$$\sigma_{12} = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$$

where σ_{12} = combined standard deviation

$d_1 = \bar{x}_{12} - \bar{x}_1$; $d_2 = \bar{x}_{12} - \bar{x}_2$

and $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ (combined arithmetic mean)

This formula for combined standard deviation of two sets of data can be extended to compute the standard deviation of more than two sets of data on the same lines.

2. **Standard deviation of natural numbers:** The standard deviation of the first n natural numbers is given by

$$\sigma = \sqrt{\frac{1}{12}(n^2 - 1)}$$

For example, the standard deviation of the first 100 (i.e., from 1 to 100) natural numbers will be

$$\sigma = \sqrt{\frac{1}{12}(100^2 - 1)} = \sqrt{\frac{1}{12}(9999)} = \sqrt{833.25} = 28.86$$

3. **Standard deviation is independent of change of origin but not of scale.**

Example 4.12: For a group of 50 male workers, the mean and standard deviation of their monthly wages are Rs 6300 and Rs 900 respectively. For a group of 40 female workers, these are Rs 5400 and Rs 600 respectively. Find the standard deviation of monthly wages for the combined group of workers. *[Delhi Univ., MBA, 2002]*

Solution: Given that $n_1 = 50,\ \bar{x}_1 = 6300,\ \sigma_1 = 900$

$n_2 = 40,\ \bar{x}_2 = 5400,\ \sigma_2 = 600$

Then, Combined mean, $\bar{x}_{12} = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \dfrac{50 \times 6300 + 40 \times 5400}{50 + 40} = 5{,}900$

and Combined standard deviation

$$\sigma_{12} = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)}{n_1 + n_2}}$$

$$= \sqrt{\frac{50(8{,}10{,}000 + 1{,}60{,}000) + 40(3{,}60{,}000 + 2{,}50{,}000)}{50 + 40}} = \text{Rs } 900$$

where $d_1 = \bar{x}_{12} - \bar{x}_1 = 5900 - 6300 = -400$

$d_2 = \bar{x}_{12} - \bar{x}_2 = 5900 - 5400 = 500$

Example 4.13: A study of the age of 100 persons grouped into intervals 20–22, 22–24, 24–26,... revealed the mean age and standard deviation to be 32.02 and 13.18 respectively. While checking, it was discovered that the observation 57 was misread as 27. Calculate the correct mean age and standard deviation. *[Delhi Univ., MBA 1997]*

Solution: From the data given in the problem, we have $\bar{x} = 32.02$, $\sigma = 13.18$ and $N = 100$. We know that

$$\bar{x} = \frac{\Sigma fx}{N} \text{ or } \Sigma fx = N \times \bar{x} = 100 \times 32.02 = 3202$$

and $$\sigma^2 = \frac{\Sigma fx^2}{N} - (\bar{x})^2 \text{ or } \Sigma fx^2 = N[\sigma^2 + (\bar{x})^2] = 100[(13.18)^2 + (32.02)^2]$$

$$= 100[173.71 + 1025.28] = 100 \times 1198.99$$

$$= 1{,}19{,}899$$

On re-placing the correct observation, we get

$$\Sigma fx = 3202 - 27 + 57 = 3232.$$

Also $\Sigma fx^2 = 1{,}19{,}899 - (27)^2 + (57)^2 = 1{,}19{,}899 - 729 + 3248 = 1{,}22{,}419$

Thus Correct A.M. is $\bar{x} = \dfrac{\Sigma fx}{N} = \dfrac{3232}{100} = 32.32.$

and Correct variance is $\sigma^2 = \dfrac{\Sigma fx^2}{N} - (\bar{x})^2 = \dfrac{1{,}22{,}419}{100} - (32.32)^2$

$$= 1224.19 - 1044.58 = 179.61$$

or Correct standard deviation is, $\sigma = \sqrt{\sigma^2} = \sqrt{179.61} = 13.402.$

Example 4.14: The mean of 5 observations is 15 and the variance is 9. If two more observations having values – 3 and 10 are combined with these 5 observations, what will be the new mean and variance of 7 observations.

Solution: From the data of the problem, we have $\bar{x} = 15$, $s^2 = 9$ and $n = 5$. We know that

$$\bar{x} = \frac{\Sigma x}{n} \quad \text{or} \quad \Sigma x = n \times \bar{x} = 5 \times 15 = 75$$

If two more observations having values – 3 and 10 are added to the existing 5 observations, then after adding these 6th and 7th observations, we get

$$\Sigma x = 75 - 3 + 10 = 82$$

Thus the new A.M. is, $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{82}{7} = 11.71$

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

$$9 = \frac{\Sigma x^2}{n} - (15)^2 \text{ or } \Sigma x^2 = 1170$$

On adding two more observations, i.e., – 3 and 10, we get

$$\Sigma x^2 = 1170 + (-3)^2 + (10)^2 = 1279$$

$$\text{Variance, } s^2 = \frac{\Sigma x^2}{n} - (\bar{x})^2 = \frac{1279}{7} - (11.71)^2 = 45.59$$

Hence the new mean and variance of 7 observations is 11.71 and 45.59 respectively.

4.5.4 Chebyshev's Theorem

Chebyshev's theorem: A statement about the proportion of observations that must lie within σ, 2σ, and 3σ deviations from the mean (population or sample distribution).

Standard deviation measures the variation among observations in a set of data. If the standard deviation value is small, then values in the data set cluster close to the mean. Conversely, a large standard deviation value indicates that the values are scattered more widely around the mean. The Russian mathematician P. L. Chebyshev (1821–1894) developed a result called **Chebyshev's theorem** that allows us to determine the proportion of data values that fall within a specified number of standard deviation from the mean value. The theorem states that:

For any set of data (population or sample) and any constant z greater than 1 (but need not be an integer), the proportion of the values that lie within z standard deviations on either side of the mean is at least $\{1 - (1/z^2)\}$. That is

$$\text{RF}\,[\,|x - \mu| \le z\,\sigma] \ge 1 - \frac{1}{z^2}$$

where RF = relative frequency of a distribution.

$z = \dfrac{x - \mu}{\sigma}$ ← population standardized score for an observation x from the population, that is, number of standard deviations a value, x is away from the mean μ (sample or population)

$= \dfrac{x - \bar{x}}{s}$ ← sample standard score

Chebyshev's theorem states *at least* what percentage of values will fall within z standard deviations in any distribution. However, for a symmetrical, bell-shaped distribution as shown in Fig. 4.4, theorem states *approximately* what percentage of values will fall within z standard deviation.

The relationships involving the mean, standard deviation and the set of observations are called the *empirical rule,* or *normal rule.*

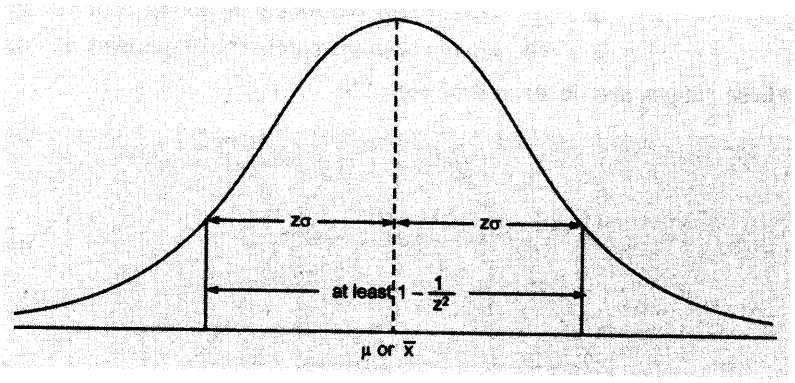


Figure 4.4
Chebyshev Theorem

Some of the implications of the statement of the theorem with $z = 2, 3$, and 4 standard deviations are as follows:

(i) The proportion of all x-values in any set of data to fall within the range $\mu \pm 2\sigma$ is at least $1 - \frac{1}{2^2} = \frac{3}{4} = 0.75$ or 75 per cent.

That is, at least three of four values or 75 per cent values must lie within ±2 standard deviations from the mean.

(ii) The proportion of all x-values in any set of data must lie within the range $\mu \pm 3\sigma$ is at least $1 - \frac{1}{3^2} = \frac{8}{9} = 88.9$ percent.

That is, at least eight of nine values or 88.9 per cent values must lie within ±3 standard deviations from the mean.

(iii) The proportion of all x-values in any set of data must lie within the range $\mu \pm 4\sigma$ is at least $1 - \frac{1}{4^2} = \frac{15}{16} = 93.75$ per cent.

This theorem has its own limitation as it emphasizes on the word, 'at least'. For example for $z = 1$, we have, $1 - \frac{1}{1^2} = 0$, which means that the proportion of all x-values to fall within the range $\mu \pm \sigma$ is zero. This result does not give any information.

The theorem is applicable to any data set regardless of the shape of the frequency distribution of values. For example, assume that the marks obtained by 100 students in business statistics had a mean of 70 per cent and standard deviation of 10 per cent. Then number of students who obtained marks between 50 and 85 will be determined as follows:

(a) For 50 per cent marks, $z = (50 - 70)/10 = -2$ indicates that 50 is 2 standard deviations below the mean,

(b) For 85 per cent marks, $z = (85 - 70)/10 = 1.5$ indicates that 85 is 1.5 standard deviations above the mean.

Now applying the Chebyshev's theorem with $z = 2.0$, we have

$$\left(1 - \frac{1}{z^2}\right) = \left[1 - \frac{1}{(2.0)^2}\right] = 0.75$$

This indicates that at least 75 per cent of the students must have obtained marks between 50 and 85.

