMONG MEASURES OF DISPERSION

For a symmetrical and moderately skewed distribution, certain measures of dispersion demonstrate close relationships among themselves. Here are some of the relationships:

$$\text{Mean deviation}=\frac{4}{5}(\text{standard deviation}) \qquad \ldots (4.24)$$

$$\text{Quartile deviation}=\frac{2}{3}(\text{standard deviation}) \qquad \ldots (4.25)$$

The implication of the above two relations is that

$$\text{Mean deviation}=\frac{6}{5}(\text{Quartile deviation}) \qquad \ldots (4.26)$$

The relations are useful in estimating one measure of dispersion when the other is known, or in verifying approximately the consistency of the measures obtained by direct calculation. If the estimated standard deviation

differs markedly from its value estimated from quartile deviation and mean deviation, one would tend to conclude that either an error has been made or the distribution, from which these estimates have been obtained, differs considerably from being symmetrical.

Example 4.18: The age distribution presented in Table 4.2 yielded the following measures: SD=6.64, Q_1=34.67, Q_2=38.83, Q_3=43.87, mean deviation=5.84. Estimate the mean deviation and quartile deviation of this distribution using the empirical relationships (4.24) and (4.25) and compare the results with the one computed directly from the distribution.

Solution: From (4.24), the mean deviation about the arithmetic mean is

$$M_d(\bar{x})=\frac{4}{5}\times SD=\frac{4}{5}\times 6.64=5.31$$

which agrees well with the actual value 5.84. This implies that the underlying distribution is moderately skewed. Again from (4.25), we have

$$Q_d=\frac{2}{3}\times SD=\frac{2}{3}\times 6.64=4.43$$

while the actual value of the Q_d is $(Q_3-Q_1)/2$=4.60. This value confirms that the underlying distribution is approximately symmetrical.

Another comparison may be made of the proportion of the items that are typically included within the range of one Q_d, or M_d, measured both above and below the mean. In symmetrical and moderately skewed distributions, the following empirical rules hold good:

$\bar{x}\pm Q_d$ includes middle 50% of the observations

$\bar{x}\pm M_d$ includes middle 57.5% of the observations

How do we interpret the above relationships? Taking again the Example 4.16 above as an illustration, the first rule viz. $\bar{x}\pm Q_d$ states that the distribution referred to above with mean = 30.89 and Q_d= 1.30, the range (30.89±1.30) i.e. 29.59 to 32.19 will contain middle 50 percent of the observations, given the distribution is symmetrical or nearly so.

Example 4.19: The mean, standard deviation and coefficient of variation of 10 observations are 15, 4.38 and 2.7% respectively. How would the results be affected if it is decided to increase each observation by a constant amount 5?

Solution: Let the original values be $x_1, x_2, \ldots, x_{10}$ and increased values be $y_1, y_2, \ldots, y_{10}$ so that $y_i = x_i + 5$. The new mean is thus $\bar{y} = \bar{x} + 5 = 15 + 5 = 20$. This confirms that the mean is affected by the changes in origin. The new variance will be

$$s_y^2 = \frac{\sum_{i=1}^{10}(y_i - \bar{y})^2}{10} = \frac{\sum_{i=1}^{10}(x_i + 5 - \bar{x} - 5)^2}{10} = \frac{\sum_{i=1}^{10}(x_i - \bar{x})^2}{10} = s_x^2$$

confirming once again that variance (and hence standard deviation) is independent of origin.

The new coefficient of variation will now be changed to:

$$\text{CV}(y) = \frac{s_y}{\bar{y}} \times 100 = \frac{4.38}{20} \times 100 = 21.9\%$$

4.6 COMPARING THE MEASURES OF DISPERSION

Like the measures of averages, a measure of dispersion should also satisfy certain criteria in order to be reckoned as an ideal measure. From this point of view, a measure of dispersion should be

- Unambiguously defined
- Easy to understand
- Based on all the observations
- Affected less due to sampling fluctuations
- Less affected by extreme values and
- Amenable to algebraic treatment.

To what extent are these conditions satisfied by the measures we have discussed so far? We provide here a brief overview of the advantages, in the light of the above criteria.

Range: The range has a clear-cut definition. It is easy to understand and is a common way to describe dispersion. It is especially useful in situations where the purpose of investigation is only to find out the extent of extreme variations. For instance, weather forecast is usually reported in terms of the lowest and the highest temperatures rather than all the hourly readings of the day. Sales in a book exhibition or transaction in a share market are usually reported in this fashion.

The range, however, has certain drawbacks that tend to limit its usefulness as a measure of variability. Since it depends solely on the largest and the

smallest values, it is highly sensitive to the presence of unusual and extreme values in a series. Furthermore, the range does not provide measurement of the dispersion of items relative to the central value. It tends to increase as the size of the sample increases. Moreover, the range cannot be used meaningfully with nominal or ordinal data. Because it is based on only two terminal observations, it is not suitable for algebraic treatment.

Mean deviation: The mean deviation possesses many of the desirable properties of an ideal measure. It takes into account every item in the distribution and shows the scatter of the items around the measure of central tendency. It has been found that if the distribution is 'normal' or nearly so, approximately 57.5 percent of the observations are included in the range $\bar{x} \pm M_d$. The chief advantage of mean deviation is that its knowledge helps us to understand the standard deviation, which is one of the most important measures of dispersion.

One of the drawbacks of the mean deviation is the ambiguity about the measure of central tendency to be used for its computation. In order to avoid confusion, it is necessary to state clearly weather the mean or the median or any other value is used in computing the average deviation.

The most serious defect of mean deviation, however, is the fact that the signs of the deviations must be ignored. The procedure of ignoring the signs makes the method non-algebraic and the measure is not amenable to mathematical manipulation. This handicap is serious in working out the theory of sampling distribution and statistical inference.

Standard deviation: Because of its high degree of accuracy and precision, standard deviation is the most prominently used measure of dispersion. It is based on all the observations, highly amenable to further algebraic treatment and is considerably less affected due to sampling fluctuations.

Quartile deviation: The quartile deviation has a special utility in measuring variation in the case of open-end distribution. It has an advantage that it is less affected by extreme values in the data set. It is also less affected by sampling variability.

The chief disadvantage is that it ignores 50% of its observations in the computation, 25% from the upper tail, and 25% from the lower tail. Further, no algebraic manipulation is possible with the quartile deviation

Coefficient of variation: The coefficient of variation is a dimensionless measure and because of this, it is regarded as the most commonly used measure of relative variation.

Example 4.20 Find the mean deviation from the mean and variance of the series a, $a+d$, $a+2d$, ...,$a+2nd$.

Solution: The series consists of $2n+1$ terms. Hence the mean of the series is

$$\bar{x} = \frac{a+(a+d)+(a+2d)+\ldots+(a+2nd)}{2n+1}$$
$$= \frac{(2n+1)a + d(1+2+\ldots+2n)}{2n+1}$$
$$= \frac{(2n+1)a + nd(2n+1)}{2n+1} = a+nd$$

The mean deviation from the mean is $M_d(\bar{x}) = \frac{1}{2n+1}\sum|x_i - \bar{x}|$

Substituting a, $a+d$, $a+2d$,.....,$a+2nd$ successively for x_i and $a+nd$ for $\bar{x}$, we have

$$M_d(\bar{x}) = \frac{1}{2n+1}\sum|x_i - \bar{x}|$$
$$= \frac{1}{2n+1}[nd+(n-1)d+.....2d+d+0+d+2d+\ldots+(n-1)d+nd]$$
$$= \frac{2d}{2n+1}(1+2+\ldots+n) = \frac{nd(n+1)}{2n+1}$$

The variance of the series is

$$s_x^2 = \frac{1}{2n+1}\sum(x_i-\bar{x})^2$$
$$= \frac{1}{2n+1}[n^2d^2+(n-1)^2d^2+....+d^2+0+d^2+....+(n-1)^2d^2+n^2d^2]$$
$$= \frac{2d^2}{2n+1}(1^2+2^2+.....+n^2)$$
$$= \frac{2d^2}{2n+1}\left[\frac{n(n+1)(2n+1)}{6}\right] = \frac{nd^2(n+1)}{3}$$

Example 4.21: (a) Find two numbers whose arithmetic mean is 6 and variance is 16 (b) Find the variance of first n natural numbers whose frequencies are equal to the corresponding numbers. .

Solution: (a) Let the numbers be a and b. Then the mean $\bar{x}$ and standard deviation s_x of the two numbers are respectively

$$\bar{x} = \frac{a+b}{2} \text{ and } s_x = \frac{|a-b|}{2}.$$

Thus

$a+b=12$ and $s_x^2 = 16 = \frac{(a-b)}{4}$. The second equation leads to $a-b=\pm 8$

Solving the equations, $a=2$, 10, $b=10$, 2. Hence the numbers are 2 and 10.

(b) The frequency distribution and the required columns needed for the computation of the variance are shown in the accompanying table:

x_i	f_i	f_ix_i	$f_ix_i^2$
1	1	1^2	1^3
2	2	2^2	2^3
...	...	...	...
...	...	...	...
n	n	n^2	n^3

$$\bar{x} = \frac{\sum f_ix_i}{n} = \frac{1^2+2^2+\ldots+n^2}{1+2+\ldots+n} = \frac{n(n+1)(2n+1)/6}{n(n+1)/2} = \frac{2n+1}{3}$$

$$s_x^2 = \frac{\sum f_ix_i^2}{n} - \bar{x}^2 = \frac{1^3+2^3+\ldots+n^3}{1+2+\ldots+n} - \left(\frac{2n+1}{3}\right)^2$$
$$= \frac{[n(n+1)/2]^2}{n(n+1)/2} - \left(\frac{2n+1}{3}\right)^2 = \frac{(n-1)(n+2)}{18}$$

4.7 A FEW MORE THEOREMS ON DISPERSION

Theorem 4.5: *For any set of values x_1, x_2..., x_n the mean deviation about the arithmetic mean cannot exceed the standard deviation.*

Proof: If s_x stands for the standard deviation of the above set and $M_d(\bar{x})$ for the mean deviation about the mean, we require to prove that $s_x \geq M_d(\bar{x})$

Let us define the above quantities as follows:

$$s_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 \quad \text{and} \quad M_d(\bar{x}) = \frac{1}{n}\sum |x_i - \bar{x}|$$

Since $\sum (x_i - \bar{x})^2 \geq 0$, we have

$$\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \geq 0, \text{ or } \sum x_i^2 \geq \frac{\left(\sum x_i\right)^2}{n}$$

The above inequality is true for any real value of x. Replacing x_i by $|x_i - \bar{x}|$ in the above inequality,

$$\sum (x_i - \bar{x})^2 \geq \frac{\left(\sum |x_i - \bar{x}|\right)^2}{n} \qquad [\because |x_i - \bar{x}|^2 = (x_i - \bar{x})^2]$$

Dividing throughout by n

$$\frac{\sum (x_i - \bar{x})^2}{n} \geq \frac{\left(\sum |x_i - \bar{x}|\right)^2}{n^2}$$

Hence

$$s_x \geq M_d(\bar{x})$$

***Theorem* 4.6**: *If $\bar{x}_1$ and s_1^2 are respectively the mean and variance of n_1 observations, $\bar{x}_2$ and s_2^2 are respectively the mean and variance of n_2 observations, then the combined variance s^2 of all the n_1+n_2 observations is given by*

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2} + \frac{n_1 n_2}{(n_1 + n_2)^2}(\bar{x}_1 - \bar{x}_2)^2$$

Proof: The observations of the two sets and their means and variances are shown below:

Set	Observations	Mean	Variance
I	$x_{11}, x_{12}, \ldots, x_{1n_1}$	$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}$	$s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2$
II	$x_{21}, x_{22}, \ldots, x_{2n_2}$	$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}$	$s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2$

