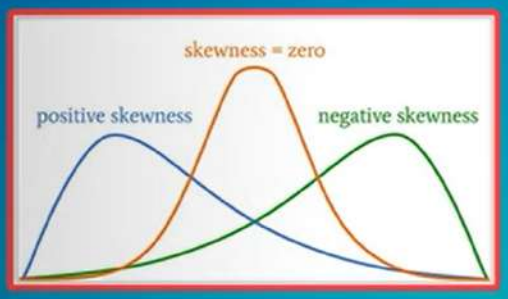
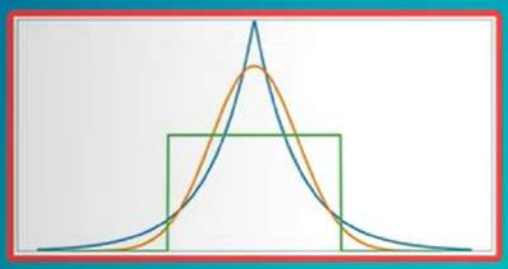
Skewness And Kurtosis

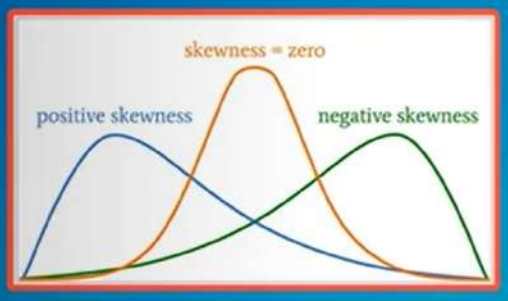


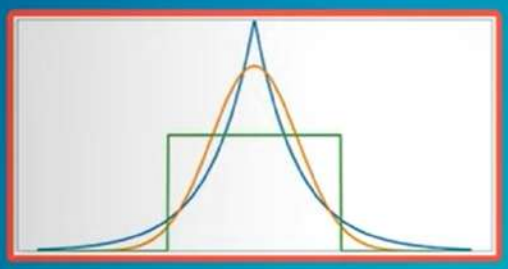


simplilearn

What's in it for you?

- Symmetrical Distribution
- Skewed Distribution
- Pearson's Coefficient Of Skewness
- Kurtosis

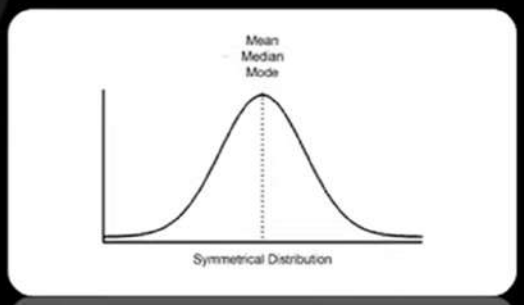




simpli

Symmetrical Distribution

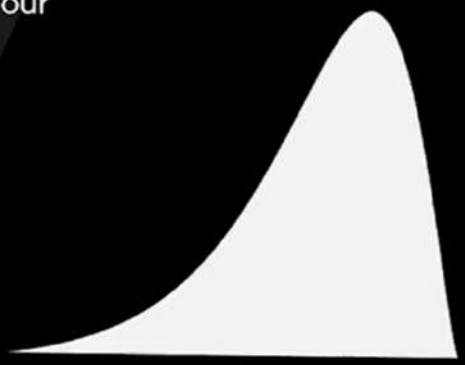
- A frequency distribution is said to be symmetrical if the frequencies are equally distributed on both the sides of central value
- A symmetrical distribution may be either bell – shaped or U-shaped
- In symmetrical distribution, the values of mean, median and mode are equal i.e., Mean=Median=Mode



Skewed Distribution

- Skewness is used to measure the level of asymmetry in our data. It is the measure of asymmetry that occurs when our data deviates from the norm.