Empirical Rule

For symmetrical, bell-shaped frequency distribution (also called normal curve), the range within which a given percentage of values of the distribution are likely to fall within a specified number of standard deviations of the mean is determined as follows:

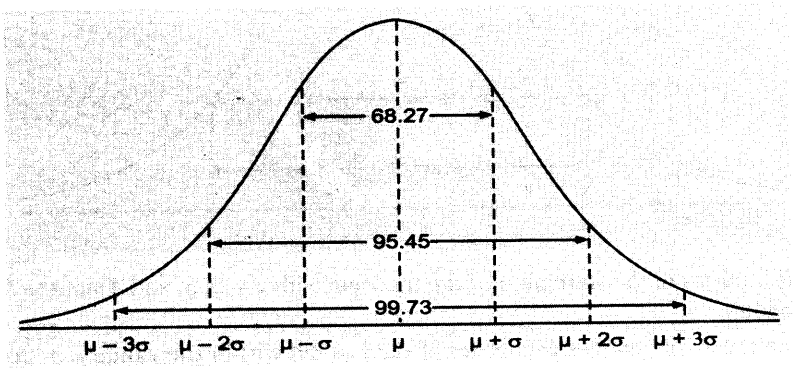
$\mu \pm \sigma$ covers approximately 68.27 per cent of values in the data set

$\mu \pm 2\sigma$ covers approximately 95.45 per cent of values in the data set

$\mu \pm 3\sigma$ covers approximately 99.73 per cent of values in the data set

These ranges are illustrated in Fig. 4.5.

Figure 4.5
Area under Normal Curve



For a symmetrical and bell-shaped distribution, relationships among three measures of variation are given in Table 4.10.

Table 4.10 Relationship Among Measures of Variation

Measures of Variation	*Percentage of Values Scatter Around the Mean Value, μ*			*Size of Measure of Variation to Standard* Deviation at
	±σ	±2σ	±2σ	
Q.D.	50.00	82.30	95.70	0.6748
MAD	57.50	88.90	98.30	0.7979
S.D.	68.27	95.45	99.73	1.0000

Relationship between Different Measures of Variation

(a) Quartile deviation (Q.D.) $= \frac{2}{3}\sigma$

Mean absolute deviation (MAD) $= \frac{4}{5}\sigma$

(b) Quartile deviation $= \frac{5}{6}$ MAD

Standard deviation $= \frac{5}{4}$ MAD or $\frac{3}{2}$ Q.D.

(c) Mean absolute deviation $= \frac{6}{5}$ Q.D.

These relationships are applicable only to symmetrical distributions.

Example 4.15: Suppose you are in charge of rationing in a state affected by food shortage. The following reports arrive from a local investigator:

Daily caloric value of food available per adult during current period:

Area	*Mean*	*Standard Deviation*
A	2500	400
B	2000	200

The estimated requirement of an adult is taken as 2800 calories daily and the absolute minimum is 1350. Comment on the reported figures and determine which area in your opinion, need more urgent attention.

Solution: Taking into consideration the entire population of the two areas, we have

Area A:
$$\mu + 3\sigma = 2500 + 3 \times 400 = 3700 \text{ calories}$$
$$\mu - 3\sigma = 2500 - 3 \times 400 = 1300 \text{ calories}$$

This shows that there are adults who are taking even less amount of calories, that is, 1300 calories as compared to the absolute minimum requirement of 1350 calories.

Area B:
$$\mu + 3\sigma = 2000 + 3 \times 200 = 2600 \text{ calories}$$
$$\mu - 3\sigma = 2000 - 3 \times 200 = 1400 \text{ calories}$$

These figures are statisfying the requirement of daily calorific need. Hence, area A needs more urgent attention.

Example 4.16: The following data give the number of passengers travelling by airplane from one city to another in one week.

115 122 129 113 119 124 132 120 110 116

Calculate the mean and standard deviation and determine the percentage of class that lie between (i) $\mu \pm \sigma$, (ii) $\mu \pm 2\sigma$, and (iii) $\mu \pm 3\sigma$. What percentage of cases lie outside these limits?

Solution: The calculations for mean and standard deviation are shown in Table 4.11.

Table 4.11 Calculations of Mean and Standard Deviation

x	$x - \bar{x}$	$(x - \bar{x})^2$
115	−5	25
122	2	4
129	9	81
113	−7	49
119	−1	1
124	4	16
132	12	144
120	0	0
110	−10	100
116	−4	16
1200	0	436

$$\mu = \frac{\Sigma x}{N} = \frac{1200}{10} = 120 \quad \text{and} \quad \sigma^2 = \frac{\Sigma (x - \bar{x})^2}{N} = \frac{436}{10} = 43.6$$

Therefore $\sigma = \sqrt{\sigma^2} = \sqrt{43.6} = 6.60$

The percentage of cases that lie between a given limit are as follows:

Interval	*Values within Interval*	*Percentage of Population*	*Percentage Falling Outside*
$\mu \pm \sigma = 120 \pm 6.60$	113, 115, 116, 119	70%	30%
$= 113.4$ and 126.6	120, 122, 124		
$\mu \pm 2\sigma = 120 \pm 2(6.60)$	110, 113, 115, 116, 119	100%	nil
$= 106.80$ and 133.20	120, 122, 124, 129, 132		

Example 4.17: A collar manufacturer is considering the production of a new collar to attract young men. Thue following statistics of neck circumference are available based on measurement of a typical group of the college students:

Mid value (in inches) :	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0
Number of students :	2	16	36	60	76	37	18	3	2

Compute the standard deviation and use the criterion $\bar{x} \pm 3\sigma$, where σ is the standard deviation and $\bar{x}$ is the arithmetic mean to determine the largest and smallest size of the collar he should make in order to meet the needs of practically all the customers bearing in mind that collar are worn on average half inch longer than neck size.

Solution: Calculations for mean and standard deviation in order to determine the range of collar size to meet the needs of customers are shown in Table 4.12.

Table 4.12 Calculations for Mean and Standard Deviation

Mid-value (in inches)	Number of students	$\frac{x-A}{h}=\frac{x-14}{0.5}$	fd	fd^2
12.0	2	−4	−8	32
12.5	16	−3	−48	144
13.0	36	−2	−72	144
13.5	60	−1	−60	60
(14.0) ← A	76	0	0	0
14.5	37	1	37	37
15.0	18	2	36	72
15.5	3	3	9	27
16.0	2	4	8	32
	N = 250		−98	548

$$\text{Mean}, \bar{x} = A + \frac{\Sigma fd}{N} \times h = 14.0 - \frac{98}{250} \times 0.5 = 14.0 - 0.195 = 13.805$$

$$\text{Standard deviation}, \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h = \sqrt{\frac{548}{250} - \left(\frac{-98}{250}\right)^2} \times 0.5$$

$$= \sqrt{2.192 - 0.153} \times 0.5 = 1.427 \times 0.5 = 0.7135$$

$$\text{Largest and smallest neck size} = \bar{x} \pm 3\sigma = 13.805 \pm 3 \times 0.173$$

$$= 11.666 \text{ and } 15.944.$$

Since all the customers are to wear collar half inch longer than their neck size, 0.5 is to be added to the neck size range given above. The new range then becomes:

(11.666 + 0.5) and (15.944 + 0.5) or 12.165 and 16.444, i.e. 12.2 and 16.4 inches.

Example 4.18: The breaking strength of 80 'test pieces' of a certain alloy is given in the following table, the unit being given to the nearest thousand grams per square inch;

Breaking Strength	*Number of Pieces*
44–46	3
46–48	24
48–50	27
50–52	21
52–54	5

Calculate the average breaking strength of the alloy and the standard deviation. Calculate the percentage of observations lying between $\bar{x} \pm 2\sigma$.

[Vikram Univ., MBA, 2000]

Solution: The calculations for mean and standard deviation are shown in in Table 4.13.

Table 4.13 Calculations for Mean and Standard Deviation

Breaking Strength	*Number of Pieces(f)*	*Mid-value (m)*	$d = (m-A)/h = (m-49/2$	fd	fd^2
44–46	3	45	–2	–6	12
46–48	24	47	–1	–24	24
48–50	27	A → (49)	0	0	0
50–52	21	51	1	21	21
52–54	5	53	2	10	20
	80			1	77

$$\text{Mean}, \bar{x} = A + \frac{\Sigma fd}{N} \times h = 49 + \frac{1}{80} \times 2 = 49.025$$

$$\text{Standard deviation}, \quad \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h = \sqrt{\frac{77}{80} - \left(\frac{1}{80}\right)^2} \times 2$$

$$= \sqrt{0.9625 - 0.000} \times 2 = 0.9810 \times 2 = 1.962$$

Breaking strength of pieces in the range, $\bar{x} \pm 2\sigma$ is

$$\bar{x} \pm 2\sigma = 49.025 \pm 2 \times 1.962$$
$$= 45.103 \text{ and } 52.949 = 45 \text{ and } 53 \text{ (approx.)}$$

To calculate the percentage of observations lying between $\bar{x} \pm 2\sigma$, we assume that the number of observations (pieces) are equally spread within lower and upper boundary of each class interval (breaking strength). Since 45 is the mid-point of the class interval 44-46 with the frequency 3, therefore there are 1.5 frequencies at 45. Similarly, at 53 the frequency would be 2.5. Hence the total number of observations (frequencies) between 45 and 53 are = 1.5 + 24 + 27 + 21 + 2.5 = 76. So the percentage of observations lying within $\bar{x} \pm 2\sigma$ would be (76/80) × 100 = 95 per cent.

4.5.5 Coefficient of Variation

Standard deviation is an absolute measure of variation and expresses variation in the same unit of measurement as the arithmetic mean or the original data. A relative measure called the **coefficient of variation** (CV), developed by Karl Pearson is very useful measure for (i) comparing two or more data sets expressed in different units of measurement (ii) comparing data sets that are in same unit of measurement but the mean values of data sets in a comparable field are widely dissimilar (such as mean wages received per month by the top management personnel and labour class personnel of a large organization).

Coefficient of variation: A measure of relative variability computed by dividing the standard deviation by the mean, then multiplying by 100.