If $\bar{x}$ stands for mean of the combined set, then, by definition, the combined variance s^2 of the two sets of n_1+n_2 observations is given by

$$s^2 = \frac{1}{n_1 + n_2}[(x_{11} - \bar{x})^2 + (x_{12} - \bar{x})^2 + \ldots\ldots\ldots + (x_{1n_1} - \bar{x})^2 + (x_{21} - \bar{x})^2 + (x_{22} - \bar{x})^2 + \ldots\ldots\ldots + (x_{2n_2} - \bar{x})^2]$$

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

where $\bar{x}$ is defined as

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

The expression for the combined variance above can be rewritten as

$$(n_1 + n_2)s^2 = \sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x})^2 \qquad \ldots (4.27)$$

But

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x})^2 = \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1 + \bar{x}_1 - \bar{x})^2$$

$$= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + n_1(\bar{x}_1 - \bar{x})^2$$

$$= n_1 s_1^2 + n_1(\bar{x}_1 - \bar{x})^2$$

Similarly

$$\sum_{i=1}^{n_2}(x_{2i} - \bar{x})^2 = n_2 s_2^2 + n_2(\bar{x}_2 - \bar{x})^2$$

Again

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

Similarly

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

Thus (4.27) can be written as

$$(n_1+n_2)s^2 = n_1s_1^2 + n_2s_2^2 + \frac{n_1n_2^2}{(n_1+n_2)^2}(\bar{x}_1-\bar{x}_2)^2 + \frac{n_2n_1^2}{(n_1+n_2)^2}(\bar{x}_1-\bar{x}_2)^2$$

$$= n_1s_1^2 + n_2s_2^2 + \frac{n_1n_2}{(n_1+n_2)}(\bar{x}_1-\bar{x}_2)^2$$

Hence

$$s^2 = \frac{n_1s_1^{\,2} + n_2s_2^2}{n_1+n_2} + \frac{n_1n_2}{(n_1+n_2)^2}(\bar{x}_1-\bar{x}_2)^2 \qquad \ldots (4.27a)$$

***Corollary* 4.4:** *An alternative form of the combined variance is as follows:*

$$s^2 = \frac{n_1(s_1^2+d_1^2)+n_2(s_2^2+d_2^2)}{n_1+n_2}$$

where $d_1 = (\bar{x}_1 - \bar{x})$ and $d_2 = (\bar{x}_2 - \bar{x})$

***Corollary* 4.5:** *If $\bar{x}_1 = \bar{x}_2$, then the expression (4.27a) above reduces to*

$$s^2 = \frac{n_1s_1^2 + n_2s_2^2}{n_2+n_2} \qquad \ldots (4.27b)$$

If further, $n_1 = n_2$, then the variance of the composite set is simply the arithmetic mean of the individual sets:

$$s^2 = \frac{s_1^2+s_2^2}{2} \qquad \ldots (4.27c)$$

***Corollary* 4.6:** *The above results can be logically extended to k sets of observations. Thus if the first set contains n_1 observations, second set n_2 observations, ..., nth set contains n_k observations, then*

$$s^2 = \frac{n_1(s_1^{\,2}+d_1^{\,2})+n_2(s_2^{\,2}+d_2^{\,2})+..+n_k(s_k^{\,2}+d_k^{\,2})}{n_1+n_2+..+n_k} \qquad \ldots (4.27d)$$

Example 4.22: A set consisting of 40 observations has a mean 25 and a variance 25. The mean and variance of 15 observations out of the same set are 20 and 2 respectively. Find the mean and variance of the remaining 25 observations.

Solution: If $\bar{x}_1$ and $\bar{x}_2$ are the means of the first and the second set of n_1 and n_2 observations respectively, the combined mean $\bar{x}$ of these two sets is given by

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1+n_2} \qquad \ldots (4.27e)$$

Here we are given the following computed values

	First sample	Second sample	Total
Size	n_1=15	n_2=25	n=40
Mean	$\bar{x}_1 = 20$	$\bar{x}_2 = ?$	$\bar{x} = 25$
Variance	$s_1^2 = 2$	$s_2^2 = ?$	s^2=25

From (4.27e) we obtain $\bar{x}_2$ on substituting the required values from the table

$$25 = \frac{15\times20 + 25\times\bar{x}_2}{15+25}$$

which yields $\bar{x}_2 = 28$. To find the variance, we use the formula

$$s^2 = \frac{n_1s_1^{\,2} + n_2s_2^{\,2}}{n_1+n_2} + \frac{n_1n_2}{(n_1+n_2)^2}(\bar{x}_1-\bar{x}_2)^2$$

Substituting the given values in the above formula

$$25 = \frac{15\times2 + 25\times s_2^{\,2}}{15+25} + \frac{15\times25}{(15+25)^2}(20-28)^2$$

Solving, we obtain $s_2^2 = 14.8$

***Theorem* 4.7:** *For a set of n values $x_1, x_2, ..., x_n$, the mean deviation is the minimum when the deviations are taken from the median.*

Proof: Let n be even such that n=2p. Let the observations be arranged as follows:

$$x_1, x_2, ..., x_p, \overset{\tilde{m}}{\uparrow} x_{p+1}, ... x_k, \overset{a}{\uparrow} x_{k+1}, ... x_{2p}.$$

Let m be the median of this series lying between x_p and x_{p+1}. Further, let a be an arbitrary value, greater than $\tilde{m}$, that lies between x_k and x_{k+1}, where $k>p$.

The mean deviation from the median m is

$$M_d(\tilde{m}) = \frac{\sum_{i=1}^{2p} |x_i - \tilde{m}|}{n} = \frac{D_1}{n} \text{ (say)}$$

The mean deviation from a is

$$M_d(a) = \frac{\sum_{i=1}^{2p} |x_i - a|}{n} = \frac{D_2}{n} \text{ (say)}$$

Now $$D_1 = \sum_{i=1}^{2p} |x_1 - \tilde{m}| = \sum_{i=1}^{p} |x_i - \tilde{m}| + \sum_{i=p+1}^{k} |x_i - \tilde{m}| + \sum_{i=k+1}^{2p} |x_i - \tilde{m}|$$

$$= \sum_{i=1}^{p} (\tilde{m} - x_i) + \sum_{i=p+1}^{k} (x_i - \tilde{m}) + \sum_{i=k+1}^{2p} (x_i - \tilde{m})$$

And

$$D_2 = \sum_{i=1}^{2p} |x_1 - a| = \sum_{i=1}^{p} |x_i - a| + \sum_{i=p+1}^{k} |x_i - a| + \sum_{i=k+1}^{2p} |x_i - a|$$

$$= \sum_{i=1}^{p} (a - x_i) + \sum_{i=p+1}^{k} (a - x_i) + \sum_{i=k+1}^{2p} (x_i - a)$$

Subtracting D_1 from D_2,

$$D_2 - D_1 = \sum_{i=1}^{p} (a - \tilde{m}) + \sum_{i=p+1}^{k} (a + \tilde{m} - 2x_i) + \sum_{i=k+1}^{2p} (\tilde{m} - a)$$

$$= p(a - \tilde{m}) + (2p - k)(\tilde{m} - a) + \sum_{i=p+1}^{k} (a + \tilde{m} - 2x_i)$$

$$= (a - \tilde{m})(k - p) + \sum_{i=p+1}^{k} (a + \tilde{m} - 2x_i)$$

$$= \sum_{i=p+1}^{k} (a - \tilde{m}) + \sum_{i=p+1}^{k} (a + \tilde{m} - 2x_i)$$

$$= 2\sum_{i=p+1}^{k} (a - x_i) > 0,$$

(since a is greater than x_i in the range x_{p+1} to x_k).

Hence $D_2 - D_1 > 0$, *i.e.* $D_2 > D_1$. This implies that

$$\frac{D_1}{n} < \frac{D_2}{n},$$

so that

$$M_d(\tilde{m}) < M_d(a)$$

Theorem 4.8: *The variance of first n natural numbers is* $\frac{n^2 - 1}{12}$

Proof: For the variable x assuming n natural numbers 1, 2,....n, the mean is

$$\bar{x} = \frac{1 + 2 + ... + n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

and the variance is

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

$$= \frac{1^2 + 2^2 + ... + n^2}{n} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n^2 - 1}{12}$$

Thus for first 5 natural numbers 1, 2, 3, 4.5, $\bar{x} = 3$ and the variance is

$$s_x^2 = \frac{n^2 - 1}{12} = \frac{5^2 - 1}{12} = 2.$$

Check that the direct computation yields the same value for the variance:

$$s_x^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5} = \frac{10}{2} = 2$$

Theorem 4.9: *If* $\bar{x}$ *and* s *denote respectively the mean and standard deviation of a set of n non-negative values, then*

$$\left(\frac{s}{\bar{x}}\right)^2 < (n-1)$$

This result implies that for set of observations, the square of the coefficient of variation cannot exceed (n-1), where n is the number of observations in the set.

Proof: Let $x_1, x_2, ..., x_n$ be a set of n non-negative values. The square of the sum of these values can be expressed as follows

$$(x_1 + x_2 + ... + x_n)^2 = x_1^2 + x_2^2 + ... + x_n^2 + \sum_{i \neq j} x_i\, x_j$$

That is

$$\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i \neq j} x_i x_j \qquad \ldots (4.28)$$

Since x_i's are non-negative, $\sum_{i \neq j} x_i\, x_j \geq 0$. Hence (4.33) can be expressed as

$$\left(\sum_{i=1}^{n} x_i\right)^2 \geq \sum_{i=1}^{n} x_i^2 \qquad \ldots (4.29)$$

Subtracting $(\Sigma x_i)^2 / n$ from both sides of (4.34)

$$\left(\sum x_i\right)^2 - \frac{\left(\sum x_i\right)^2}{n} \geq \sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}$$

or

$$\left(\sum x_i\right)^2 \left(1 - \frac{1}{n}\right) \geq ns^2 \Rightarrow \frac{n-1}{n}(n\bar{x})^2 \geq ns^2$$

from which

$$\bar{x}\sqrt{(n-1)} \geq s$$

A simple rearrangement of the result yields

$$\left(\frac{s}{\bar{x}}\right)^2 < (n-1)$$

***Theorem* 4.10:** *The sum of squares of a set of n observations is given by*

$$\frac{1}{n}\sum_{i \neq j}(x_i - x_j)^2, \qquad (i, j = 1, 2, \ldots, n)$$

Proof: The sum of squares of n observations $x_1, x_2, \ldots, x_n$ is defined as $\Sigma(x_i - \bar{x})^2$. Thus we require to prove that

$$n\sum(x_i - \bar{x})^2 = \sum(x_i - x_j)^2 \qquad \ldots (4.30)$$

The right hand side of (4.30) can be expanded as follows:

$$\begin{aligned}&(x_1 - x_2)^2 + (x_1 - x_3)^2 + ... + (x_1 - x_n)^2\\ &+ (x_2 - x_3)^2 + (x_2 - x_4)^2 + ... + (x_2 - x_n)^2\\ &+ (x_3 - x_4)^2 + ... + (x_3 - x_n)\end{aligned}$$

$$+ (x_{n-1} - x_n)^2$$

Performing the squares and collecting the similar quantities, we get

$$\begin{aligned}\sum(x_i - x_j)^2 &= (n-1)(x_1^2 + x_2^2 + ... + x_n^2)\\ &\quad - 2(x_1x_2 + x_1x_3 + ... + x_1x_n + x_2x_3 + ... + x_2x_n + ... + x_{n-1}x_n)\\ &= (n-1)\sum x_i^2 - 2\sum x_i x_j\end{aligned}$$

The left-hand side of (4.30) can be written as

$$\begin{aligned}n\sum(x_i - \bar{x})^2 &= n\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\\ &= n\left(\sum x_i^2 - \frac{\sum x_i^2 + 2\sum_{i<j} x_i x_j}{n}\right)\\ &= n\sum x_i^2 - \sum x_i^2 - 2\sum x_i x_j\\ &= (n-1)\sum x_i^2 - 2\sum x_i x_j = \text{RHS}\end{aligned}$$

Hence the proof.