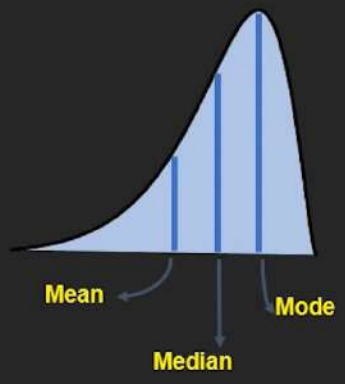
- A skewed distribution may be –
  - Positively Skewed
  - Negatively Skewed



Skewed Distribution

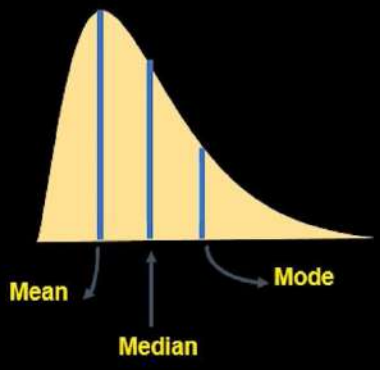
❑ Negatively Skewed

- In this, the distribution is skewed to the left
- Here, Mode exceeds Mean and Median



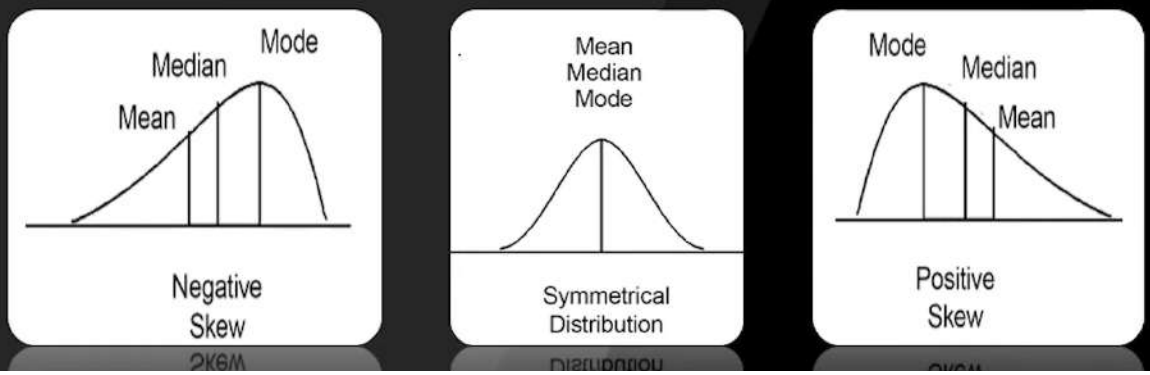
❑ Positively Skewed

- In this, the distribution is skewed to the right
- Here, Mean exceeds Mode and Median



Graphical Measure Of Skewness

- Measures of skewness help us to know to what degree and in which direction (positive or negative) the frequency distribution has a departure from symmetry
- Positive or negative skewness can be detected graphically (as below) depending on whether the right tail or the left tail is longer but, we don't get idea of the magnitude



Pearson's Coefficient Of Skewness

- This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

$$\text{Pearson's Coefficient} = \frac{Mean - Mode}{Standard\ Deviation}$$

- Normally, the coefficient of skewness lies between -3 to +3

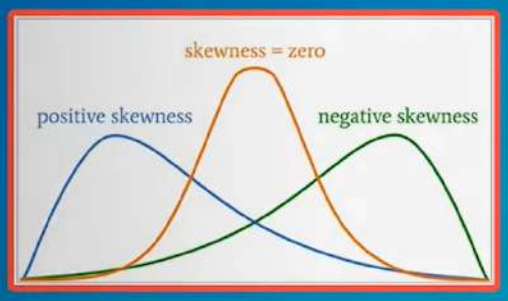
Pearson's Coefficient Of Skewness

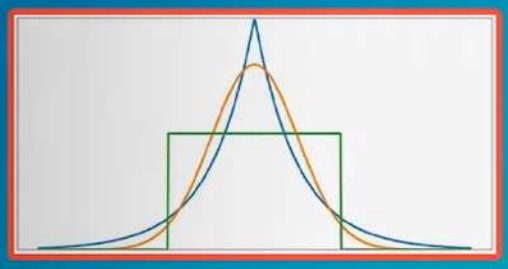
In case the mode is indeterminate

$$\text{Pearson's Coefficient} = \frac{\text{Mean} - (3\,\text{Median} - 2\,\text{Mean})}{\text{Standard Deviation}}$$