Thus, in view of this limitation we need to convert absolute measure of variation, that is, S.D. into a relative measure, which can be helpful in comparing the variability of two or more sets of data. The new measure, coefficient of variation (CV) measures the standard deviation relative to the mean in percentages. In other words, CV indicates how large the standard deviation is in relation to the mean and is computed as follows:

$$\text{Coefficient of variation (CV)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

Multiplying by 100 converts the decimal to a percent.

The set of data for which the coefficient of variation is low is said to be more uniform (consistent) or more homogeneous (stable).

Example 4.19: The weekly sales of two products A and B were recorded as given below:

Product A	:	59	75	27	63	27	28	56
Product B	:	150	200	125	310	330	250	225

Find out which of the two shows greater fluctuation in sales.

Solution: For comparing the fluctuation in sales of two products we will prefer to calculate coefficient of variation for both the products.

Product A: Let A = 56 be the assumed mean of sales for product A.

Table 4.14 Calculations of the Mean and Standard Deviation

Sales (x)	Frequency (f)	$d = x - A$	fd	fd^2
27	2	−29	−58	1682
28	1	−28	−28	784
(56) ← A	1	0	0	0
59	1	3	3	9
63	1	7	7	49
75	1	19	19	361
	7		−57	2885

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} = 56 - \frac{57}{7} = 47.86$$

$$s_A^2 = \frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2 = \frac{2885}{7} - \left(-\frac{57}{7}\right)^2$$

$$= 412.14 - 66.30 = 345.84$$

$$s_A = \sqrt{345.84} = 18.59$$

Then $$CV(A) = \frac{s_A}{\bar{x}} \times 100 = \frac{18.59}{47.86} \times 100 = 38.84 \text{ per cent}$$

Product B: Let A = 225 be the assumed mean of sales for product B.

Table 4.15 Calculations of Mean and Standard Deviation

Sales (x)	Frequency (f)	$d = x - A$	fd	fd^2
125	1	−100	−100	10,000
150	1	−75	−75	5625
200	1	−25	−25	625
225	1	0	0	0
250	1	25	25	625
310	1	85	85	7225
330	1	105	105	11,025
	7		15	35,125

$$\bar{x} = A + \frac{\Sigma fd}{\Sigma f} = 225 + \frac{15}{7} = 227.14$$

$$s_B^2 = \frac{\Sigma fd^2}{\Sigma f} - \left(\frac{\Sigma fd}{\Sigma f}\right)^2 = \frac{35{,}125}{7} - \left(\frac{15}{7}\right)^2$$

$$= 5017.85 - 4.59 = 5013.26$$

or $$s_B = \sqrt{5013.26} = 70.80$$

Then $$CV(B) = \frac{s_B}{\bar{x}} \times 100 = \frac{70.80}{227.14} \times 100 = 31.17 \text{ per cent}$$

Since the coefficient variation for product A is more than that of product B, therefore the sales fluctuation in case of product A is higher.

Example 4.20: From the analysis of monthly wages paid to employees in two service organizations X and Y, the following results were obtained:

	Organization X	*Organization Y*
Number of wage-earners	550	650
Average monthly wages	5000	4500
Variance of the distribution of wages	900	1600

(a) Which organization pays a larger amount as monthly wages?

(b) In which organization is there greater variability in individual wages of all the wage earners taken together?

Solution: (a) For finding out which organization X or Y pays larger amount of monthly wages, we have to compare the total wages:

Total wage bill paid monthly by X and Y is

$$X \ : \ n_1 \times \bar{x}_1 = 550 \times 5000 = \text{Rs. } 27,50,000$$

$$Y \ : \ n_2 \times \bar{x}_2 = 650 \times 4500 = \text{Rs. } 29,25,000$$

Organization Y pays a larger amount as monthly wages as compared to organization X.

(b) For calculating the combined variation, we will first calculate the combined mean as follows:

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{27,50,000 + 29,25,000}{1200} = \text{Rs } 4729.166$$

$$\sigma_{12} = \sqrt{\frac{n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_1^2\right)}{n_1 + n_2}}$$

$$= \sqrt{\frac{550(900 + 73,351.05) + 650(1600 + 52,517.05)}{550 + 650}}$$

$$= \sqrt{\frac{4,08,38,080.55 + 3,51,76,082.50}{1200}} = \sqrt{63345.13} = 251.68$$

where $d_1 = \bar{x}_{12} - \bar{x}_1 = 4729.166 - 5000 = -270.834$

$d_2 = \bar{x}_{12} - \bar{x}_2 = 4729.166 - 4500 = 229.166$

Conceptual Questions 4B

8. What purpose does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.

9. What do you understand by 'coefficient of variation'? Discuss its importance in business problems.

[UP Tech. Univ., MBA, 2000]

10. When is the variance equal to the standard deviation? Under what circumstances can variance be less than the standard deviation? Explain.

11. (a) Explain and illustrate how the measures of variation afford a supplement to the information about frequency distribution furnished by averages.

[Delhi Univ., MBA, 2001]

(b) Describe various methods of measuring variation. Which of these do you consider as the best and why?

12. Explain the advantages of standard deviation as a measure of variation over range and the average deviation. Under what circumstances will the variance of a variable be zero?

13. Comment on the comparative merits and demerits of measures of variation.

14. Explain the term 'variation'. What purpose does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.

[Delhi Univ., MBA, 1998]

15. Describe the various methods of measuring variation along with their respective merits and demerits.

[Delhi Univ., MBA, 1998]

16. It has been said that the lesser the variability that exists, the more an average is representative of a set of data. Comment.

17. (a) What information is provided by variance or standard deviation?

(b) What additional information about a set of data is provided by a measure of variability that is not obtained from an average?

18. What advantages are associated with variance and standard deviation relative to range as the measure of variability?

19. Suppose you read a published statement that the average amount of food consumption in this country is adequate; the overall conclusion based upon the statement is that

everyone is properly fed. Criticize the conclusion in terms of the concept of variability as it relates to the use of averages. [*Delhi Univ., MBA, 2000*]

20. The Vice-President, Sales has been studying records regarding the performance of his sales representatives. He has noticed that in the last 2 years, the average level of sales per representative has remained the same, while the distribution of the sales levels has widened. The sales levels from this period have significantly larger variations from the mean than in any of the previous 2 year periods for which he has records. What conclusions might be drawn from these observations? [*Delhi Univ., MBA, 1999*]

21. Explain Chebyshev's theorem which provides an approximation to the spread of a set of observations on either side of the mean.

22. Two economists are studying fluctuations in the price of gold. One is examining the period of 1998–2002. The other is examining the period of 1995–1999. What differences would you expect to find in the variability of their data?

23. How would you reply to the following statement: 'Variability is not an important factor because even though the outcome is more uncertain, you still have an equal chance of falling either above or below the median. Therefore on an average, the outcome will be the same.'

24. A retailer uses two different formulas for predicting monthly sales. The first formula has an average miss of 700 records, and a standard deviation of 35 records. The second formula has an average miss of 300 records, and a standard deviation of 16. Which formula is relatively less accurate?

Self-Practice Problems 4B

4.9 Find the average deviation from mean for the following distribution:

Quantity demanded (in units) :

60	61	62	63	64	65	66	67	68
Frequency :								
2	0	15	29	25	12	10	4	3

4.10 Find the average deviation from mean for the following distribution:

Dividend yield :

0–3	3–6	6–9	9–12	12–15	15–18	18–21
Number of companies :						
2	7	10	12	9	6	4

4.11 Find the average deviation from median for the following distribution:

Sales (Rs '000) :

1–3	3–5	5–7	7–9	9–11	11–13	13–15	15–17
Number of shops :							
6	53	85	56	21	26	4	4

4.12 In a survey of 48 engineering companies following data was collected:

Level of profit (Rs in lakh) :	10	11	12	13	14
Number of companies :	3	12	18	12	3

Calculate the variance and standard deviation for the distribution.

4.13 A manufacturer of T-shirts approaches you with the following information

Length of shoulder (in inches) :

12.0	12.5	13.0	13.5	14	14.5	15	15.5	16
Frequency:								
5	20	30	43	60	56	37	16	3

Calculate the standard deviation and advice the manufacturer as to the largest and the smallest shoulder size T-shirts he should make in order to meet the needs of his customers.

4.14 A charitable organization decided to give old-age pension to people over sixty years of age. The scales of pension were fixed as follows:

Age Group	*Pension/month (Rs)*
60–65	200
65–70	250
70–75	300
75–80	350
80–85	400

The ages of 25 persons who secured the pension are as given below:

74	62	84	72	61	83	72	81	64
71	63	61	60	67	74	64	79	73
75	76	69	68	78	66	67		

Calculate the monthly average pension payable per person and the standard deviation.

4.15 Two automatic filling machines A and B are used to fill tea in 500 g cartons. A random sample of 100 cartons on each machine showed the following:

Tea Contents (in g)	*Machine A*	*Machine B*
485–490	12	10
490–495	18	15
495–500	20	24
500–505	22	20
505–510	24	18
510–515	4	13

Comment on the performance of the two machines on the basis of average filling and dispersion.

4.16 An analysis of production rejects resulted in the following observations

No. of Rejects per Operator	*No. of Operator*	*No. of Rejects per Operator*	*No. of Operator*
21–25	5	41–45	15
26–30	15	46–50	12
31–35	28	51–55	3
36–40	42		

Calculate the mean and standard deviation.

[Delhi Univ., MBA, 2000]

4.17 Blood serum cholestrerol levels of 10 persons are as under:

240 260 290 245 255 288 272 263 277 250

Calculate the standard deviation with the help of assumed mean

4.18 32 trials of a process to finish a certain job revealed the following information:

Mean time taken to complete the job = 80 minutes
Standard deviation = 16 minutes

Another set of 8 trials gave mean time as 100 minutes and standard deviation equalled to 25 minutes.

Find the combined mean and standard deviation.