***Theorem* 4.11:** *For three numbers $x_1, x_2\, x_3$, the variance can be expressed as follows:*

$$s_x^2 = \frac{(x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_1)^2}{9} = \frac{1}{9}\sum_{i \neq j}^{3}(x_i - x_j)^2$$

Proof: If $\bar{x}$ is the mean of three observations $x_1, x_2\, x_3$, then the usual formula for computing their variance is

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}{3} = \frac{1}{3}\sum_{i=1}^{3}(x_i - \bar{x})^2$$

where

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}$$

Using Theorem 4.10

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i\neq j}^{n}(x_i-x_j)^2$$

Dividing throughout by n

$$\frac{1}{n}\sum(x_i-\bar{x})^2=\frac{1}{n^2}\sum(x_i-x_j)^2 \qquad \ldots (4.31)$$

For n=3, the equation (4.31) above can be written as

$$\frac{1}{3}\sum(x_i-\bar{x})^2=\frac{1}{9}\sum(x_i-x_j)^2, \qquad (i=1,2,3)$$

Hence the result.

You can directly establish the above result as follows:

Let x_1, x_2 and x_3 be three values with mean $\bar{x}$. Then the variance of these three quantities s_x^2 is

$$s_x^2=\frac{x_1^2+x_2^2+x_3^2}{3}-\left(\frac{x_1+x_2+x_3}{3}\right)^2$$

$$=\frac{(x_1-x_2)^2+(x_2-x_3)^2+(x_3-x_1)^2}{9}$$

$$=\frac{1}{9}\sum_{i\neq j}^{3}(x_i-x_j)^2$$

Thus for three numbers 2, 5, 8, the variance can be found using the above formula:

$$s_x^2=\frac{(2-5)^2+(5-8)^2+(8-2)^2}{9}=\frac{54}{9}=6$$

***Theorem 4.12**: For two unequal observations, $M_d(\bar{x})$ =SD=R/2, where $M_d(\bar{x})$ =Mean deviation about mean, SD=Standard deviation and R=Range*

Proof: Let the two observations be x_1 and x_2 $(x_1>x_2)$. By definition, the mean, range, and the standard deviation of x_1 and x_2 are respectively

$$\bar{x}=\frac{x_1+x_2}{2}, \quad R=x_1-x_2 \text{ and } s_x=\sqrt{\frac{\sum(x_i-\bar{x})^2}{2}}.$$

Now

$$M_d(\bar{x})=\frac{\sum|x_i-\bar{x}|}{2}=\frac{|x_1-\bar{x}|+|x_2-\bar{x}|}{2}$$

$$=\frac{\left|x_1-\frac{x_1+x_2}{2}\right|+\left|x_2-\frac{x_1+x_2}{2}\right|}{2}$$

$$=\frac{\left|\frac{x_1-x_2}{2}\right|+\left|\frac{x_2-x_1}{2}\right|}{2}$$

$$=\frac{\frac{x_1-x_2}{2}+\frac{x_1-x_2}{2}}{2}=\frac{x_1-x_2}{2}=\frac{R}{2}$$

Again

$$s_x^2=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2}{2}$$

$$=\left(\frac{x_1-x_2}{2}\right)^2=\left(\frac{R}{2}\right)$$

so that $s_x=\frac{R}{2}$

Combining the results, it follows that $s_x=M_d=R/2$. This relation also demonstrates that $R>s$ and also that $R>M_d$.

4.8 THE MOMENTS

The term **moment** in statistical usage is analogous to **moment of forces** in physics. In statistics, moments are certain constant values in a given distribution and as we will see, they clearly fall under descriptive statistics. Because of this nature, the moments help us to ascertain the nature and form of the underlying distribution.

Moments of a distribution may be calculated from arithmetic mean of the distribution or from any arbitrarily chosen value including zero (**origin**). When the moments are computed about the arithmetic mean of the distribution, we call them **moments about mean**, or **central moments**. When they are computed about an arbitrary value, we call them **raw moments**. When they are computed about zero, they are called **moment about origin** or **raw moments** You can compute an infinite number of moments for a given distribution, but in practice, we need only four to investigate the form and characteristics of a distribution.

4.8.1 Moments about an Arbitrary Value

Consider a variable X, assuming values $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$. Let 'a' be any arbitrarily chosen value. Then the first four raw moments about the value 'a', designated by μ_1', μ_2', μ_3' and μ_4' are defined as

$$\mu_1' = \frac{1}{n}\sum(x_i - a), \quad \mu_2' = \frac{1}{n}\sum(x_i - a)^2$$

$$\mu_3' = \frac{1}{n}\sum(x_i - a)^3, \text{ and } \mu_4' = \frac{1}{n}\sum(x_i - a)^4$$

4.8.2 Central Moments

Replacing 'a' by $\bar{x}$ in the above expressions, we arrive at what is referred to as the **central moments** or **moments about the mean**. These moments are usually denoted by μ_1, μ_2, μ_3 and μ_4 and are defined as:

$$\mu_1 = \frac{\sum(x_i - \bar{x})}{n}, \qquad \mu_2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

$$\mu_3 = \frac{\sum(x_i - \bar{x})^3}{n}, \text{ and } \mu_4 = \frac{\sum(x_i - \bar{x})^4}{n}$$

Clearly, the first central moment μ_1 is zero and the second central moment μ_2 is the variance s^2. The third and fourth moments, as we shall see later, assist us to determine the shape characteristics of the distribution.

If again 'a' is replaced by 'zero' we get moments about the origin. Denoting these moments respectively by ν_1, ν_2, ν_3 and ν_4, we get the following expressions for the first four moments about the origin:

$$\nu_1 = \frac{\sum x_i}{n}, \quad \nu_2 = \frac{\sum x_i^2}{n}, \quad \nu_3 = \frac{\sum x_i^3}{n} \text{ and } \nu_4 = \frac{\sum x_i^4}{n}$$

Note that $\nu_1 = \bar{x}$, showing that the first moment about the origin is the arithmetic mean of the distribution. In general, the rth moment about the mean (μ_r), about 'a' (μ_r') and about the origin (ν_r) respectively are symbolically expressed as follows:

$$\mu_r = \frac{1}{n}\sum(x_i - \bar{x})^r, \mu_r' = \frac{1}{n}\sum(x_i - a)^r \text{ and } \nu_r = \frac{1}{n}\sum x_i^r$$

a frequency distribution, the r-th moment defined above will be pressed as follows:

r^{th} moment about 'a': $\mu_r' = \frac{1}{n}\sum f_i(x_i - a)^r$,

r^{th} central moment: $\mu_r = \frac{1}{n}\sum f_i(x_i - \bar{x})^r$,

r^{th} moment about origin: $\nu_r = \frac{1}{n}\sum f_i x_i^r$,

ere $n = \sum f_i$

ample 4.23 Compute the first four central moments for the following quency distribution

x_i :	2	3	4	5	6
f_i :	1	3	7	3	1

lution: We prepare the following table for computing the moments:

x_i	f_i	$x_i - \bar{x}$	$f_i(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^3$	$f_i(x_i - \bar{x})^4$
2	1	−2	−2	4	−8	16
3	3	−1	−3	3	−3	3
4	7	0	0	0	0	0
5	3	+1	+3	3	+3	3
6	1	+2	+2	4	+8	16
Total	15	-	0	14	0	38

Here $\bar{x} = 4$

Thus

$$\mu_1 = \frac{\sum f_i(x_i - \bar{x})}{n} = 0$$

$$\mu_2 = \frac{\sum f_i(x_i - \bar{x})^2}{n} = \frac{14}{15} = 0.933$$

$$\mu_3 = \frac{\sum f_i(x_i - \bar{x})^3}{n} = 0$$

$$\mu_4 = \frac{\sum f_i(x_i - \bar{x})^4}{n} = \frac{38}{15} = 2.533$$

4.9 CENTRAL MOMENTS IN TERMS OF RAW MOMENTS

4.9.1 Central Moment and Moment about Arbitrary Value

The computation of central moments can be effected through the computation of the raw moments about any arbitrary origin. This implies that a relationship does exist between central moments and raw moments. We show these relationships below only for the first four moments:

Let X be a discrete variable assuming values $x_1, x_2, \ldots x_n$ with mean $\bar{x}$. Let 'a' be an arbitrary value. Then

$$\mu_1' = \frac{\sum(x_i - a)}{n} = \frac{\sum x_i - na}{n} = \bar{x} - a$$

$$\mu_1 = \frac{\sum(x_i - \bar{x})}{n} = \frac{\sum x_i - n\bar{x}}{n} = 0$$

$$\mu_2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{\sum\{(x_i - a) - (\bar{x} - a)\}^2}{n}$$

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

$$= \mu_2' - 2\mu_1'\mu_1' + \mu_1'^2 = \mu_2' - \mu_1'^2$$

$$\mu_3 = \frac{\sum(x_i - \bar{x})^3}{n} = \frac{\sum\{(x_i - a) - (\bar{x} - a)\}^3}{n}$$

$$= \frac{\sum(x_i - a)^3}{n} - 3\frac{\sum(x_i - a)^2}{n}(\bar{x} - a)$$

$$+ 3\frac{\sum(x_i - a)}{n}(\bar{x} - a)^2 - (\bar{x} - a)^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 3\mu_1'\mu_1'^2 - \mu_1'^3$$

$$= \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\mu_4 = \frac{\sum(x_i - \bar{x})^4}{n} = \frac{\sum\{(x_i - a) - (\bar{x} - a)\}^4}{n}$$

$$= \frac{\sum(x_i - a)^4}{n} - 4\frac{\sum(x_i - a)^3}{n}(\bar{x} - a)$$

$$+ 6\frac{\sum(x_i - a)^2}{n}(\bar{x} - a)^2 - 4\frac{\sum(x_i - a)}{n}(\bar{x} - a)^3 + (\bar{x} - a)^4$$

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 4\mu_1'\mu_1'^3 + \mu_1'^4$$

$$= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

In general

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

$$= \frac{\sum(u_i - d)^r}{n}, \quad \text{where } u_i = x_i - a,\ d = \bar{x} - a$$

$$= \frac{1}{n}\left[\sum u_i^r - {}^rC_1 d\sum u_i^{r-1} + {}^rC_2 d^2\sum u_i^{r-2} - \ldots + (-1)\sum d^r\right]$$

$$= \mu_r' - {}^rC_1 d\mu_{r-1}' + {}^rC_2 d^2\mu_{r-2}' - \ldots + (-1)^r d^r$$

$$= \mu_r' - {}^rC_1\mu_1'\mu_{r-1}' + {}^rC_2\mu_1'^2\mu_{r-2}' - \ldots + (-1)^r\mu_1'^r$$

Moments of desired order can now be obtained substituting r=1, 2, 3, 4,....in the expression above

Example 4.24: Compute first four central moments for the observations 7, 8, 9, 12, and 14.

Solution: The mean of these observations is 10. The moments now can be computed by constructing the following table.

i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
1	7	−3	9	−27	81
2	8	−2	4	- 8	16
3	9	−1	1	−1	1
4	12	2	4	8	16
5	14	4	16	64	256
Total	50	0	34	36	370

$\bar{x} = 10$

The central moments are

$$\mu_1 = \frac{\sum(x_i - \bar{x})}{n} = \frac{0}{5} = 0, \qquad \mu_2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{34}{5} = 6.8$$

$$\mu_3 = \frac{\sum(x_i - \bar{x})^3}{n} = \frac{36}{5} = 7.2, \quad \mu_4 = \frac{\sum(x_i - \bar{x})^4}{n} = \frac{370}{5} = 74$$

Example 4.25: Compute the first four moments about an arbitrary value 12 using the data in Example 4.24 and hence the central moments.