- The value of coefficient of skewness is zero, when the distribution is symmetrical
- The value of coefficient of skewness is positive, when the distribution is positively skewed
- The value of coefficient of skewness is negative, when the distribution negatively skewed

Kurtosis



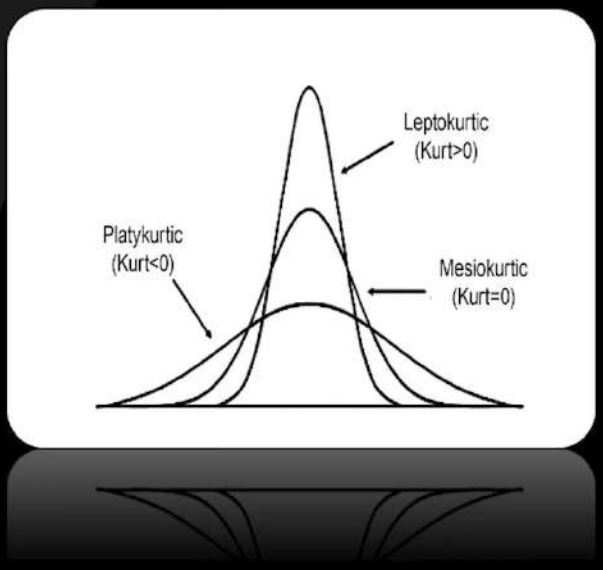


Kurtosis

- Kurtosis is another measure of the shape of a frequency curve. It is a Greek word, which means bulginess
- While skewness signifies the extent of asymmetry, kurtosis measures the degree of peakness of a frequency distribution

Kurtosis

- When the peak of a curve becomes relatively high then that curve is called Leptokurtic
- When the curve is flat-topped, then it is called Platykurtic
- The normal curve is called Mesokurtic



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.

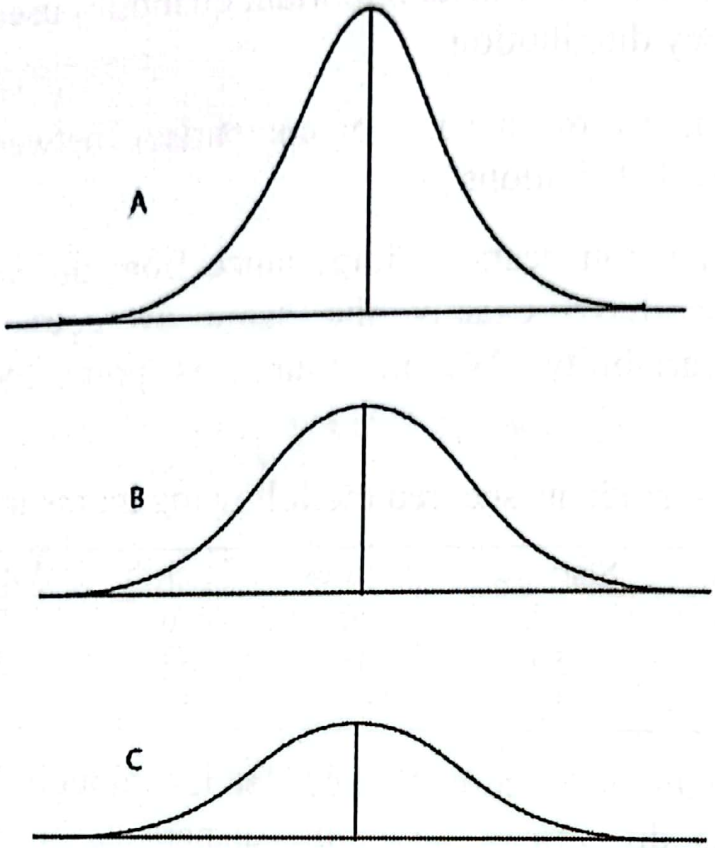


Figure 4.1: Distributions with unequal variability but with identical mean

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 $x_1, x_2,.....,x_n$ are the values of n observations, then range R of the variable x is given by

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

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

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

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

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).