4.19 From the analysis of monthly wages paid to workers in two organizations X and Y, the following results were obtained:

	X	Y
Number of wage-earners	: 550	600
Average monthly wages (Rs)	: 1260	1348.5
Variance of distribution of wages (Rs)	: 100	841

Obtain the average wages and the variability in individual wages of all the workers in the two organizations taken together.

4.20 An analysis of the results of a budget survey of 150 families showed an average monthly expenditure of Rs 120 on food items with a standard deviation of Rs 15. After the analysis was completed it was noted that the figure recorded for one household was wrongly taken as Rs 15 instead of Rs 105. Determine the correct value of the average expenditure and its standard deviation.

4.21 The standard deviation of a distribution of 100 values was Rs 2. If the sum of the squares of the actual values was Rs 3,600, what was the mean of this distribution?

4.22 An air-charter company has been requested to quote a realistic turn-round time for a contract to handle certain imports and exports of a fragile nature.

The contract manager has provided the management accountant with the following analysis of turn-round times for similar goods over a given twelve-monthly period.

Turn-round Time (in hours)	*Frequency*
Less than 2	25
2 and < 4	36
4 and < 6	66
6 and < 8	47
8 and < 10	26
10 and < 12	18
12 and < 14	2

(a) Calculate mean and standard deviation.

(b) Advice the contract manager about the turn-round time to be quoted using

(i) mean plus one standard deviation;

(ii) mean plus two standard deviations.

4.23 The following relationship holds between two measures of temperature:

$$F^\circ = 32 + \frac{9}{5} C^\circ$$

where F° and C° denote the degree in daily average temperature measured in Fahrenheit and Centigrade.

If the variance of daily average temperature in a city throughout the year is 25°C, what is the variance in F° for that year and vice-versa.

4.24 The hourly output of a new machine is four times that of the old machine. If the variance of the hourly output of the old machine in a period of n hours is 16, what is the variance of the hourly output of the new machine in the same period of n hours.

4.25. The number of cheques cashed each day at the five branches of a bank during the past month has the following frequency distribution:

Number of Cheques	*Frequency*
0–199	10
200–399	13
400–599	17
600–799	42
800–999	18

The General manager, operations for the bank, knows that a standard deviation in cheque cashing of more than 200 checks per day creates staffing problem at the branches because of the uneven workload. Should the manager worry about staffing next month?

4.26. Mr. Gupta, owner of a Bakery, said that the average weekly production level of his company was 11,398 loaves, and the variance was 49,729. If data used to compute the results were collected for 32 weeks, during how many weeks was the production level below 11,175? and Above 11,844?

Coefficient of Variance

4.27 Two salesmen selling the same product show the following results over a long period of time:

	Salesman X	*Salesman Y*
Average sales volume per month (Rs)	30,000	35,000
Standard deviation	2,500	3,600

Which salesman seems to be more consistent in the volume of sales?

4.28 Suppose that samples of polythene bags from two manufacturers A and B are tested by a buyer for bursting pressure, giving the following results:

Bursting Pressure	Number of Bags	
	A	B
5.0–9.9	2	9
10.0–14.9	9	11
15.0–19.9	29	18
20.0–24.9	54	32
25.0–29.9	11	27
30.0–34.9	5	13

(a) Which set of bags has the highest bursting pressure?

(b) Which has more uniform pressure? If prices are the same, which manufacturer's bags would be preferred by the buyer? Why?

[Delhi Univ., MBA 1997]

4.29 The number of employees, average daily wages per employee, and the variance of daily wages per employee for two factories are given below:

		Factory A	Factory B
Number of employees	:	50	100
Average daily wages (Rs)	:	120	85
Variance of daily wages (Rs)	:	9	16

(a) In which factory is there greater variation in the distribution of daily wages per employee?

(b) Suppose in factory B, the wages of an employee were wrongly noted as Rs 120 instead of Rs 100. What would be the correct variance for factory B?

4.30 The share prices of a company in Mumbai and Kolkata markets during the last ten months are recorded below:

Month	Mumbai	Kolkata
January	105	108
February	120	117
March	115	120
April	118	130
May	130	100
June	127	125
July	109	125
August	110	120
September	104	110
October	112	135

Determine the arithmetic mean and standard deviation of prices of shares. In which market are the share prices more stable? *[HP Univ., MBA 2002]*

4.31 A person owns two petrol filling stations A and B. At station A, a representative sample of 200 consumers who purchase petrol was taken. The results were as follows:

Number of Litres of Petrol Purchased	Number of Consumers
0 and < 2	15
2 and < 4	40
4 and < 6	65
6 and < 8	40
8 and < 10	30
10 and over	10

A similar sample at station B users showed a mean of 4 litres with a standard deviation of 2.2 litres. At which station is the purchase of petrol relatively more variable?

Hints and Answers

4.9 MAD = 1.239; $\bar{x}$ = 63.89

4.10 $\bar{x}$ = 10.68; MAD = 3.823

4.11 Med = 6.612; MAD = 2.252

4.12 $\sigma^2 = 1$ and $\sigma = 1$

4.13 $\bar{x}$ = 14.013 inches; σ = 0.8706 inches; $\bar{x} + 3\sigma$ = 14.884 (largest size); $\bar{x} - 3\sigma$ = 13.142 (smallest size)

4.14 $\bar{x}$ = Rs 280.2; σ = Rs 60.765

4.15 Machine A: $\bar{x}_1 = 499.5$; $\sigma_1 = 7.14$; Machine B : $\bar{x}_2 = 500.5$; $\sigma_2 = 7.40$

4.16 $\bar{x}$ = 36.96; σ = 6.375

4.17 σ = 16.48

4.18 $\bar{x}_{12}$ = 84 minutes; σ_{12} = 19.84

4.19 $\bar{x}_{12}$ = Rs 1306; σ_{12} = Rs 53.14

4.20 Corrected $\bar{x}$ = Rs 120.6 and corrected σ = Rs 12.4

4.21 $\bar{x}$ = 5.66

4.22 (a) (i) $\bar{x}$ = 5.68 ; (ii) σ = 2.88

(b) (i) $\bar{x} + \sigma = 5.68 + 2.88 = 8.56$ hours

The chance of this turn-round time cover approx. 84%

(ii) $\bar{x} + 2\sigma = 5.68 + 2\,(2.88) = 11.44$ hours

The chance of this turn round time cover approx. 97.7%

4.24 Variance (new machine) = 256 hours

4.25 Mean $\mu = \dfrac{\Sigma fx}{N} = \dfrac{59{,}000}{100} = 590$ cheques per day

Standard deviatio, $\sigma = \sqrt{\dfrac{\Sigma f(x-\mu)^2}{N}} = \sqrt{\dfrac{58{,}70{,}000}{100}}$

= 242.48 cheques per day

Since standard deviation σ value is more than 200, the manager should worry.

4.26 The standard deviation for the distribution is $\sigma = \sqrt{\sigma^2} = \sqrt{49{,}729} = 223$. A production of 11,175 loaves is one standard deviation below the mean (11,398 – 11,175) = 223. Assuming that the distribution is symmetrical, we know that within $\mu \pm \sigma$ per cent about 68% of all observations fall. The interval from the mean to one standard deviation below the mean would contain about 34 per cent (68 per cent ÷ 2) of the data. Therefore, (50 – 34) = 16

per cent (or approx 5 weeks) of the data would be below 11,175 loaves.

4.27 Salesman X

4.28 Manufacture A: $\bar{x}_1 = 21, \sigma_1 = 4.875$ and C.V. = 23.32%

Manufacturer B : $\bar{x}_2 = 21.81$, $\sigma_2 = 7.074$ and C.V. = 32.44%; (a) Bags of manufacturer B have higher bursting pressure; (b) Bags of manufacturer A have more uniform pressure; (c) Bags of manufacturer A should be preferred by buyer as they have uniform pressure.

4.29 (a) CV(A) = 2.5 ;

CV(B) = 4.7. Variation in the distribution of daily wages per employee in factory B is more.

(b) Correct $\Sigma x = 100 \times 85 - 120 + 100 = 8{,}480$

Correct mean $\bar{x} = 8480/100 = 84.8$

Since $\sigma^2 = (\Sigma x^2/N) - (\bar{x})^2$

or $16 = (\Sigma x^2/100) - (85)^2$

$= \Sigma x^2 - 7{,}22{,}500$

or $\Sigma x^2 = 7{,}24{,}100$

Correct $\Sigma x^2 = 7{,}24{,}100 - (120)^2 + (100)^2$

$= 7{,}19{,}700$

Correct $\sigma^2 = (7{,}19{,}700/100) - (84.8)^2 = 5.96$

4.30 CV(Mumbai) = 7.24% ; CV (Kolkata) = 8.48%. This shows more stability in Mumbai stock market.

4.31 CV(A) = 46.02% ; CV(B) = 55%. The purchase of petrol is relatively more variable at station B.

Formulae Used

1. Range, R

   Value of highest observation – Value of lowest observation = H – L

   $$\text{Coefficient of range} = \frac{H-L}{H+L}$$

2. Interquartile range $= Q_3 - Q_1$

   $$\text{Quartile deviation, QD} = \frac{Q_3 - Q_1}{2}$$

   $$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

3. Mean average deviation

   For ungrouped data

   (i) $\text{MAD} = \frac{\Sigma |x - \bar{x}|}{n}$, for sample

   (ii) $\text{MAD} = \frac{\Sigma |x - \mu|}{N}$, for population

   (iii) $\text{MAD} = \frac{\Sigma |x - \text{Me}|}{n}$, from median

   For grouped data $\text{MAD} = \frac{\Sigma f|x - \bar{x}|}{\Sigma f}$

4. Coefficient of $\text{MAD} = \frac{\text{MAD}}{\bar{x} \text{ or Me}} \times 100$

5. Variance

   Ungrouped data

   $$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{N} = \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

   $$= \frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2$$

   where $d = x - A$; A is any assumed A.M. value

   $$\text{Grouped data, } \sigma^2 = \left[\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2\right] h$$

   where $d = (m - A)/h$; h is the class interval and m is the mid-value of class intervals.