Solution: The accompanying table shows the computational procedures with a=12.

i	x_i	(x_i-12)	$(x_i-12)^2$	$(x_i-12)^3$	$(x_i-12)^4$
1	7	−5	25	−125	625
2	8	−4	16	−64	256
3	9	−3	9	−27	81
4	12	0	0	0	0
5	14	2	4	8	16
Total	50	−10	54	−208	978

$$\mu_1' = \frac{\sum(x_i-12)}{n} = -\frac{10}{5} = -2, \qquad \mu_2' = \frac{\sum(x_i-12)^2}{n} = \frac{54}{5} = 10.8$$

$$\mu_3' = \frac{\sum(x_i-12)^3}{n} = \frac{-208}{5} = -41.6, \quad \mu_4' = \frac{\sum(x_i-12)^4}{n} = \frac{978}{5} = 195.6$$

Hence

$$\begin{aligned}\mu_2 &= \mu_2' - \mu_1'^2 \\ &= 10.8-(-2)^2 \\ &= 6.8\end{aligned}$$

$$\begin{aligned}\mu_3 &= \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 \\ &= -41.6-3(-2)(10.8)+2(-2)^3 \\ &= 7.2\end{aligned}$$

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_1'\mu_3' + 6\mu_1'^2\mu_2' - 3\mu_1'^4 \\ &= 195.6-4(-2)(-41.6)+6(-2)^2(10.8)-3(-2)^4 \\ &= 74\end{aligned}$$

The central moments calculated in this example from raw moments are the same as those obtained directly in Example 4.14 as ought to be.

Example 4.26: The accompanying table shows the distribution of 131 employees of a department store by their hourly wages in US dollar. Compute first four moments about an arbitrary value 10 and hence the corresponding central moments.

Wages in US $ (x_i)	Number of employees (f_i)
5	1
6	2
7	5
8	10
9	20
10	51
11	22
12	11
13	5
14	3
15	1

Let us first calculate moments about an arbitrary origin set at 10. The necessary calculations are shown in the table below:

x_i	f_i	x_i-10	$f_i(x_i-10)$	$f_i(x_i-10)^2$	$f_i(x_i-10)^3$	$f_i(x_i-10)^4$
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	0	0	0	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
Total	131	-	8	346	74	3718

The required moments are

$$\mu_1' = \frac{\sum f_i(x_i-10)}{n} = \frac{8}{131} = 0.06, \quad \mu_2' = \frac{\sum f_i(x_i-10)^2}{n} = \frac{346}{131} = 2.64$$

$$\mu_3' = \frac{\sum f_i(x_i-10)^3}{n} = \frac{74}{131} = 0.56, \quad \mu_4' = \frac{\sum f_i(x_i-10)^4}{n} = \frac{3718}{131} = 28.38$$

Hence the moments about the mean are

$$\mu_2 = \mu_2' - \mu_1'^2 = 2.64-(.06)^2 = 2.64$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = .56-3(.06)(2.64)+2(.06)^3 = .08$$

$$\begin{aligned}\mu_4 &= \mu_4' - 4\mu_1'\mu_3' + 6\mu_1'^2\mu_2' - 3\mu_1'^4 \\ &= 28.38-4(.06)(.56)+6(.06)(2.64)-3(.06)^4 = 29.19\end{aligned}$$

4.9.2 Central Moments and Moments about Origin

The central moments bear the following relationships with those of the moments about the origin:

$$\mu_1 = \frac{\sum(x_i - \bar{x})}{n} = \frac{\sum x_i}{n} - \bar{x} = \nu_1 - \nu_1 = 0$$

$$\mu = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \bar{x}^2 = \nu_2 - \nu_1^2$$

$$\mu = \frac{\sum(x_i - \bar{x})^3}{n} = \frac{\sum x_i^3}{n} - 3\bar{x}\frac{\sum x_i^2}{n} + 3\bar{x}^2\frac{\sum x_i}{n} - \bar{x}^3$$

$$= \nu_3 - 3\nu_1\nu_2 + 3\nu_1^2\nu_1 - \nu_1^3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$$

And

$$\mu_4 = \frac{\sum(x_i - \bar{x})^4}{n}$$

$$= \frac{\sum x_i^4}{n} - 4\bar{x}\frac{\sum x_i^3}{n} + 6\bar{x}^2\frac{\sum x_i^2}{n} - 4\bar{x}^3\frac{\sum x_i}{n} + \bar{x}^4$$

$$= \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 4\nu_1^4 + \nu_1^4$$

$$= \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$$

Example 4.27: Using the following data, compute the first four moments about the origin and hence the central moments.

x_i	0	1	2	3	4	5	6	7
f_i	364	376	218	89	33	13	2	1

Solution: The table below shows the necessary calculations:

x_i	f_i	f_ix_i	$f_ix_i^2$	$f_ix_i^3$	$f_ix_i^4$
0	364	0	0	0	0
1	376	376	376	376	376
2	218	436	872	1744	3488
3	89	267	801	2403	7209
4	33	132	528	2112	8448
5	13	65	325	1625	8125
6	2	12	72	432	2592
7	1	7	49	343	2401
Total	1096	1295	3023	9035	32639

Thus

$$\nu_1 = \frac{\sum f_i x_i}{n} = \frac{1295}{1096} = 1.18, \qquad \nu_2 = \frac{\sum f_i x_1^2}{n} = \frac{3023}{1096} = 2.76$$

$$\nu_3 = \frac{\sum f_i x_1^3}{n} = \frac{9035}{1096} = 8.24, \qquad \nu_4 = \frac{\sum f_i x_1^4}{n} = \frac{32639}{1096} = 29.78$$

Hence the central moments are

$$\mu_1 = 0$$

$$\mu_2 = \nu_2 - \nu_1^2 = 2.76 - (1.18)^2 = 1.37.$$

$$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3 = 8.24 - 3(2.76)(1.18) + 2(1.18)^3 = 1.76$$

$$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4$$

$$= 29.78 - 4(8.24)(1.18) + 6(2.76)(1.18)^2 - 3(1.18)^4 = 8.13$$

4.10 EFFECTS OF CHANGES IN ORIGIN AND SCALE ON MOMENTS

In discussing measures of central tendency and dispersion, we demonstrated that certain algebraic manipulations, such as changes in the origin and scale of measurement in the original variable, make the calculations of these measures easier, convenient and time saving, without any imposition on the accuracy of the results. For moments too, the same manipulations are possible without affecting the results. This is shown below.

The *rth* central moment for a discrete variable x is defined as follows:

$$\mu_r(x) = \frac{1}{n}\sum(x_i - \bar{x})^r \qquad \ldots (a)$$

Let u be a new variable taking on values $u_1, u_2, \ldots, u_n$ and for the ith value of u, we define

$$u_i = \frac{x_i - a}{h} \qquad \ldots (b)$$

where a and h are two arbitrary constants, referred to as origin and scale factors. From (b) above, we have

$$x_i = hu_i + a \Rightarrow \bar{x} = h\bar{u} + a \qquad \ldots (c)$$

Substituting (c) in (a)

$$\mu_r(x) = \frac{1}{n}\sum(x_i - \bar{x})^r = \frac{1}{n}\sum(hu_i - h\bar{u})^r = h^r\mu_r(u)$$

This shows that change in the origin does not have any effect on the moments but scale has an effect in that the rth moment obtained under the new variable u gets multiplied by an amount h^r to be equal to the moment of the original variable x.

Specifically, for r=1, 2, 3 and 4

$$\mu_1(x) = h\mu_1(u) = 0, \quad \mu_2(x) = h^2\mu_2(u),$$
$$\mu_3(x) = h^3\mu_3(u), \quad \mu_4(x) = h^4\mu_4(u).$$

The following example illustrates applications of the above relationships.

Example 4.28: A variable x takes on values 25, 30, 40, 50, 75, Verify the relation $\mu_r(x) = h^r\mu_r(u)$ for r=1, 2, 3, 4.

Solution: This example is designed to demonstrate the effect of changes in the origin and scale of measurement on the central moment directly without computing the raw moments. To accomplish this, we compute $\bar{x}$ and construct the accompanying table.

i	x_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^3$	$(x_i-\bar{x})^4$
1	25	−19	361	−6859	130321
2	30	−14	196	−2744	38416
3	40	−4	16	−64	256
4	50	+6	36	+216	1296
5	75	+31	961	+29791	923521
Total	220	0	1570	20340	1093810

$^*\bar{x}$=44

$$\mu_1(x) = \frac{\sum(x_i-\bar{x})}{n} = 0$$

$$\mu_2(x) = \frac{\sum(x_i-\bar{x})^2}{n} = \frac{1570}{5} = 314$$

$$\mu_3(x) = \frac{\sum(x_i-\bar{x})^3}{n} = \frac{20340}{5} = 4068$$

$$\mu_4(x) = \frac{\sum(x_i-\bar{x})^4}{n} = \frac{1093810}{5} = 218762$$

The x variable is now changed to a new variable u defined as $u = (x-a)/h$ with a=40 and h=5, and the following table is constructed to obtain the moments $\mu_r(u)$.

x_i	u_1	$u_i-\bar{u}$	$(u_i-\bar{u})^2$	$(u_i-\bar{u})^3$	$(u_i-\bar{u})^4$
25	−3	−3.8	14.44	−54.872	208.5136
30	−2	−2.8	7.84	−21.952	61.4656
40	0	−0.8	0.64	−0.512	0.4096
50	+2	+1.2	1.44	+1.728	2.0736
75	+7	+6.2	38.44	+238.328	1477.6336
Total	+4	0	62.8	162.72	1750.096

$^*\bar{u} = 0.8$

$$\mu_1(u) = \frac{\sum(u_i-\bar{u})}{n} = 0$$

$$\mu_2(u) = \frac{\sum(u_i-\bar{u})^2}{n} = \frac{62.8}{5} = 12.56$$

$$\mu_3(u) = \frac{\sum(u_i-\bar{u})^3}{n} = \frac{162.72}{5} = 32.544$$

$$\mu_4(u) = \frac{\sum(u_i-\bar{u})^4}{n} = \frac{1750.096}{5} = 350.02$$

Multiplying the moments of u values successively by h, h^2, h^3, and h^4, we obtain.

$$h\mu_1(u) = 0 = \mu_1(x)$$
$$h^2\mu_2(u) = 5^2(12.56) = 314 = \mu_2(x)$$
$$h^3\mu_3(u) = 5^3(32.544) = 4068 = \mu_3(x)$$
$$h^4\mu_4(u) = 5^4(350.02) = 218762 = \mu_4(x)$$

This numerically establishes the relationships.

We now turn to show that one can establish the same relationships calculating the raw moments. This involves computing raw moments of the x variable through u variable, where u is a transformed variable as defined before. Since a is an arbitrary constant, we choose a=0. That is the transformation is $u_i = x_i/h$ so that $x_i = hu_i$ in which case

$$\frac{\sum x_i^r}{n} = \frac{h^r\sum u_i^r}{n} \quad \ldots (a)$$

For a frequency distribution, the above relationship can be expressed as

$$\frac{\sum f_i x_i^r}{n} = \frac{h^r\sum f_i u_i^r}{n} \quad \ldots (b)$$

The notational representation of (a) or (b) is

$$\nu_r(x) = h^r \nu_r(u)$$

Now we express $\nu_r(x)$ in terms of $\nu_r(u)$ and obtain the central moments in terms of $\nu_r(x)$:

$$\mu_1 = 0$$
$$\mu_2(x) = \nu_2(x) - \nu_1^2(x)$$
$$\mu_3(x) = \nu_3(x) - 3\nu_2(x)\nu_1(x) + 2\nu_1^3(x)$$
$$\mu_4(x) = \nu_4(x) - 4\nu_3(x)\nu_1(x) + 6\nu_2(x)\nu_1^2(x) - 3\nu_1^4(x)$$

Example 4.29: Compute the first four moments about the origin for the observations in Example 4.28 and hence the central moments.