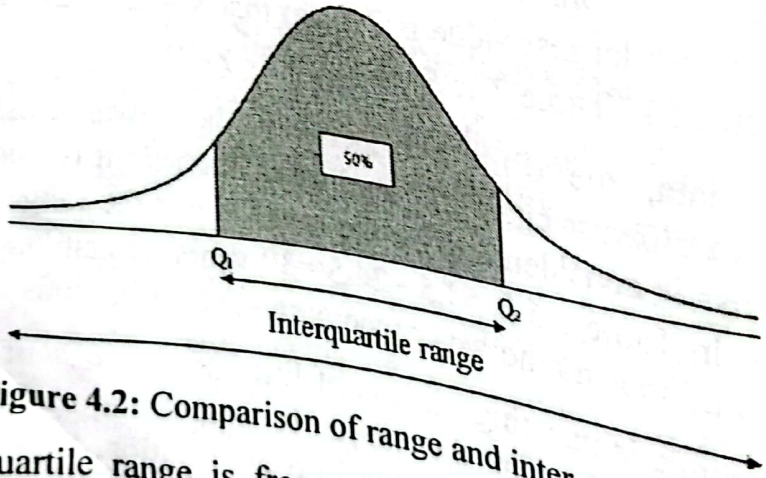


Figure 4.2: Comparison of range and inter-quartile range

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.

4.3.4 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 all 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.

4.3.5 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

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

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

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

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

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

and

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

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

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

Solution: To compute the mean deviation about the mean, follow the steps below:

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.

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

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:

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

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

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:

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

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

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

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:

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

$$S_2=10+12=22$$

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 deviati about the median value of the distribution using (4.7).

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

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 4 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. using usual method.

4.3.7 Variance and Standard Deviation

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

the results are added[1]. The sum of squares can be regarded as a measure of the 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

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

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

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

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

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

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

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, $E(s^2)=\sigma^2$.

4.3.8 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, ,,.,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 $(x_1-\bar{x})$, $(x_2-\bar{x})$,...,$(x_n-\bar{x})$. 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, ..,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:

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

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

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

Example 4.6: Compute the variance and standard deviation for the following frequency distribution:

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

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

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

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

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

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

Solution: In order to compute the required measures, we construct the following table:

Length	x_i	x_i^2	f_i	$f_i x_i$	$f_i x_i^2$
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

4.3.10 Properties of Variance

(i) Effect of changes in origin on variance

Variance and hence standard deviation have certain appealing properties. Consider that each of the numbers $x_1, x_2,....x_n$ increases or decreases by a constant amount c. What is the effect of this change on the value of the variance? Specifically, if the variance of $x_1, x_2,....x_n$ is denoted by s_x^2 and that of the new numbers by s_y^2, then how are these two quantities related: is $s_x^2=s_y^2$, or $s_x^2>s_y^2$ or $s_x^2<s_y^2$? We show below that such a change in the value of x's does not have any effect on their variance. In other words, $s_x^2= s_y^2$. We demonstrate below this invariance property of the variance

Let y be the transformed variable defined as follows:

$$y_i = x_i \pm c \qquad (i=1,2,...,n)$$

where c is a constant.

Summing the above expression and dividing throughout by n–1, $\bar{y} = \bar{x} \pm c$,

(a) When $y_i = x_i + c$, we have $\bar{y} = \bar{x} + c$, so that

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

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

(b) When $y_i = x_i - c$, than $\bar{y} = \bar{x} - c$, so that

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

This establishes the fact that any linear change in the variable x does not have any effect on its variance.

(ii) Effect of changes in the scale of measurement on variance

What happens when each observation of the variable is multiplied or divided by a constant amount c? Let us examine the case when each observation is divided by c. Under this transformation, the new set of observations is

$$y_i = \frac{x_i}{c} \qquad (i=1, 2,....n)$$

Summing both sides and dividing throughout by n, we have $\bar{y} = \frac{\bar{x}}{c}$. Us 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 chang in origin and the scale of measurement on the variance.

***Theorem* 4.1:** *The variance is independent of origin but dependent on th 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 ar 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 $$s_x^2 = h^2 s_y^2 \text{ (Proved)}$$

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 bu 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.

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).