6. Standard deviation

   Ungrouped data $\sigma = \sqrt{\sigma^2}$

   $$\text{Grouped data } \sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2} \times h$$

7. Coefficient of variation (CV) $= \frac{\sigma}{\bar{x}} \times 100$

Chapter Concepts Quiz

True or False

1. Range is a measure of variation which gives us information about scatter of values around a measure of central tendency. (T/F)
2. When a distribution consists of different observations, s or σ are relatively large. (T/F)
3. The interquartile range is based upon only two values in the data set. (T/F)
4. Absolute measures of variation are used for comparing variability among observations in a data set. (T/F)
5. The semi-interquartile range is inappropriate to use with skewed distributions. (T/F)

6. Mean absolute deviation taken from median is least. (T/F)
7. The standard deviation is measured in the same unit as the observations in the data set. (T/F)
8. In a symmetrical distribution, semi-interquartile range is one fourth of the range. (T/F)
9. The coefficient of variation is a relative measure of dispersion. (T/F)
10. The inter-quartile range measures the average range of the lower fourth of a distribution. (T/F)
11. For a symmetrical distribution, mean absolute deviation equals 4/5 of standard deviation. (T/F)
12. Variance indicates the average distance of any observation in the data set from the mean. (T/F)
13. Sample standard deviation provides an accurate estimate of the population standard deviation. (T/F)
14. Variance is the square of the standard deviation. (T/F)
15. Standard deviation can be calculated by taking deviation from any measure of central tendency. (T/F)

Multiple Choice

16. The standard deviation of a set of 50 observations is 8. If each observation is multiplied by 2, then the new value of standard deviation will be:
(a) 4 (b) 8
(c) 16 (d) none of the above
17. If mean and coefficient of variation of a set of data is 10 and 5, respectively, then the standard deviation is:
(a) 10 (b) 50
(c) 5 (d) none of the above
18. The semi-interquartile range is preferred to standard deviation as a measure of dispersion when:
(a) sample size is small
(b) distribution is standardized
(c) distribution is highly skewed
(d) range is small
19. In a more dispersed (spread out) set of data:
(a) difference between the mean and the median is greater
(b) value of the mode is greater
(c) standard deviation is greater
(d) inter-quartile range is smaller
20. Which of the following is a relative measure of dispersion:
(a) standard deviation
(b) variance
(c) coefficient of variation
(d) all of the above
21. In a normal frequency distribution, the number oif observations included in $\bar{x} \pm$ MAD are:
(a) 50 per cent (b) 57.51 per cent
(c) 68.51 per cent (d) none of the above
22. If quartile deviation is 8, then value of the standard deviation will be:
(a) 12 (b) 16
(c) 24 (d) none of the above
23. If mean absolute deviation is 8, then value of the standard deviation will be:
(a) 15 (b) 12
(c) 10 (d) none of the above
24. If the first and third quartiles are 22.16 and 56.36, respectively, then the quartile deviation is:
(a) 17.1 (b) 34.2
(c) 51.3 (d) none of the above.
25. The standard deviation of a set of 50 observations is 6.5. If value of each observation is increased by 5, then the standard deviation is:
(a) 2.5 (b) 1.5
(c) 3.5 (d) none of the above
26. The standard deviation of the first n natural numbers is:
(a) $\sqrt{\frac{1}{6}(n^2-1)}$ (b) $\sqrt{\frac{1}{6}(n^2+1)}$
(c) $\sqrt{\frac{1}{12}(n^2-1)}$ (d) $\sqrt{\frac{1}{12}(n^2+1)}$
27. The number of observations in a set of data covered by the interval, $\bar{x} \pm$ Q.D. are:
(a) 50 per cent (b) 57.73 per cent
(c) 59.23 per cent (d) none of the above
28. In a symmetrical distribution, observations covered in the interval $\bar{x} \pm 3\sigma$ are:
(a) 99.475 per cent (b) 99.65 per cent
(c) 99.73 per cent (d) none of the above
29. The relationship between mean absolute deviation and the quartile deviation is:
(a) MAD $= \frac{5}{6}$ Q.D. (b) MAD $= \frac{6}{5}$ Q.D.
(c) MAD $= \frac{4}{5}$ Q.D. (d) MAD $= \frac{5}{4}$ Q.D.
30. Which of the following measures of dispersion is least affected by extreme values of observations in a data set?
(a) range
(b) quartile deviation
(c) mean absolute deviation
(d) standard deviation
31. Which of the following is not a valid reason for measuring the dispersion of distribution?
(a) It provides an indication of the reliability of the statistic used to measure central tendency.
(b) It enables us to compare several samples with similar averages.
(c) It uses more data in discribing a distribution.
(d) It draws attention to problems associated with very small or very large variability in distributions.
32. Why is it necessary to square the differences from the mean when computing the population variance?
(a) So that extreme values will not affect the calculation.
(b) Because it is possible that N could be very small.
(c) Some of the differences will be positive and some will be negative.
(d) None of these.
33. Assume that a population has $\mu = 100$ and $\sigma = 10$. If a particular observation has a standard score of 1, it can be concluded that
(a) its value is 110.

(b) it lies between 90 and 110, but its exact value cannot be determined.
(c) its value is greater than 110.
(d) nothing can be determined without knowing N.

34. How does the computation of a sample variance differ from the computation of a population variance?
(a) μ is replaced by $\bar{x}$.
(b) N is replaced by $n - 1$.
(c) both (a) and (b)
(d) none of these.

35. Chebyshev's theorem says that 99 percent of the values will lie within ± 3 standard deviations from the mean for
(a) bell-shaped distributions.
(b) positively skewed distributions.
(c) negatively skewed distributions.
(d) all distributions.

Concepts Quiz Answers

1. F	**2.** T	**3.** T	**4.** F	**5.** T	**6.** T	**7.** T	**8.** T	**9.** T
10. F	**11.** T	**12.** T	**13.** F	**14.** F	**15.** F	**16.** (c)	**17.** (b)	**18.** (b)
19. (c)	**20.** (c)	**21.** (b)	**22.** (a)	**23.** (c)	**24.** (a)	**25.** (b)	**26.** (c)	**27.** (a)
28. (c)	**29.** (b)	**30.** (c)	**31.** (a)	**32.** (c)	**33.** (d)	**34.** (c)	**35.** (a)	

Review Self-Practice Problems

4.32 A petrol filling station has recorded the following data for litres of petrol sold per automobile in a sample of 680 automobiles:

Petrol Sold (Litres)	*Frequency*
0– 4	74
5– 9	192
10–14	280
15–19	105
20–24	23
25–29	6

Compute the mean and standard deviation for the data.

4.33 A frequency distribution for the duration of 20 long-distance telephone calls in minutes is as follows:

Call Duration (Minutes)	*Frequency*
4– 7	4
8–11	5
12–15	7
16–19	2
20–23	1
24–27	1

Compute the mean, variance, and standard deviation.

4.34 Automobiles travelling on a highway are checked for speed by the police. Following is a frequency distribution of speeds:

Speed (km per hours)	*Frequency*
45–49	10
50–54	40
55–59	150
60–64	175
65–69	75
70–74	15
75–79	10

What is the mean, variance, and standard deviation of speed for the automobiles travelling on the highway?

4.35 A work-standards expert observes the amount of time (in minutes) required to prepare a sample of 10 business letters in the office with observations in ascending order: 5, 5, 5, 7, 9, 14, 15, 15, 16, 18.
(a) Determine the range and middle 70 per cent range for the sample.
(b) If the sample mean of the data is 10.9, then calculate the mean absolute deviation and variance.

4.36 ABC Stereos, a wholesaler, was contemplating becoming the supplier to three retailers, but inventory shortages have forced him to select only one. ABC's credit manager is evaluating the credit record of these three retailers. Over the past 5 years these retailers' accounts receivable have been outstanding for the following average number of days. The credit manager feels that consistency, in addition to lowest average, is important. Based on relative dispersion, which retailer would make the best customer?

Lee :	62.20	61.80	63.40	63.00	61.70
Forest :	62.50	61.90	63.80	63.00	61.70
Davis :	62.00	61.90	63.00	63.90	61.50

[*Delhi Univ., MBA, 1999*]

4.37 A purchasing agent obtained samples of 60 watt bulbs from two companies. He had the samples tested in his own laboratory for length of life with the following results:

Length of Life (in hours)	*Samples from*	
	Company A	*Company B*
1700–1900	10	3
1900–2100	16	40
2100–2300	20	12
2300–2500	8	3
2500–2700	6	2

(a) Which company's bulbs do you think are better in terms of average life?

(b) If prices of both the companies are same, which company's bulbs would you buy and why?

[Delhi Univ., MBA, 2000]

4.38 The Chief Medical Officer of a hospital conducted a survey of the number of days 200 randomly chosen patients stayed in the hospital following an operation. The data are given below

Hospital stay (in days) :

1–3	4–6	7–9	10–12	13–15	16–18	19–21	22–24
Number of patients:							
18	90	44	21	9	9	4	5

(a) Calculate the mean number of days patients stay in the hospital along with standard deviation of the same.

(b) How many patients are expected to stay between 0 and 17 days.

4.39 A nursing home is well-known in effective use of pain killing drugs for seriously ill patients. In order to know approximately how many nursing staff to employ, the nursing home has began to keep track of the number of patients that come every week for checkup. Each week the CMO records the number of seriously ill patients and the number of routine patients. The data for the last 5 weeks is as follows:

Seriously ill patients :	33	50	22	27	48
Routine patients :	34	31	37	36	27

(a) Find the limits within which the middle 75 per cent of seriously ill patients per week should fall.

(b) Find the limits within which the middle 68 per cent of routine patients per week should fall.