Solution: Let h=5. With this change in the scale, the u values are calculated and necessary computations are shown in the following table:

i	x_i	$u_i = x_i/h$	u_i^2	u_i^3	u_i^4
1	25	5	25	125	625
2	30	6	36	216	1296
3	40	8	64	512	4096
4	50	10	100	1000	10000
5	75	15	225	3375	50625
Total	220	44	450	5228	66642

$$\nu_1(x) = h\nu_1(u) = \frac{5\sum u_i}{n} = \frac{5\times 44}{5} = 44$$

$$\nu_2(x) = h^2\nu_2(u) = \frac{25\sum u_i^2}{n} = \frac{25\times 450}{5} = 2250$$

$$\nu_3(x) = h^3\nu_3(u) = \frac{125\sum u_i^3}{n} = \frac{125\times 5228}{5} = 130700$$

$$\nu_4(x) = h^4\nu_4(u) = \frac{625\sum u_i^4}{n} = \frac{625\times 66642}{5} = 8330250$$

Hence the central moments are

$$\mu_2(x) = \nu_2(x) - \nu_1^2(x) = 2250 - 44^2 = 314.$$
$$\mu_3(x) = \nu_3(x) - 3\nu_2(x)\nu_1(x) + 2\nu_1^3(x)$$

$$= 130700 - 3(2250)(44) + 2(44)^3 = 4068$$
$$\mu_4(x) = \nu_4(x) - 4\nu_3(x)\nu_1(x) + 6\nu_2(x)\nu_1^2(x) - 3\nu_1^4(x)$$
$$= 8330250 - 4(130700)(44) + 6(2250)(44)^2 - 3(44)^4$$
$$= 218762$$

Check that these values are the same as that obtained in Example 4.26 directly.

Example 4.30: Given the following data on the ages of 75 persons in years. Calculate the first four central moments changing suitably the origin and scale.

Age	Number of persons
07.5–10.5	3
10.5–13.5	7
13.5–16.5	12
16.5–19.5	15
19.5–22.5	18
22.5–25.5	10
25.5–28.5	8
28.5–31.5	2
Total	75

Solution: Define a new variable u with a=21 and h=3, so that

$$u_i = \frac{x_i - 21}{3}$$

The accompanying table is constructed to facilitate the computation of the moments.

x_i	f_i	u_i	f_iu_i	$f_iu_i^2$	$f_iu_i^3$	$f_iu_i^4$
9	3	−4	−12	48	−192	768
12	7	−3	−21	63	−189	567
15	12	−2	−24	48	−96	192
18	15	−1	−15	15	−15	15
21	18	0	0	0	0	0
24	10	+1	+10	10	+10	10
27	8	+2	+16	32	+64	128
30	2	+3	+6	18	+54	162
Total	75	–	−40	234	−364	1842

$$\mu_1'(x) = h\mu_1'(u) = \frac{h\sum f_iu_i}{n} = 3\times\frac{(-40)}{75} = -1.6$$

$$\mu_2'(x) = h^2\mu_2'(u) = \frac{h^2\sum f_i u_i^{\,2}}{n} = 3^2 \times \frac{(234)}{75} = 28.08$$

$$\mu_3'(x) = h^3\mu_3'(u) = \frac{h^3\sum f_i u_i^{\,3}}{n} = 3^3 \times \frac{(-364)}{75} = -131.04$$

$$\mu_4'(x) = h^4\mu_4'(u) = \frac{h^4\sum f_i u_i^{\,4}}{n} = 3^4 \times \frac{(1842)}{75} = 1989.36$$

Hence,

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 28.08 - (-1.6)^2 = 25.52$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= -131.04 - 3\times 28.08\times(-1.6) + 2(-1.6)^3 = -4.49$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 1989.36 - 4\times(-131.04)\times(-1.6)$$

$$+ 6\times 28.08\times(-1.6)^2 - 3\times(-1.6)^4$$

$$= 1562.35$$

Example 4.31: The first three moments of a distribution about the value 2 of a variable x are 1, 16, and –40. Find the mean, the variance and the third central moment. Also show that the first three moments about x=0 are 3, 24, and 76.

Solution: As stated in the problem

$$\mu_1' = \frac{\sum f_i(x_i - 2)}{n} = 1 \qquad \ldots \text{(a)}$$

$$\mu_2' = \frac{\sum f_i(x_i - 2)^2}{n} = 16. \qquad \ldots \text{(b)}$$

$$\mu_3' = \frac{\sum f_i(x_i - 2)^3}{n} = -40 \qquad \ldots \text{(c)}$$

From relation (a) above

$$\frac{\sum f_i x_i}{n} - 2 = 1 \Rightarrow \bar{x} = 3$$

Hence the variance is $s^2 = \mu_2' - \mu_1'^2 = 16 - 1^2 = 15$

d the third moment is

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = -40 - 3(1)(16) + 2(1)^3 = -86$$

find the moments about the origin (i.e. about x=0), we evaluate $\frac{\sum f_i x_i}{n}$, $\frac{\sum f_i x_i^2}{n}$, and $\frac{\sum f_i x_i^3}{n}$. These quantities are directly obtainable from st three expressions labeled (a), (b) and (c) above.

e first expression (a) yields

$$\nu_1 = \frac{\sum f_i x_i}{n} = \bar{x} = 3$$

om (b)

$$\frac{\sum f_i(x_i - 2)^2}{n} = \frac{\sum f_i(x_i^2 - 4x + 4)}{n} = 16$$

$$\frac{\sum f_i x_i^{\,2}}{n} - 4\frac{\sum f_i x_i}{n} + 4 = 16$$

$$\frac{\sum f_i x_i^{\,2}}{n} - 4\times 3 + 4 = 16$$

$$\frac{\sum f_i x_i^{\,2}}{n} = 24$$

his by definition is the second moment ν_2 about x=0. That is $\nu_2 = 24$

om (c)

$$\frac{\sum f_i x_i^{\,3}}{n} - 6\frac{\sum f_i x_i^{\,2}}{n} + 12\frac{\sum f_i x_i}{n} - 8 = -40$$

$$\frac{\sum f_i x_i^{\,3}}{n} = -40 + 8 + 12(3) + 6(24) = 76$$

which is ν_3, the third moment about the origin. Hence the first three moments about the origin are

$$\nu_1 = 3, \quad \nu_2 = 24 \text{ and } \nu_3 = 76.$$

4.11 SHEPPARD'S CORRECTION FOR MOMENTS

While calculating moments, it is assumed that all the values of the variable in a class lie at the centre of that class. This assumption leads to consider the mid-value of that interval as representing the class. This means that all the values falling in the class coincide with the mid-value. In practice, however, this is not the case. The assumption is a simplifying one and it merely serves as an approximation to facilitate calculation and hence it introduces some errors. This error is sometimes known as **grouping error**, since it arises due to grouping of the values. Obviously, these errors need correction. Sheppard suggested some corrections for these errors. These corrections, as Sheppard observed, apply only to second and fourth moments. Thus if h stands for the class width, the second and fourth corrected moments due to Sheppard are

$$\bar{\mu}_2 = \mu_2 - \frac{h^2}{12} \quad \ldots \text{(a)}$$

$$\bar{\mu}_4 = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4 \quad \ldots \text{(b)}$$

Sheppard's correction is easy to apply once the central moments and the size of the class (h) are known.

The above corrections are valid only when the

(a) frequency distribution is continuous;
(b) frequency curves taper off to zero to both ends.

The second condition implies that the curve should approach the baseline gradually and slowly at each end of the distribution.

As we will from our illustrated examples below that the corrected version of the formula as suggested by Sheppard does not work well particularly for the 4th moment. In recent years some modification has been suggested by Kendall and others to improve upon the 4th moment, which appears to be more consistent and close to the moments obtained from ungrouped data (Western, 2011). The revised version of the 4th moment is as follows:

$$\bar{\mu}_4 = \mu_4 - \frac{1}{2}\mu_2 + \frac{7}{240}h^2 \quad \ldots \text{(c)}$$

Example 4.32: Attempt an adjustment of the moments for the data in Example 4.30 for grouping errors following Sheppard.

Solution: The distribution is continuous and the both ends of the distribution approach zero. Hence we can apply Sheppard's correction to the computed moments.

$$\bar{\mu}_2 = \mu_2 - \frac{h^2}{12} = 25.52 - \frac{3^2}{12} = 24.77$$

$$\bar{\mu}_4 = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4$$

$$= 1562.39 - \frac{3^2}{2}(25.52) + \frac{7}{240}(3)^4$$

$$= 1449.91$$

How do we interpret these corrected moments? The interpretation is simple. It says that if we compute moments from raw data from which the grouped distribution has been formed, we would have probably obtained $\mu_2 = 24.77$ and $\mu_4 = 1449.91$ in place of 25.52 and 1562.39 respectively. Any deviation from these expected values will result from the way the raw data are organized into frequency distributions.

When we apply the modified version of the 4th moment as suggested by Kendall and others, we obtain

$$\bar{\mu}_4 = \mu_4 - \frac{1}{2}\mu_2 + \frac{7}{240}h^2$$

$$= 1562.39 - \frac{1}{2}(25.52) + \frac{7}{240}(3^2)$$

$$= 1549.89$$

This is in fact remains close to the uncorrected moment and differ by a wide margin of 112.48 from the Sheppard's estimate. But we do not exactly know which of two formulae yields the correct moment unless we compute the same from the raw data. The following examples will give us some clue on the validity of the corrections suggested by the above formula.

Example 4.33: Check the validity of Sheppard's corrections for the second and fourth moments with the raw data on age in Table 2.1 and the corresponding grouped data in Example 4.2.