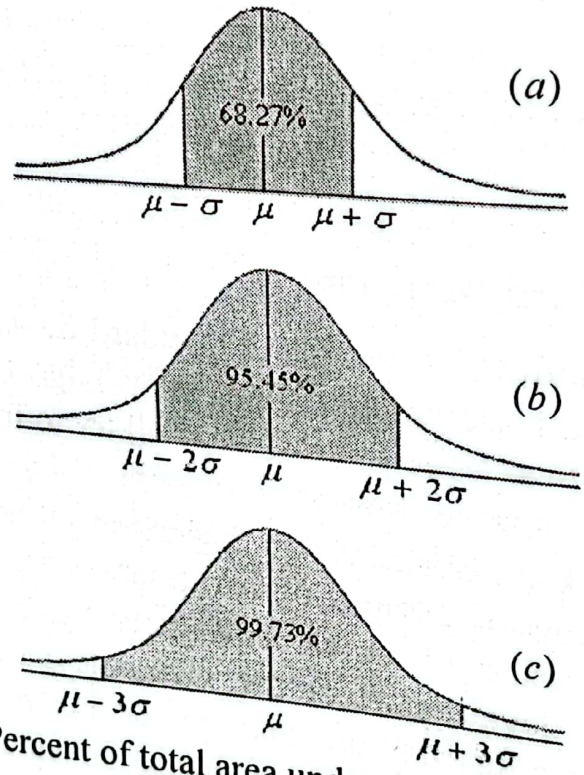


Figure 4.3(a-c): Percent of total area under the curve between various parts

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.

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:

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

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:

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

$$\text{Quartile deviation} = \frac{2}{3}(\text{standard deviation}) \qquad \ldots (4.24)$$

The implication of the above two relations is that

$$\text{Mean deviation} = \frac{6}{5}(\text{Quartile deviation}) \qquad \ldots (4.25)$$

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

$$\ldots (4.26)$$

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?

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, h

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

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

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

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 be any arbitrarily chosen value. Then the first four raw moments about 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 referre 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:

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

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

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:

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

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:

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

For a frequency distribution, the r-th moment defined above will be expressed 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: $\nu_r = \frac{1}{n}\sum f_i x_i^r$,

where $n = \sum f_i$

Example 4.23: Compute the first four central moments for the following frequency distribution

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

Solution: 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 RA

4.9.1 Central Moment and Moment about Arbitrary Value

The computation of central moments can be effected through th computation of the raw moments about any arbitrary origin. This implie that a relationship does exist between central moments and raw moment 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 $\bar{x}$. L '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

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

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

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 procedure 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

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

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

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

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

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

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

$$= 28.38 - 4(.06)(.56) + 6(.06)(2.64) - 3(.06)^4 = 29.19$$

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

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

This example also demonstrates the same feature: the fourth moment seriously underestimated by Sheppard's adjustment and the third mome 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 4th 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.

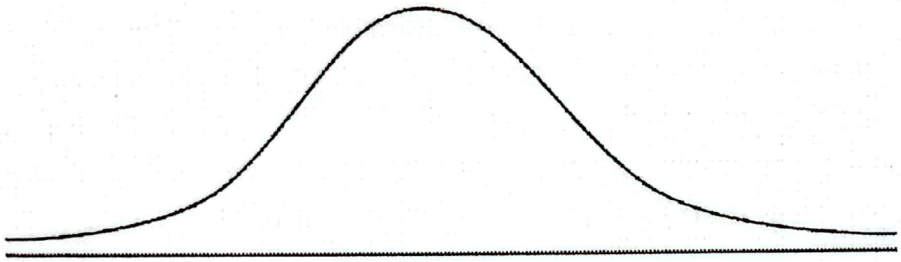


Figure 4.4: Normal curve: symmetrical distribution

(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:

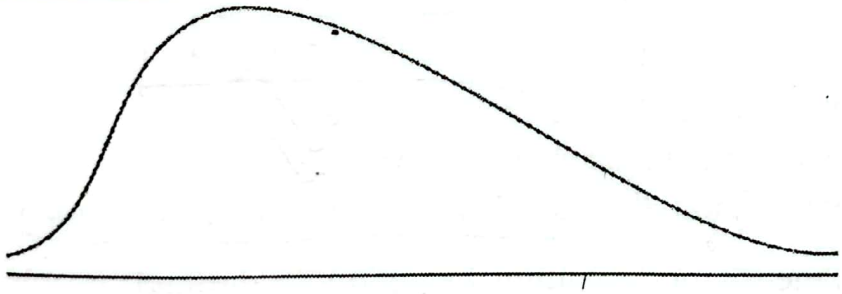


Figure 4.5: Curve representing a positively skewed distribution

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:

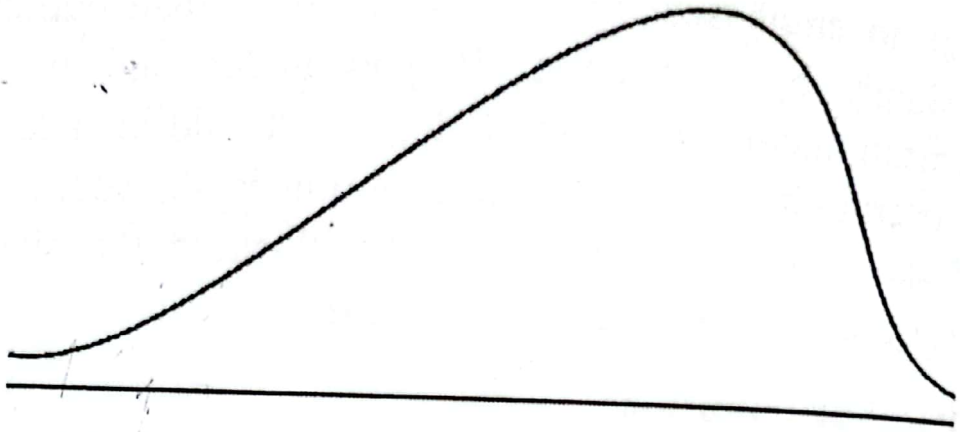


Figure 4.6: Curve representing a negatively skewed distribution

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

of 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:

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

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

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

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:

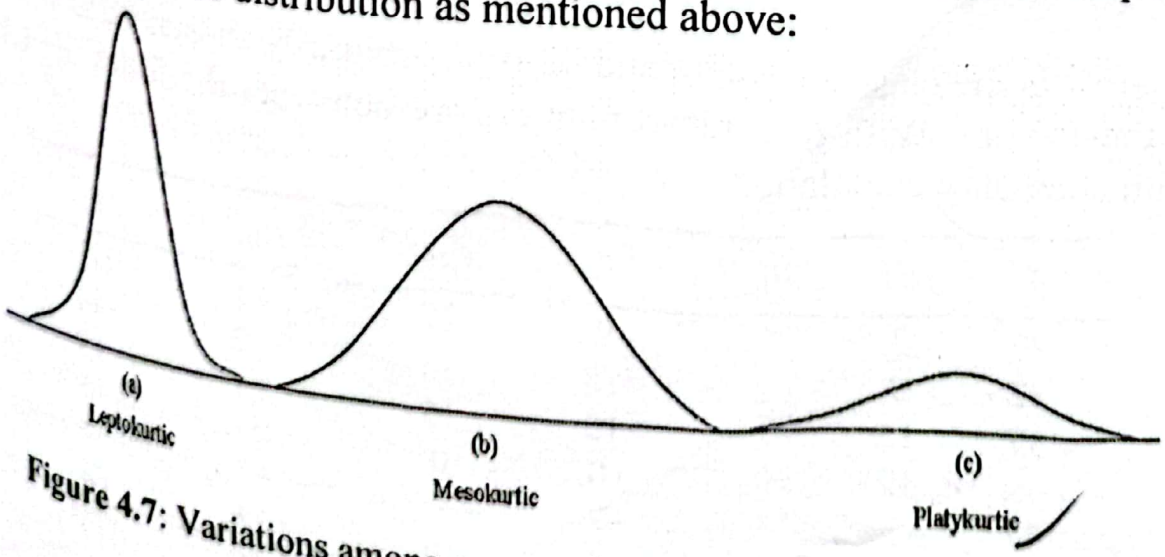


Figure 4.7: Variations among symmetrical or bell-shaped distributions:

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 are

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

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

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

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

ence 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?