4.40 There are a number of possible measures of sales performance, including how consistent a sales person is, in meeting established sales goals. The following data represent the percentage of goal met by each of three sales persons over the last five years

Raman :	88	68	89	92	103
Sindhu :	76	88	90	86	79
Prasad :	104	88	118	88	123

Which salesman is most consistent. Suggest an alternative measure of consistency (if possible).

4.41 Gupta Machine Company has a contract with one of its customers to supply machined pump gears. One requirement is that the diameter of its gears be within specific limits. The following data is of diameters (in inches) of a sample of 20 gears:

4.01 4.00 4.02 4.03 4.00 3.98 3.99 3.99
4.01 4.02 3.99 3.98 3.97 4.00 4.02 4.01
4.02 4.00 4.01 3.99

What can Gupta say to his customers about the diameters of 95 per cent of the gears they are receiving?

[Delhi Univ., MBA, 1998]

4.42 A production department uses a sampling procedure to test the quality of newly produced items. The department employs the following decision rule at an inspection station: If a sample of 14 items has a variance of more than 0.005, the production line must be shut down for repairs. Suppose the following data have just been collected:

3.43 3.45 3.43 3.48 3.52 3.50 3.39
3.48 3.41 3.38 3.49 3.45 3.51 3.50

Should the production line be shut down? Why or why not?

4.43 Police records show the following numbers of daily crime reports for a sample number of days during the winter months and a sample number of days during the summer months.

Winter :	18	20	15	16	21	20	12	16	19	20
Summer :	28	18	24	32	18	29	23	38	28	18

(a) Compute the range and inter-quartile range for each period.

(b) Compute the variance and standard deviation for each period.

(c) Compute the coefficient of variation for each period.

4.44 Public transportation and the automobiles are two options an employee can use to get to work each day. Samples of time (in minutes) recorded for each option are shown below:

Public transportation :

28	29	32	37	33	25	29	32	41	34

Automobile :

29	31	33	32	34	30	31	32	35	33

(a) Compute the sample mean time to get to work for each option.

(b) Compute the sample standard deviation for each option.

(c) On the basis of your results from parts (a) and (b), which method of transportation should be preferred? Explain.

4.45 The mean and standard deviation of a set of 100 observations were worked out as 40 and 5 respectively by a computer which, by mistake, took the value 50 in place of 40 for one observation. Find the correct mean and variance. *[Lucknow Univ., MBA, 1989]*

4.46 The number of employees, wages per employee and the variance of the wages per employee for two factories is given below:

	Factory A	Factory B
Number of employees :	100	150
Average wage per employee per month (Rs) :	3200	2800
Variance of the wages per employee per month (Rs) :	625	729

(a) In which factory is there greater variation in the distribution of wages per employee?

(b) Suppose in factory B, the wages of an employee were wrongly noted as Rs 3050 instead of Rs 3650, what would be the correct variance for factory B?

[Kumaun Univ., MBA, 1998]

4.47 In two factories A and B engaged in the same industry, the average weekly wages and standard deviations are as follows:

Factory	*Average Weekly Wages (Rs)*	*S.D. of Wages (Rs)*	*No. of Wage Earners*
A	460	50	100
B	490	40	80

(a) Which factory, A or B, pays a higher amount as weekly wages?

(b) Which factory shows greater variability in the distribution of wages?

(c) What is the mean and standard deviation of all the workers in two factories taken together?

[HP Univ., MBA ; Vikram Univ., MBA, 1997]

4.48 The mean of 5 observations is 4.4 and the variance is 8.24. If three of the five observations are 1, 2 and 6, find the other two.

4.49 The mean and standard deviation of normal distribution are 60 and 5 respectively. Find the inter-quartile range and the mean deviation of the distribution:

[Delhi Univ., BCom (H), 1997]

4.50 Mean and standard deviation of the following continuous series are 31 and 5.9 respectively. The distribution after taking step deviations is as follows:

Step deviation, d :	-3	-2	-1	0	1	2	3
Frequency, f :	10	15	25	25	10	10	5

Determine the actual class intervals.

[Delhi Univ., BCom (H) 1998]

4.51 The value of the arithmetic mean and standard deviation of the following frequency distribution of a continuous variable derived from the use of working origin and scale are Rs. 107 and 13.1 respectively. Determine the actual classes.

Step deviation, d :	-3	-2	-1	0	$+1$	$+2$
Frequency, f :	1	3	4	7	3	2

[Ranchi Univ., MBA, 1998]

4.52 The mean and standard deviation of a set of 100 observations were found to be 40 and 5 respectively. But by mistake a value 50 was taken in place of 40 for one observation. Re-calculate the correct mean and standard deviation. *[Lucknow Univ., MBA, 1999]*

4.53 The mean and the standard deviation of a sample of 10 sizes were found to be 9.5 and 2.5 respectively. Later on, an additional observation became available. This was 15.0 and was included in the original sample. Find the mean and the standard deviation of 11 observations.

4.54 The Shareholders Research Centre of India has recently conducted a research-study on price behaviour of three leading industrial shares, A, B, and C for the period 1979 to 1985, the results of which are published as follows in its Quarterly Journal:

Share	*Average Price (Rs)*	*Standard Deviation*	*Current Selling Price (Rs)*
A	18.2	5.4	36.00
B	22.5	4.5	34.75
C	24.0	6.0	39.00

(a) Which share, in your opinion, appears to be more stable in value?

(b) If you are the holder of all the three shares, which one would you like to dispose of at present, and why?*[HP Univ., MCom; Jammu Univ., MCom, 1997]*

4.55 Find the missing information from the following:

	Group I	*Group II*	*Group III*	*Combined*
Number	50	?	90	200
Std. dev.	6	7	?	7.746
Mean	113	?	115	116

[HP Univ., MBA; Osmania Univ., MBA, 1997]

4.56 An analysis of the weekly wages paid to workers in two firms A and B belonging to the same industry, gives the following results:

	Firm A	*Firm B*
Number of wage-earners	550	650
Average daily wages	50	45
Standard deviation of the distribution of wages	$\sqrt{90}$	$\sqrt{120}$

(a) Which firm, A or B, pays out a larger amount as daily wages?

(b) In which firm, A or B, is there greater variability in individual wages?

(c) What are the measures of (i) average daily wages and (ii) standard deviation in the distribution of individual wages of all workers in the two firms taken together?

[M.D. Univ., MBA; Diploma in Mgt., AIMA, Dec., 1999]

Hints and Answers

4.32 $\bar{x}$ = 10.74 litres per automobile, σ = 5.00 litres

4.35 (a) Range = H – L = 18 – 5 = 15 minutes

Middle 70% of R = $P_{85} - P_{15}$

$= x_{(85/100)+(1/2)} - x_{(15/100)+(1/2)}$

$= x_{(8.5+0.5)} - x_{(1.5+0.5)} = x_9 - x_2$

= 16 – 5 = 11 minutes

(b) MAD $= \dfrac{\Sigma|x-\bar{x}|}{n} = \dfrac{47}{10}$ = 4.7 minutes

$$s^2 = \frac{\Sigma x^2 - n\bar{x}^2}{n-1} = \frac{1{,}431 - 10(10.9)^2}{10-1}$$

= 26.99 minutes

4.36 Lee : $\bar{x}$ = 62.42, s = 0.7497,

CV = $(s/\bar{x}) \times 100$ = 1.20%

Forest: $\bar{x}$ = 62.18, s = 0.9257,

CV = $(s/\bar{x}) \times 100 = 1.49\%$

Davis : $\bar{x} = 62.46$, $s = 0.9762$,

CV = $(s/\bar{x}) \times 100 = 1.56\%$

Based on CV, Lee would be the best customer

4.37 For company A:

$$\bar{x} = A + \frac{\Sigma fd}{N} \times h = 2200 - \frac{16}{60} \times 200 = 2146.67;$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{88}{60} - \left(\frac{-16}{60}\right)^2} \times 200 = 236.4$$

CV = $(\sigma/\bar{x}) \times 100 = 11\%$

For company B: $\bar{x} = 2070$; $\sigma = 158.8$ and CV = 7.67%.

(a) Bulbs of company A are better.

(b) CV(B) < CV(A): Buy company B bulbs as their burning hours are more uniform.

4.38 (a) $\bar{x} = \frac{\Sigma fm}{n} = \frac{1543}{200} = 7.715$ days;

$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n-1} = \frac{4384.755}{199} = 22.033$$

$s = \sqrt{22.033} = 4.69$ days

(b) $z = \frac{x - \bar{x}}{s} = \frac{0 - 7.715}{4.69} = -1.644$, i.e. zero day stay is 1.64 standard deviation below the mean, and

$z = \frac{x - \bar{x}}{s} = \frac{17 - 7.715}{4.69} = 1.97$, i.e. 17 days stay in 1.97 standard deviation above the mean.

Applying the Chebyshev's theorem with $z = 1.97$, we have

$$\left(1 - \frac{1}{z^2}\right) = \left[1 - \frac{1}{(1.97)^2}\right] = 0.743$$

This indicates that at least 75% patients, i.e. 0.75 (200) = 150 patients should stay between 0 and 17 days.

4.39 (a) $\bar{x} = \frac{\Sigma x}{n} = 180 \div 5 = 36$ patients

$$s = \sqrt{\frac{\Sigma x^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{7106 - 5(36)^2}{4}}$$

$= \sqrt{156.5} = 12.51$ patients

The middle 75% of data should be in the interval, $\bar{x} \pm 2s = 36 \pm 2\,(12.51) = (11, 61)$ patients.

(b) $\bar{x} = \frac{\Sigma x}{n} = 165 \div 5 = 33$ patients

$$s = \sqrt{\frac{\Sigma x^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{5511 - 5(33)^2}{4}}$$

$= \sqrt{16.5} = 4.06$ patients

If distribution is normal, then middle 68% of data should be in the interval $\bar{x} \pm s = 33 \pm 4.06 = (29, 37)$ patients.