Solution: The arithmetic mean for the raw data on age in Table 2.1 is computed to be $\bar{x} = 39.2$ so that the second, third and fourth central moments are

$$\mu_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{2098}{50} = 41.96$$

$$\mu_3 = \frac{\sum (x_i - \bar{x})^3}{n} = \frac{1918.80}{50} = 38.38$$

$$\mu_4 = \frac{\sum (x_i - \bar{x})^4}{n} = \frac{218389}{50} = 4367.78$$

The group distribution gave a mean of 39.3 with which the central moments were computed. These moments were as follows:

$$\mu_2 = \frac{\sum f_i (x_i - \bar{x})^2}{n} = \frac{2160}{50} = 43.21$$

$$\mu_3 = \frac{\sum f_i (x_i - \bar{x})^3}{n} = \frac{2359.20}{50} = 47.18$$

$$\mu_4 = \frac{\sum f_i (x_i - \bar{x})^4}{n} = \frac{223947}{50} = 4478.94$$

The Sheppard's corrections for the second and fourth moments for the group distribution are

$$\bar{\mu}_2 = \mu_2 - \frac{h^2}{12} = 43.21 - \frac{5^2}{12} = 41.13$$

$$\bar{\mu}_4 = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4$$

$$= 4478.94 - \frac{25}{2}(43.21) + \frac{7}{240}(5)^4$$

$$= 3957.04$$

The modified version (c) of the 4th moment yields

$$\bar{\mu}_4 = \mu_4 - \frac{1}{2}\mu_2 + \frac{7}{240}h^2$$

$$= 4478.94 - \frac{1}{2}(43.21) + \frac{7}{240}(5^2)$$

$$= 4458.06$$

Note that while the second adjusted moment is closer to the one obtained from the raw data, but the fourth adjusted moment due to Sheppard is far apart from the corresponding moments calculated from the raw data, rather it is closer to the one calculated from the group distribution. Hence

eppard's adjustment does not work well for the fourth moment in this rticular instance. Incidentally, the new version of the correction as own in (c) is pleasingly consistent with the one obtained from the raw a. This gives us an impression that the Kendall's correction performs tter so far as the correction of the 4th moment due to grouping is ncerned.

rther note that the value of the third central moment based on the ouped data substantially varies from the one calculated from raw data 7.18 VS 38.38). Thus it is our feeling that third moment also needs rrection for grouping. To further strengthen our feeling, we present one re example.

ample 4.34: Use the data in Example 2.17 to validate the Sheppard's rmula in adjusting the moments for grouping errors.

lution: The moments computed from the raw data presented in Example 17 are

$$\mu_2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{1004.25}{65} = 15.45$$

$$\mu_3 = \frac{\sum (x_i - \bar{x})^3}{n} = \frac{2309.52}{65} = 35.53$$

$$\mu_4 = \frac{\sum (x_i - \bar{x})^4}{n} = \frac{45599.50}{65} = 701.54$$

ence

$$\bar{x} = 18.11.$$

e group distribution gave the same mean value and hence the moments re as follows:

$$\mu_2 = \frac{\sum f_i (x_i - \bar{x})^2}{n} = 15.66$$

$$\mu_3 = \frac{\sum f_i (x_i - \bar{x})^3}{n} = 33.71$$

$$\mu_4 = \frac{\sum f_i (x_i - \bar{x})^4}{n} = 707.52.$$

ile the adjusted moments due to Sheppard were computed as follows:

$$\overline{\mu}_2 = \mu_2 - \frac{h^2}{12} = 14.91$$

$$\overline{\mu}_4 = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7}{240}h^4 = 639.41.$$

This example also demonstrates the same feature: the fourth moment is seriously underestimated by Sheppard's adjustment and the third moment needs considerable correction.

Application of the modified correction gives

$$\overline{\mu}_4 = \mu_4 - \frac{1}{2}\mu_2 + \frac{7}{240}h^2$$

$$= 707.52 - \frac{1}{2}(15.66) + \frac{7}{240}(3^2)$$

$$= 699.95$$

which is more closer to the moment obtained from the raw data. The foregoing examples tend to suggest that the Sheppard's formula seriously underestimates the 4^{th} moment.

One important point is in order. Groupings of data not only have effect on the moments, it may have enormous effect on the other measures of central tendency and dispersion if we fail to organize the raw data in a frequency distribution with appropriate class widths.

4.12 SHAPE CHARACTERISTICS OF A DISTRIBUTION

The study of the shape characteristics of a distribution is of crucial importance in comparing a distribution with other distributions. By shape characteristic of a distribution, we refer to the extent of its **asymmetry** and **peakedness** relative to an agreed upon standard. The asymmetry of a distribution is studies through what we refer to as the measures of skewness, while peakedness of a distribution is studied through the measures of kurtosis. The accompanying sections are devoted to the study of these characteristics of a frequency distribution.

4.12.1 Skewness

The term **skewness** refers to the lack of symmetry. The lack of symmetry in a distribution is always determined with reference to a normal distribution, which is always symmetrical. Any departure of a distribution from symmetry leads to an asymmetric distribution and in such cases, we call this distribution as skewed. The skewness may be either positive or negative. Absence of skewness makes the distribution **symmetrical**.

It is important to emphasize that skewness of a distribution cannot be determined simply by inspection. If you understand the differences between the mean, median and the mode, you should be able to suggest a method for determining whether a distribution is skewed, and if so, the direction of skew. The following graphs illustrate the skewness of a frequency distribution in three different shapes.

(a) Symmetrical distribution:

The type of this distribution is known as normal. One would obtain such a distribution with height, weight, examination scores and many other real life data. An important characteristic of such distributions is that mean, median and mode have identical value.

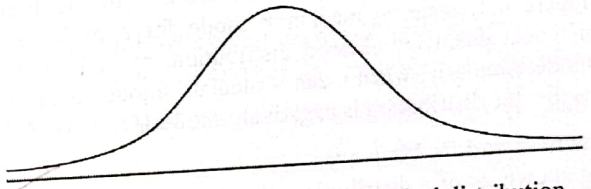


Figure 4.4: Normal curve: symmetrical distribution

(b) Positively skewed distribution

In this distribution, the long tail to the right indicates the presence of extreme values at the positive end of the distribution. This pulls the mean to the right. The frequency curve would look like as follows:

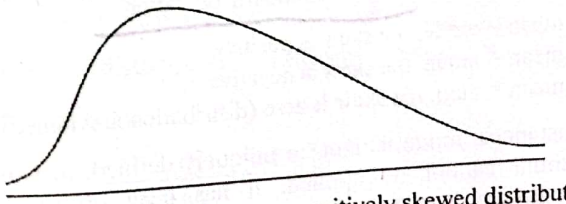


Figure 4.5: Curve representing a positively skewed distribution

This type of distribution is known as positively skewed distribution. These distributions occur with, for example, family size, female age at marriage, wages of the employees etc.

(c) Negatively skewed distribution

In a negatively skewed distribution, the mean is pulled in a negative direction. The frequency curve would look like:

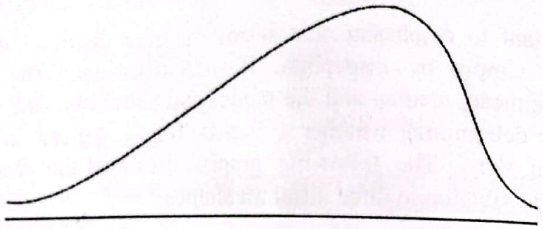


Figure 4.6: Curve representing a negatively skewed distribution

Reaction times for an experiment, daily maximum temperature for a month in winter will result in such a negatively skewed curve.

It is apparent from the above figures that the measures mean, median and mode provide a way to study the shape characteristics of a distribution. As we can see from Figure 4.4, mean = median = mode for a perfectly symmetrical distribution. For a positively skewed distribution, as in Figure 4.5, mean > median > mode. Similarly, when mean < median <mode, as in Figure 4.6, the indication is that the distribution is negatively skewed.

4.12.2 Skewness and its Measures

In studying skewness of a distribution, the first thing that we want to know whether the distribution is positively skewed or negatively skewed. The second thing is to measure the degree of skewness. The simplest measure of skewness is the **Pearson's coefficient of skewness** defined as:

$$S_k(P) = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} \quad \ldots (a)$$

- If mean > mode, the skew is positive
- If mean < mode, the skew is negative
- If mean = mod, the skew is zero (distribution is symmetrical)

In many instances, mode cannot be uniquely defined, in which case, the above formula cannot be applied. It has been observed that for a moderately skewed distribution, the following relationship holds:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median}) \quad \ldots (b)$$

Using this relation, the Pearson's coefficient of skewness is redefined as follows:

$$S_k(P) = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} \quad \ldots (c)$$

Another measure of skewness due to Bowley, is defined in terms of the quartile values. Since there is no difference between the distances of either the first quartile (Q_1) or the third quartile (Q_3) from the median (Q_2) in a symmetrical distribution, any difference in the distances from the median is a reasonable basis for measuring skewness in a distribution. Thus, in terms of the three quartiles Q_1, Q_2 and Q_3, the Bowley's quartile coefficient of skewness is

$$S_k(B) = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \quad \ldots (d)$$

This is evidently a pure number lying between –1 and +1 and is zero for a symmetrical distribution.

- If $Q_3 - Q_2 = Q_2 - Q_1$, skewness = 0 and the distribution is symmetrical
- If $Q_3 - Q_2 > Q_2 - Q_1$, skewness>0 and the distribution is positively skewed
- If $Q_3 - Q_2 < Q_2 - Q_1$, skewness<0 and the distribution is negatively skewed

Example 4.35: The following are some of the measures of locations obtained from a distribution of current flow in ampere based on 125 fuses.

Mean=30.89, Median (Q_2)=30.58, s^2=4.93, s=2.22, Q_1=29.50, and Q_3=32.1.

Compute (i) Pearson's coefficient of skewness $S_k(P)$ and
(ii) Bowley's coefficient of skewness $S_k(B)$

Comment also on the nature of the underlying frequency distribution

Solution: For the above distribution, following Pearson:

$$S_k(P) = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} = \frac{3(30.89 - 30.58)}{2.22} = 0.42$$

Since skewness >0, the distribution is positively skewed.

The Bowley's coefficient is

$$S_k(B) = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{1.52 - 1.08}{2.6} = 0.17$$

Here $Q_3 - Q_2 = 1.52$ and $Q_2 - Q_1 = 1.08$, showing that $Q_3 - Q_2 > Q_2 - Q_1$. Hence the Bowley's coefficient also shows that the distribution is positively skewed.

4.12.3 Use of Moments in Assessing the Skewness of a Distribution

The skewness of a distribution may also be measured by making use of moments. A relative measure of skewness denoted by β_1, is defined as follows:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \ldots \text{(e)}$$

The value of β_1 shall be zero for a perfectly symmetrical distribution. Instead of β_1, Karl Pearson suggested γ_1 to be used as a measure of skewness, where

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\mu_2^{3/2}} \quad \ldots \text{(f)}$$

Obviously, for a symmetrical distribution, $\gamma_1=0$. Clearly, γ_1 measures the skewness more directly as compared to β_1.

The value of β_1 will give the magnitude of the skewness, while the value of μ_3 will determine the nature of the distribution, positive or negative.

Example 4.36: The measure of skewness of a distribution is 0.3. The mode and the median are 50 and 55. Find the mean and standard deviation of the distribution.

Solution: Assuming that the distribution under reference is moderately skewed, we can use the following empirical rule:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median}),$$

from which

$$\text{Mean} - 50 = 3(\text{Mean} - 55)$$

Solving, we have

$$\text{Mean} = 57.5$$

Using Pearson's measure of skewness

$$S_k(\text{P}) = \frac{\text{Mean} - \text{Mode}}{s_x}$$

so that

$$0.3 = \frac{57.5 - 50}{s_x}$$

from which

$$s_x = 25 \Rightarrow s_x^2 = 625$$

4.12.4 Skewness and Empirical Rules

The extent of skewness of a distribution can be examined by empirical rule outlined above. The rule says that if the distribution is not very skewed to the right or left, then

- 68.27% of all measurements will lie within plus or minus one standard deviation of the mean
- 95.44% of all measurements will lie within plus or minus two standard deviation of the mean
- 99.73 % of all measurements will lie within plus or minus three standard deviation of the mean

Any departure from these limits will indicate presence of skewness in the distribution under investigation.

4.12.5 Kurtosis and its Measures

There are considerable variations among symmetrical distributions. For instance, they can differ markedly in terms of **peakedness**. This is what we call **kurtosis**. Kurtosis, as defined by Spiegel (Spiegel: *Theory and Problems of Statistics*) is the degree of peakedness of a distribution, usually taken in relation to a normal distribution. A curve having relatively higher peak than the normal curve, is known as **leptokurtic**. On the other hand, if the curve is more flat-topped than the normal curve, it is called **platykurtic**. A normal curve itself is called **mesokurtic**, which is neither too peaked nor too flat-topped. The following curves illustrate the shape of 3 different types of distribution as mentioned above:

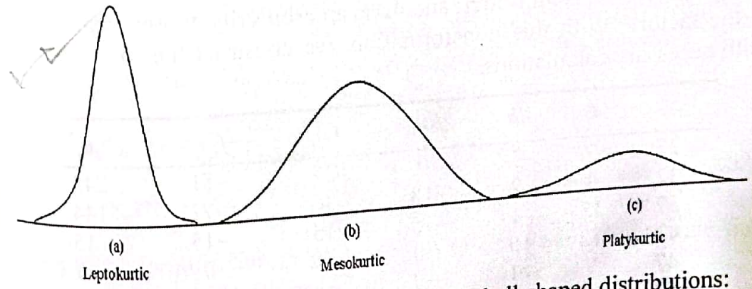


Figure 4.7: Variations among symmetrical or bell-shaped distributions:

Measures of Kurtosis

The most important measure of kurtosis is β_2, defined as the ratio of fourth moment to the square of the second moment:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

This measure is a pure number and is always positive.