4.40

Sales Person	$\bar{x}$	s	$CV = (s/\bar{x}) \times 100$
Raman	88	12.67	14.4%
Sindhu	83.8	6.02	7.2%
Prasad	104.2	16.35	15.7%

Sindhu is most consistent both in terms of s and CV.

4.41

Diameter :	3.97	3.98	3.99	4.00	4.01	4.02	4.03
Frequency:	1	2	4	4	4	4	1

$$\bar{x} = \frac{\Sigma x}{n} = 80.04 \div 20 = 4.002 \text{ inches}$$

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

$$= \sqrt{\frac{320.325 - 20(4.002)^2}{19}} = 0.016 \text{ inches}$$

If distribution is bell-shaped, then 95% of the gears will have diameters in the interval: $\bar{x} \pm 2s = 4.002 \pm 2\,(0.016) = (3.970, 4.034)$ inches.

4.44 (a) Public : 32; Auto : 32 (b) Public : 4.64; Auto : 1.83 (c) Auto has less variability.

4.45 (i) $\bar{x} = \frac{\Sigma x}{n}$ or $\Sigma x = \bar{x}\,N = 40 \times 100 = 4{,}000$

Correct $\Sigma x = 4000 - 50 + 40 = 3990$. Thus Correct, $\bar{x} = 3990 \div 100 = 39.9$

(ii) $\sigma^2 = \frac{\Sigma x^2}{N} - (\bar{x})^2$ or $25 = \frac{\Sigma x^2}{100} - (40)^2$

or $\Sigma x^2 = 1{,}62{,}500$

Correct $\Sigma x^2 = 1{,}62{,}500 - (50)^2 + (40)^2$

$= 1{,}62{,}500 - 2500 + 1600$

$= 1{,}61{,}600$

Correct $\sigma^2 = \frac{\text{Correct } \Sigma x^2}{N} - (\text{Correct } \bar{x})^2$

$= \frac{1{,}61{,}600}{100} - (39.9)^2 = 23.99$

4.46 (a) CV (A) = $\frac{\sigma}{\bar{x}} \times 100 = \frac{\sqrt{625}}{3200} \times 100 = 0.781$;

CV(B) = $\frac{\sqrt{729}}{2800} \times 100 = 0.964$

There is greater variation in the distribution of wages per employee in factory B.

(b) $\bar{x} = \frac{\Sigma x}{N}$ or $\Sigma x = N\bar{x} = 150 \times 2800 = 4{,}20{,}000$

Correct $\Sigma x = 4{,}20{,}000 - 3050 + 3650 = 4{,}20{,}600$;

Correct, $\bar{x} = \frac{4{,}20{,}600}{150} = 2{,}804$

Variance, $\sigma^2 = \frac{\Sigma x^2}{N} - (\bar{x})^2$

or $729 = \frac{\Sigma x^2}{150} - (2800)^2$

or $\Sigma x^2 = 1{,}17{,}61{,}09{,}350$

$$\text{Correct } \Sigma x^2 = 1,17,61,09,350 - (3050)^2 + (3650)^2 = 1,18,01,29,350$$

$$\text{Correct } \sigma^2 = \frac{\text{Correct } \Sigma x^2}{N} - (\text{Correct } \bar{x})^2$$

$$= \frac{1,18,01,29,350}{150} - (2804)^2 = 5113$$

4.47 (a) Total weekly wages: Factory A = 460 × 100 = Rs 46,000; Factory B = 490 × 80 = Rs 39,200
Factory A pays a larger amount.

(b) $CV(\text{Factory A}) = \frac{\sigma}{\bar{x}} \times 100 = \frac{50}{460} \times 100 = 10.87\%$;

$CV(\text{Factory B}) = \frac{40}{490} \times 100 = 8.16\%$

Since CV (A) × CV (B), factory A shows greater variability in wages.

(c) $\bar{x}_{12} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{100 \times 460 + 80 \times 490}{100 + 80}$

= Rs 473.33

$$\sigma_{12}^2 = \frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N_1 + N_2};$$

$$= \frac{100\{(50)^2 + (13.33)^2\} + 80\{(40)^2 + (16.67)^2\}}{100 + 80}$$

= 48.19

$d_1 = |\bar{x}_1 - \bar{x}_{12}| = |460 - 473.33| = 13.33$

$d_2 = |\bar{x}_2 - \bar{x}_{12}| = |490 - 473.33| = 16.67$

4.48 $\bar{x} = \frac{\Sigma x}{N}$ or $\Sigma x = N\bar{x} = 5 \times 4.4 = 22$

Let two terms x_1 and x_2 are missing. Then
$x_1 + x_2 + 1 + 2 + 6 = 22$ or $x_1 + x_2 = 13$

Also $\sigma^2 = \frac{\Sigma x^2}{N} - (\bar{x}^2)$ or $8.24 = \frac{\Sigma x^2}{5} - (4.4)^2$

or $\Sigma x^2 = 138$

$\therefore \Sigma x^2 = x_1^2 + x_2^2 + (1)^2 + (2)^2 + (6)^2$

$= 138$ or $x_1^2 + x_2^2 = 97$

Now $x_1^2 + x_2^2 = (x_1 + x_2)^2 - 2x_1x_2$

or $97 = (13)^2 - 2x_1x_2$ or $x_1x_2 = 36$

$(x_1 - x_2)^2 = x_1^2 + x_2^2 - 2x_1x_2 = 97 - 2(36) = 25,$

or $x_1 - x_2 = 5$

Solving two equations $x_1 + x_2 = 13$ and $x_1 - x_2 = 5$, we have $x_1 = 9$ and $x_2 = 4$.

4.49 $QD = \frac{2}{3}\sigma = \frac{2}{3} \times 5 = \frac{10}{3}$;

$QD = \frac{Q_3 - Q_2}{2} = \frac{10}{3}$ or $Q_3 - Q_1 = \frac{20}{3} = 6.67$

Thus interquartile range is 6.67

4.51 $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h$

or $13.1 = \sqrt{\frac{36}{20} - \left(\frac{-6}{20}\right)^2} \times h$ or $h = 10$

$\bar{x} = A + \frac{\Sigma fd}{N} \times h$

or $107 = A - \frac{6}{20} \times 10$ or A = 110 (assumed mean)

Since deviations are taken from A = 110 and class interval is, $h = 10$, therefore the class corresponding to $d = 0$ will be 105–115. Other classes will be:

Class :					
75–85	85–95	95–105	105–115	115–125	125–135
Frequency :					
1	3	4	7	3	2

4.52 Correct $\bar{x} = 39.9$ and $\sigma = 4.9$

4.53 $\bar{x} = \frac{\Sigma x}{N}$ or $\Sigma x = N\bar{x} = 10 \times 9.5 = 95$

Adding the 11th observation,

We get $\Sigma x = 95 + 15 = 110$.

Then $\bar{x} = \frac{\Sigma x}{N} = 110 \div 11 = 10$

Also, $\sigma^2 = \frac{\Sigma x^2}{N} - (\bar{x})^2$ or $(2.5)^2 = \frac{\Sigma x^2}{10} - (9.5)^2$

or $\Sigma x^2 = 965$

Now value of $\Sigma x^2 = 965 + (15)^2 = 1190$. Then

$$\sigma^2 = \frac{\Sigma x^2}{N} - (\bar{x})^2 = \frac{1190}{11} - (10)^2 = 2.86$$

4.54 (a) CV(A) = 30, CV(B) = 20 and CV(C) = 25; Share B is more stable.

(b) Dispose share A because of high variability in its price.

4.55 Given $N_1 + N_2 + N_3 = 200$, $N_1 = 50$, $N_3 = 90$, therefore $N_2 = 60$

$$\bar{x}_{123} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

$$\text{or } 116 = \frac{50 \times 113 + 60 \times \bar{x}_2 + 90 \times 115}{200}$$

or $\bar{x}_2 = 120$

$$\sigma_{123}^2 = \frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2) + N_3(\sigma_3^2 + d_3^2)}{N_1 + N_2 + N_3}$$

$$60 = \frac{50\{(6)^2 + (-3)^2\} + 60\{(7)^2 + (4)^2\} + 90\{\sigma_3^2 + (-1)^2\}}{200}$$

or $\sigma_3^2 = 64$ or $\sigma_3 = 8$; where $d_1 = \bar{x}_1 - \bar{x}_{123}$; $d_2 = \bar{x}_2 - \bar{x}_{123}$; $d_3 = \bar{x}_3 - \bar{x}_{123}$

4.56 (a) Firm B pays more wages;
(b) Firm B has greater variability in individual wages
(c) $\bar{x}_{12} = 47.29$ and $\sigma_{12} = 10.605$

Case Studies

Case 4.1: Himgiri Hospital

The hospital recently has installed a new computer-based, interactive, hospital communication system. The system fully integrates the communication activities of admitting, nursing, physician services, laboratory, radiology, pharmacy and assorted medication services, business office, medical records, central supply, dietary services, emergency, and outpatient.

In special training sessions with physicians who were to use the system, the director of the hospital observed that one of the key variables affecting the physicians was the 'waiting time' they experienced between inputting data or information requests at a video matrix terminal and the response by the mainframe computer. One of the doctor who is cardiologist was particularly vocal in his complaints about the system: 'Look, I can't wait all day for a machine. I need information that is accurate and in a form I can use. You can't expect me to also spend time learning how to use your machine—I have enough to do.'

To the physicians, sitting at a terminal and waiting for the computer to respond was simply 'intolerable.' The director of the hospital was sympathetic to the physicians' attitude and had negotiated a contract with the computer hardware vendor specifying that the average waiting time not to exceed 10 seconds.

After the system has been operating nearly 15 months, the director conducted a full-scale evaluation. In general, all aspects of the system looked either good or excellent with the exception that only about 60 percent of the physicians were actually using it, and over the past several months there had been a number of complaints about excessive waiting times.