For normal distribution $\beta_2 = 3$. When the value of β_2 is greater than 3, the curve is more peaked than the normal curve. When the value of β_2 is less than 3, the curve is less peaked than the normal curve. Based on the β_2 values, we classify a distribution as follows:

- if $\beta_2 > 3$, the distribution is leptokurtic;
- if $\beta_2 < 3$ the distribution is platykurtic;
- if $\beta_2 = 3$, the distribution is mesokurtic.

The deviation of β_2 from 3 is sometimes denoted by γ_2, *i.e* $\gamma_2 = \beta_2 - 3$ and is called excess of kurtosis.

Example 4.37: Compute the first four moments and hence examine the shape characteristics of the age distribution as shown Example 4.2 by all possible measures.

Solution: The following transformation is made in the variable x for computing the raw moments:

$$u_i = \frac{x_i - 42}{5}$$

where x is the class mid-point and 42 is an arbitrarily chosen value while 5 is the factor. With this transformation, we construct the following table with necessary calculations.

x_i	f_i	u_i	f_iu_i	$f_iu_i^2$	$f_iu_i^3$	$f_iu_i^4$
27	3	–3	–9	27	–81	243
32	9	–2	–18	36	–72	144
37	15	–1	–15	15	–15	15
42	12	+0	+0	0	0	0
47	7	+1	+7	7	+7	7
52	4	+2	+8	16	+32	64
Total	50	–	–27	101	–129	473

The raw moments about 42 ar

$$\mu_1'(x) = h\mu_1'(u) = \frac{h\sum f_iu_i}{n} = 5 \times \frac{(-27)}{50} = -2.7$$

$$\mu_2'(x) = h^2\mu_2'(u) = \frac{h^2\sum f_iu_i^2}{n} = 5^2 \times \frac{(101)}{50} = 50.5$$

$$\mu_3'(x) = h^3\mu_3'(u) = \frac{h^3\sum f_iu_i^3}{n} = 5^3 \times \frac{(-129)}{50} = -322.5$$

$$\mu_4'(x) = h^4\mu_4'(u) = \frac{h^4\sum f_iu_i^4}{n} = 5^4 \times \frac{(473)}{50} = 5912.5$$

Hence the corrected moments are

$$\mu_2 = \mu_2' - \mu_1'^2 = 50.5 - (-2.7)^2 = 43.21$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$= -322.5 - 3 \times 50.5 \times (-2.7) + 2(-2.7)^3 = 47.18$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 5912.5 - 4 \times (-322.5) \times (-2.7)$$

$$+ 6 \times 50.5 \times (-2.7)^2 - 3 \times (-2.7)^4$$

$$= 4478.94$$

Based on the above measures .

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{47.18^2}{43.21^3} = 0.03 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{4478.94}{43.21^2} = 2.40$$

Clearly, the distribution is slightly skewed to the right as implied by the β_1 value. It is platykurtic since $\beta_2 < 3$.

Our previous calculations show that for this distribution, Q_1=34.67, Q_2=38.83, Q_3=43.87, mean=39.3, s=6.6, so that the Pearson's and Bowley's coefficient of skewness are respectively.

$$S_k(P) = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviatiobn}} = \frac{3(39.3 - 38.83)}{6.6} = 0.21$$

$$S_k(B) = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{5.04 - 4.16}{9.2} = 0.10$$

The empirical formula due to Pearson and Bowley also demonstrate the same feature of the shape characteristic of the distribution.

Example 4.38: The second moment about the mean of a symmetrical distribution is 25. What must be its fourth moment about the mean for the distribution to be (i) Leptokurtic (ii) Platykurtic and (iii) Mesokurtic?

Solution: A distribution is leptokurtic, platykurtic or mesokurtic according as the value of β_2 is greater than, less than or equal to 3, where

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

The distribution will be

(i) Leptokurtic if $\mu_4 > 3\mu_2^2$, i.e. if $\mu_4 > 3(25)^2$ i.e. if $\mu_4 > 1875$.

(ii) Platykurtic if $\mu_4 < 3\mu_2^2$, i.e. if $\mu_4 < 3(25)^2$ i.e. if $\mu_4 < 1875$.

(iii) Mesokurtic if $\mu_4 = 3\mu_2^2$, i.e. if $\mu_4 = 3(25)^2$ i.e. if $\mu_4 = 1875$.

4.13 BOX AND WHISKER PLOT

4.13.1 Five-Number Summary

The box and whisker plot is an elegant technique of studying the shape characteristics of a distribution. To accomplish this task, the plot makes use of a set of measures collectively called **five-number summary**.

The summary consists of

a) the lowest measurement (L)
b) the first quartile (Q_1) measurement
c) the median or the second quartile (Q_2) measurement
d) the third quartile (Q_3) measurement, and
e) the highest measurement (H)

It is easy to display a five-number summary graphically. We illustrate by an example below how such a summary can be displayed.

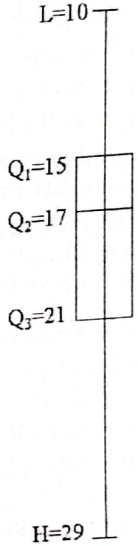
Example 4.39: The following data are borrowed from example 2.17 in chapter 2 which refer to the payment times in days of 65 telephone customers.

22	29	16	15	18	17	12	13	17	16	15	19	17
10	21	15	14	17	18	12	20	14	16	15	16	20
22	14	25	19	23	15	19	18	23	22	16	16	19
13	18	24	24	26	13	18	17	15	24	15	17	14
18	17	21	16	21	25	19	20	27	16	17	16	21

Display the data by five-number summary.

Solution: To understand the nature of the given distribution, we begin with constructing a stem-and-leaf plot as displayed in the accompanying range. We note that the highest observation (H) and lowest observation (L) in the data set are respectively 29 and 10 resulting in a range of 19. The other measures are: Q_1=15, Q_2=17 and Q_3=21. With these values, we draw a vertical line extending from L=10 to H=29. In addition, a rectangle is drawn that extends from Q_1=15 to Q_3=21 and a horizontal line is drawn to indicate the location of the median (Q_2).

Stem	Leaf
10	0
11	
12	00
13	000
14	0000
15	0000000
16	000000000
17	00000000
18	000000
19	00000
20	000
21	0000
22	000
23	00
24	000
25	00
26	0
27	0
28	
29	0



The summary divides the measurements into four parts, with the middle 50 percent of the measurements depicted by the rectangle. The summary indicates that the largest 25 percent of the measurements is more spread out than the smallest 25 percent of the measurements, and that the second-largest 25 percent of the measurements is more spread out than the second smallest 25 percent of the measurements. Overall, the summary indicates that the measurements are somewhat skewed to the right. The stem-and-leaf plot displays that the distribution is indeed positively skewed

4.13.2 Box–and–Whisker Plot

A more sophisticated modification of the graphical five-number summary is the **box-and whiskers** plot (also known as **box plot**). A box plot

reduces the details of the stem and leaf plot and provides a different visual image of the distribution's location, spread, shape, tail, length and outliers. Such a plot is constructed by using Q_1, Q_2, Q_3 and the inter-quartile range (IQR). The plot also demonstrates the concentration of the values in the tails of the distribution. Box and whiskers plot also conveys to us an impression of the location or centering through a special measure of central tendency known as **trimean** and is defined as

$$\text{Trimean} = \frac{\text{Lower hinge} + 2(\text{median}) + \text{upper hinge}}{4}$$

To construct a box-and -whiskers display, the following steps are involved:

1. Draw a box that extends from Q_1 to Q_3 representing the inter-quartile range and so encloses the middle 50% of the measurements. The edges of the box are known as the **hinges.**
2. Next draw a vertical line through the box at the value of the median (Q_2). This line divides the data set into two roughly equal parts.
3. We next define what we call **inner fence.** There are two inner fences. The first inner fence is located 1.5(IQR) below Q_1 and the second inner fence is located 1.5(IQR) above Q_3. That is the inner fences are

$$Q_1 - 1.5\,(\text{IQR}) \text{ and } Q_3 + 1.5\,(\text{IQR})$$

4. Next we define what we call **outer fence**. Like inner fences, there are two outer fences. Of these two outer fences, one is located 3(IQR) below Q_1 and the other is located 3(IQR) above Q_3. That is the outer fences are

$$Q_1 - 3\,(\text{IQR}) \text{ and } Q_3 + 3(\text{IQR})$$

The inner and outer fences are used to identify **outliers**. An outlier is a measurement that is unusually different from most of the other measurements in the data set.

5. Draw two **whiskers** which are the dashed lines extending below Q_1 and above Q_3. One whisker is drawn from Q_1 to the smallest measurement between the inner fences. The other whisker is drawn from Q_3 to the largest measurement between the inner fences.
6. Now we turn to identify the outliers (if any) in the data set. Measurements that are located between the inner and the outer fences are known as **mild outliers.** Plot these measurements by the symbol *. Measurements that are located outside the outer fences are known as **extreme outliers.** Plot these measurements by the symbol Θ.

Example 4.40: The following are the number of miles traveled by 51 people by car in a given week

52	46	50	78	72	62	93	78	41	47	60
66	56	63	66	81	58	77	78	85	67	
94	70	68	43	42	82	48	44	44	93	
57	58	85	52	72	87	54	48	47	80	
76	74	52	67	86	63	74	53	86	70	

Display the data in a box and whisker plot and comment..

Solution: The plot is constructed by drawing a box between the lower and upper quartiles, i.e. Q_1 and Q_3 with a solid line drawn across the box to locate the median (Q_2). Here Q_1=52, Q_2=66, Q_3=78 so that the inner fences (I_f) are

$$52 - 1.5(78 - 52) = 13 \text{ and } 78 + 1.5(78 - 52) = 117$$

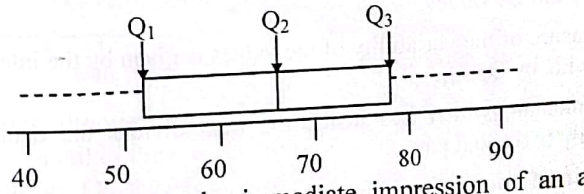
The outer fences (O_f) are

$$52 - 3(78 - 52) = -26 \text{ and } 78 + 3\,(78 - 52) = 156$$

We next draw the whiskers (W_h). For the data in hand, one whisker extends from Q_1=52 down to 41, because 41 is the smallest measurement between the first inner fence I_f =13 and the second inner fence I_f=117

The other whisker is drawn from Q_3 to the largest measurement between the inner fences. This whisker extends from Q_3=78 up to 91, because 91 is the largest measurement between I_f=13 and I_f=117. Both the whiskers are shown by dashed lines.

Is there any outlier mild or extreme? Mild outliers are located between inner and outer fences. Examine that the (I_f, O_f) pairs are (–26, 13) and ((117, 156). No values lie between these pairs. The data thus do not have any mild outliers. No value (s) lie outside the outer fences, hence there is no extreme outliers too. The complete box plot appears below.

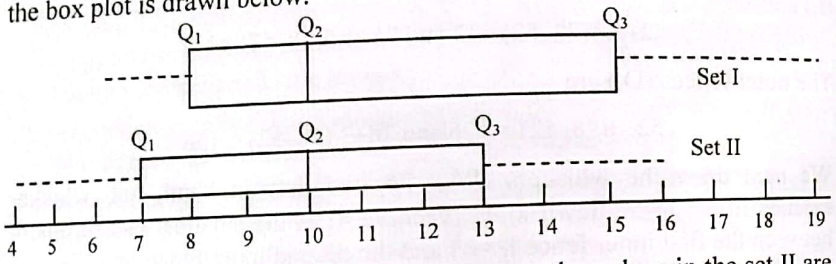


The box plot gives us the immediate impression of an approximately symmetrical distribution with the middle 50% of the values lying between 52 and 78 resulting in an inter-quartile range of 26.

Example 4.41: Given the following information obtained from two sets of data. Draw a box plot to represent these data and comment on the distributions.

Set I: Median=10, Lower quartile = 8, Upper quartile = 15, Lowest value = 6, Highest value = 19.
Set II: Median=10, Lower quartile = 7, Upper quartile = 13, Lowest value = 4, Highest value = 16.

Solution: Check that for the set 1, the inner fences are –2.5 and 25.5, while the outer fences are –13 and 36. For the second set, the inner fences are –2 and 22, while the outer fences are –11 and 31. With these values, the box plot is drawn below.



The median for both sets is the same. However, the values in the set II are more evenly distributed with a smaller range. There is a bigger spread of values for set I and the distribution for this set is positively skewed.

4.13.3 Interpreting a Box and Whiskers Plot

While interpreting a box plot, we require certain points to be kept in mind.

These are

1. The box between Q_1 and Q_3 contains the middle 50% of the data.
2. The edges of the box are called hinges, which are approximated by the Q_1 and Q_2 values.
3. A measure of the variability of the values is given by the inter-quartile range i.e. by Q_3–Q_1.
4. The median, which lies inside the box, divides the data set into roughly two equal parts.
5. Additional information about skewness is obtained from the lengths of the whiskers. If one of the whiskers is longer than the other, the data set is probably skewed in the direction of the longer whisker.

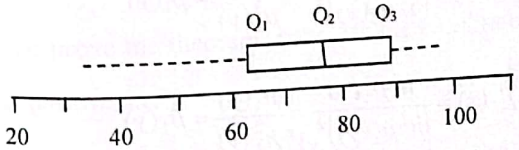
6. A general assessment can be made about the presence of outliers by examining the number of values classified as mild outliers and the number classified as extreme outliers.
7. Observations identified as outliers should be investigated (for its accuracy in measurement, reporting or plotting).

Generally, a box plot displays the central tendency, variability, and overall range of a set of data. It also displays whether the measurements are symmetrically distributed. However, the exact shape of the distribution is better understood by a stem-and-leaf plot, and/ or a histogram.

Consider the following stem-and-leaf plot and the corresponding box and whisker plots dra' n alongside.

Stem	Leaf
3	2
3*	
4	
4*	5
5	0
5*	6 8
6	0 1 1 3 4 4
6*	5 6 7 7 8 9 9
7	2
7	6 8
8	1 3 3
8*	5 6 7 7 8 9 9
9	0 0 1 2 2 3 3 4
9*	6 8

The corresponding box plot with Q_1=64, Q_2=77 and Q_3=89.5 is



We note that, although the box plot tells us that the scores are somewhat skewed with a tail to the left, it does not reveal the double-peaked nature of the distribution. On the other hand, the stem-and-leaf plot clearly shows that this distribution is double-peak.

4.14 MORE THEOREMS AND EXAMPLES

***Theorem* 4.13**: *Prove that β_1 and β_2 are invariant to the changes in origin and scale of measurement.*

Proof: Let $\beta_1(x)$ and $\beta_2(x)$ denote the values of β_1 and β_2 calculated from a set of observations $x_1, x_2, \ldots x_n$ pertaining to a variable x. Then by definition

$$\beta_1(x)=\frac{\mu_3^{\,2}(x)}{\mu_2^{\,3}(x)} \quad \text{and} \quad \beta_2(x)=\frac{\mu_4(x)}{\mu_2^{\,2}(x)}$$

Let y be a transformed variable assuming values $y_1, y_2, \ldots y_n$. Under this transformation, let the ith observation y_i be defined as follows:

$$y_i=\frac{x_i-a}{h}, \qquad h>0$$

where a and h are respectively the origin and scale factors. The corresponding beta values for the variable x are:

$$\beta_1(y)=\frac{\mu_3^{\,2}(y)}{\mu_2^{\,3}(y)} \quad \text{and} \quad \beta_2(y)=\frac{\mu_4(y)}{\mu_2^{\,2}(y)}$$

We will have to prove that $\beta_1(x)=\beta_1(y)$ and $\beta_2(x)=\beta_2(y)$

We know from section (4.13) that $\mu_r(x) = h^r \mu_r(y)$, from which

$$\mu_2(x)=h^2\mu_2(y),\ \mu_3(x)=h^3\mu_3(y), \text{ and } \mu_4(x)=h^4\mu_4(y),$$

Hence

$$\beta_1(x)=\frac{\left(h^3\mu_3(y)\right)^2}{\left(h^2\mu_2(y)\right)^3}=\frac{\mu_3^2(y)}{\mu_2^3(y)}=\beta_1(y)$$

and

$$\beta_2(x)=\frac{h^4\mu_4(y)}{\left(h^2\mu_2(y)\right)^2}=\frac{\mu_4(y)}{\mu_2^2(y)}=\beta_2(y)$$

Hence the proof.

***Theorem* 4.14:** *For any set of observations $x_1, x_2, \ldots x_n$, prove that $\beta_2 \geq 1 + \beta_1$.*

Proof: Let us recall that

$$\mu_2=\frac{1}{n}\sum(x_i-\bar{x})^2, \quad \mu_3=\frac{1}{n}\sum(x_i-\bar{x})^3, \quad \mu_4=\frac{1}{n}\sum(x_i-\bar{x})^4$$

Consider the expression $\left(ax_i^2+bx_i+c\right)^2$ where a, b, c are arbitrary constants to be chosen later. If a, b, and c are assumed to be real, then the above expression is always positive. That is

$$\left(ax_i^2+bx_i+c\right)^2\geq 0 \qquad \ldots \text{(a)}$$

Replacing x in (a) by $(x_i-\bar{x})$, summing all along and dividing throughout by n,

$$\frac{a^2\sum(x_i-\bar{x})^4}{n}+\frac{b^2\sum(x_i-\bar{x})^2}{n}+c^2+\frac{2ab\sum(x_i-\bar{x})^3}{n}+\frac{2ac\sum(x_i-\bar{x})^2}{n}+\frac{2bc\sum(x_i-\bar{x})}{n}\geq 0$$

or
$$a^2\mu_4+b^2\mu_2+c^2+2ab\mu_3+2ac\mu_2+0\geq 0$$

Choosing $a=1$, $b=-\mu_3/\mu_2$ and $c=-\mu_2$, the above expression becomes

$$\mu_4-\frac{\mu_3^{\,2}}{\mu_2}-\mu_2^{\,2}\geq 0 \qquad \ldots \text{(b)}$$

Dividing both sides of (b) by $\mu_2^{\,2}$

$$\frac{\mu_4}{\mu_2^{\,2}}-\frac{\mu_3^{\,2}}{\mu_2^{\,3}}-1\geq 0 \Rightarrow \beta_2-\beta_1-1\geq 0$$

Hence
$$\beta_2\geq 1+\beta_1$$

***Theorem* 4.15**: *Prove that $\beta_2 \geq 1$*

Proof: To prove the theorem, consider the expression $\frac{\sum(x_i-\bar{x})^2}{n}$, which is always positive, i.e.

$$\frac{\sum(x_i-\bar{x})^2}{n}\geq 0$$

or
$$\frac{\sum x_i^{\,2}}{n}-\left(\frac{\sum x_i}{n}\right)^2\geq 0 \Rightarrow \frac{\sum x_i^{\,2}}{n}\geq\left(\frac{\sum x_i}{n}\right)^2$$

Replacing x_i by $(x_i-\bar{x})^2$ in the last two terms of the above expression,

$$\frac{\sum(x_i-\bar{x})^4}{n} \geq \left\{\frac{\sum(x_i-\bar{x})^2}{n}\right\}^2 \Rightarrow \mu_4 \geq {\mu_2}^2$$

Dividing both sides by ${\mu_2}^2$, $\beta_2 \geq 1$

The above theorem can also be proved by considering the expression $\{a(x_i-\bar{x})^2+c\}^2$ and proceeding as in Theorem (4.13) above:

Expanding the expression, we have

$$a^2(x_i-\bar{x})^4+2ac(x_i-\bar{x})^2+c^2 \geq 0$$

Summing and dividing throughout by n

$$\frac{a^2\sum(x_i-\bar{x})^4}{n}+2ac\frac{\sum(x_i-\bar{x})^2}{n}+c^2 \geq 0$$

or

$$a^2\mu_4+2ac\mu_2+c^2 \geq 0$$

Setting a=1 and $c=-\mu_2$,

we have $\mu_4-\mu_2^2 \geq 0$. Dividing throughout by μ_2^2

$$\beta_2 \geq 1$$

EXERCISES 4

1. What do you mean by dispersion of a variable? What are its measures? Why do you need these measures at all? Explain.
2. What are the important measures of dispersion? Give a brief description of these measures. What are the criteria by which you can judge the adequacy of a measure of dispersion?
3. How do you arrive at an estimate of the standard deviation from sample data? What are the advantages and disadvantages of the standard deviation over the range and the average deviation respectively?
4. What is variance? How do you compute it from raw data? If you have a grouped distribution, how do you modify your formula?

5. Given below are the monthly household incomes (in Tk.) for ten families.

10,648	17,416	6,517	13,555	14,821
9,226	152,936	11,800	18,527	12,222

Compute the range, inter-quartile range and standard deviation as measures of variability.

Which of the above measures you feel is the best measure of variability in the data? Why?

6. What is coefficient of variation? What are its advantages over the other measures of dispersion? If the average score of male students is 3.0 and the standard deviation of their scores is 0.25, and if the corresponding figures for female students are 2.9 and 0.25, do the scores of male students are in greater variability? Why?
7. Two salesmen selling the same product have the following records of sales over a long period of time:

	Salesman I	Salesman II
Average sales per month	Tk. 30,000	Tk. 35,000
Standard deviation	Tk.2,500	Tk.3,600

Which salesman seems to be more consistent in respect to the volume of sales? Give reasons in support of your answer.

8. What are the properties of variance? If $Z=X+Y$, then under what condition is the variance of Z equal to the sum of the variance of X and variance of Y? Explain.
9. Why are the variance and the standard deviation, computed in terms of the arithmetic mean rather than of any other measures of central tendency? Suppose that the variable x has a known mean 15 and a variance 25. Find the mean and standard deviation of y under each of the following relationships: (a) $y=4+16x$, (b) $y=16-4x$ **[Ans.:**(a) 244, 80 (b) –44, 20]
10. Distinguish between a population variance and a sample variance. Show that the variance is independent of origin but it depends on the scale of measurement. Is the same true for coefficient of variation and standard deviation?
11. Deduce a formula to compute the combined variance of two sets data having n_1 and n_2 observations, with means $\bar{x}_1$ and $\bar{x}_2$ and variances ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively. Write down the formula when $\bar{x}_1=\bar{x}_2$ and that $n_1=n_2$.

The means of two samples of sizes 50 and 100 are 54.1 and 50.3 and the standard deviations are 8 and 7 respectively. Find the combined mean and the standard deviation of these two samples. **[Ans.:** Mean=51.57, sd=7.60]

13. Deduce a formula to compute the combined variance of two sets data having