The director was considering the possibility of holding a new series of training sessions for the physicians, but he decided to first review the data collected on actual waiting times experienced by the physicians. These sets of data were available: those collected during the original training session in January 2003 and those collected by staff analysts in March 2004 and May, 2004. These waiting-time (in seconds) data are given below:

January 2003			*March 2004*				*May 2004*			
9	8	5	8	6	14	12	7	7	15	16
6	6	7	12	8	7	10	13	7	17	15
9	9	8	12	10			14	7		
7										

Questions for Discussion

1. Calculate the mean waiting time for each of the three sets of data. Do the mean waiting times appear to be in conformance with the established standard?
2. Calculate the median waiting times for each of the three sets of data. What general conclusions can you draw?
3. Determine the range and standard deviation for each of the three sets of data and consider the implications of the results.

Chapter 5

While an individual is an insolvable puzzle, in an aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to.

—Arthur Conan Doyle

Skewness, Moments, and Kurtosis

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- know the complementary relationship of skewness with measures of central tendency and dispersion in describing a set of data.
- understand 'moments' as a convenient and unifying method for summarizing several descriptive statistical measures.

5.1 INTRODUCTION

In Chapter 4 we discussed measures of variation (or disperson) to describe the spread of individual values in a data set around a central value. Such descriptive analysis of a frequency distribution remains incomplete until we measure the degree to which these individual values in the data set deviate from symmetry on both sides of the central value and the direction in which these are distributed. This analysis is important due to the fact that data sets may have the same mean and standard deviation but the frequency curves may differ in their shape. A frequency distribution of the set of values that is not 'symmetrical (normal)' is called *asymmetrical or skewed*. In a skewed distribution, extreme values in a data set move towards one side or tail of a distribution, thereby lengthening that tail. When extreme values move towards the upper or right tail, the distribution is positively skewed. When such values move towards the lower or left tail, the distribution is negatively skewed. As discussed, the mean, median, and mode are affected by the high-valued observations in any data set. Among these measures of central tendency, the mean value gets affected largely due to the presence of high-valued observations in one tail of a distribution. The mean value shifted substantially in the direction of high-values. The mode value is unaffected, while the median value, which is affected by the numbers but not the values of such observations, is also shifted in the direction of high-valued observations, but not as far as the mean. The median value changes about 2/3 as far as the mean value in the direction of high-valued observations (called extremes). Symmetrical and skewed distributions are shown in Fig. 5.1.

Measure of skewness is the statistical technique to indicate the direction and extent of skewness in the distribution of numerical values in the data set.

For a positively skewed distribution A.M. > Median > Mode, and for a negatively skewed distribution A.M. < Median < Mode. The relationship between these measures of central tendency is used to develop a **measure of skewness** called the *coefficient of skewness* to understand the degree to which these three measures differ.

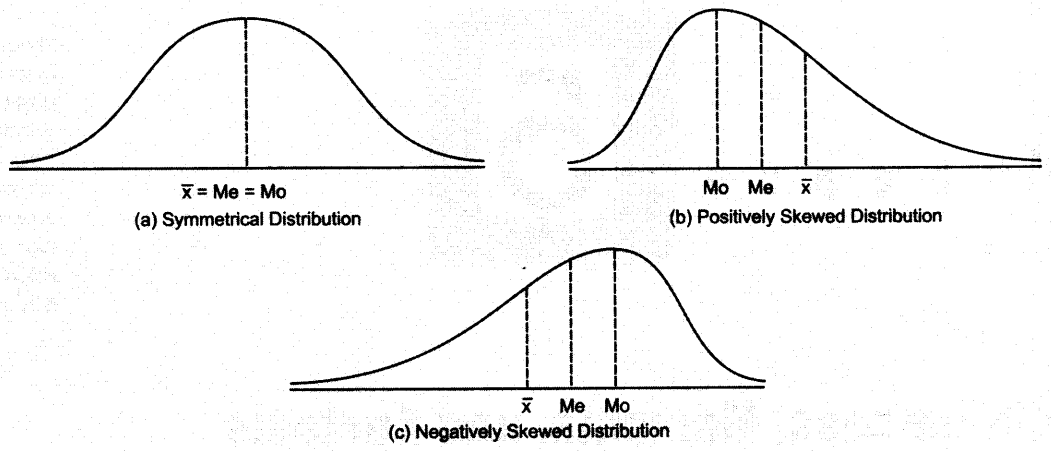


Figure 5.1
Comparison of Three Data Sets Differing in Shape

From the above discussion two points of difference emerge between variation and skewness:

(i) Variation indicates the amount of spread or dispersion of individual values in a data set around a central value, while skewness indicates the direction of dispersion, that is, away from symmetry.

(ii) Variation is helpful in finding out the extent of variation among individual values in a data set, while skewness gives an understanding about the concentration of higher or lower values around the mean value.

5.2 MEASURES OF SKEWNESS

The degree of skewness in a distribution can be measured both in the *absolute* and *relative* sense. For an asymmetrical distribution, the distance between mean and mode may be used to measure the degree of skewness because the mean is equal to mode in a symmetrical distribution. Thus,

$$\text{Absolute } S_k = \text{Mean} - \text{Mode}$$
$$= Q_3 + Q_1 - 2 \text{ Median (if measured in terms of quartiles).}$$

For a positively skewed distribution, Mean > Mode and therefore S_k is a positive value, otherwise it is a negative value. This difference is taken to measure the degree of skewness because in an asymmetrical distribution mean moves away from the mode. Larger the difference between mean and mode, whether positive or negative, more is the asymmetrical distribution or skewness. This difference, however, may not be desirable for the following reasons:

(i) The difference between mean and mode is expressed in the same units as the distribution and therefore cannot be used for comparing skewness of two or more distributions having different units of measurement.

(ii) The difference between mean and mode may be large in one distribution and small in another, although the shape of their frequency curves is the same.

In order to overcome these two shortcomings and to make valid comparisons between skewness of two or more distributions, the absolute difference has to be expressed in relation to the standard deviation— a measure of dispersion. Since we want to express any measure of skewness as a pure (relative) number, therefore this distance is expressed in terms of the unit of measurement in units of the standard deviation.

5.2.1 Relative Measures of Skewness

The following are three important relative measures of skewness.

Karl Pearson's coefficient of skewness

The measure suggested by Karl Pearson for measuring coefficient of skewness is given by:

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{\bar{x} - \text{Mo}}{\sigma} \tag{5-1}$$

where Sk_p = Karl Pearson's coefficient of skewness.

Since a mode does not always exist uniquely in a distribution, therefore it is convenient to define this measure using median. For a moderately skewed distribution the following relationship holds:

$$\text{Mean} - \text{Mode} = 3(\text{Mean} - \text{Median}) \quad \text{or} \quad \text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

When this value of mode is substituted in the Eqn. (5-1) we get

$$Sk_p = \frac{3(\bar{x} - \text{Me})}{\sigma} \tag{5-2}$$

Theoretically, the value of Sk_p varies between $\pm$ 3. But for a moderately skewed distribution, value of $Sk_p = \pm 1$. Karl Pearson's method of determining coefficient of skewness is particularly useful in open-end distributions.

Bowley's Coefficients of Skewness

The method suggested by Prof. Bowley is based on the relative positions of the median and the quartiles in a distribution. If a distribution is symmetrical, then Q_1 and Q_3 would be at equal distances from the value of the median, that is,

$$\text{Median} - Q_1 = Q_3 - \text{Median}$$

or $$Q_3 + Q_1 - 2\,\text{Median} = 0 \quad \text{or} \quad \text{Median} = \frac{Q_3 + Q_1}{2}$$

This shows that the value of median is the mean value of Q_1 and Q_3. Obviously in such a case the absolute value of the coefficient of skewness will be zero.

When a distribution is asymmetrical, quartiles are not at equal distance from the median. The distribution is positively skewed, if $Q_1 - \text{Me} > Q_3 - \text{Me}$, otherwise negatively skewed.

The absolute measure of skewness is converted into a relative measure for comparing distributions expressed in different units of measurement. For this, absolute measure is divided by the inter-quartile-range. That is,

$$\text{Relative } Sk_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1} = \frac{(Q_3 - \text{Med}) - (\text{Med} - Q_1)}{(Q_3 - \text{Med}) + (\text{Med} - Q_1)} \tag{5-3}$$

In a distribution, if Med = Q_1, then $Sk_b = \pm 1$, but of Med = Q_3 then $Sk_b = -1$. This shows that the value of Sk_b varies between $\pm$ 1 for moderately skewed distribution. This method of measuring skewness is quite useful in those cases where (*i*) mode is ill-defined and extreme observations are present in the data, (*ii*) the distribution has open-end classes. These two advantages of Bowley's coefficient of skewness indicate that it is not affected by extreme observations in the data set.

Remark: The values of Sk_b obtained by Karl Pearson's and Bowley's methods cannot be compared. On certain occasions it is possible that one of them gives a positive value while the other gives a negative value.

Kelly's Coefficient of Skewness

The relative measure of skewness suggested by Prof. Kelly is based on percentiles and deciles:

$$Sk_k = \frac{P_{10} + P_{90} - 2P_{50}}{P_{90} - P_{10}} \quad \text{or} \quad \frac{D_1 + D_9 - 2D_5}{D_9 - D_1} \tag{5-4}$$

This method is an extension of Bowley's method in the sense that Bowley's method is based on the middle 50 per cent of the observations while this method is based on the observations between the 10th and 90th percentiles (or first and nineth deciles).

Example 5.1: Data of rejected items during a production process is as follows:

No of rejects (per operator)	21–25	26–30	31–35	36–40	41–45	46–50	51–55
No. of operaters	5	15	28	42	15	12	3

Calculate the mean, standard deviation, and coefficient of skewness and comment on the results.

Solution: The calculations for mean, mode, and standard deviation are shown in Table 5.1

Table 5.1 Calculations for Mean, Mode and Standard Deviation

Class	Mid-value (m)	Frequency (f)	$d = \frac{m-A}{h} = \frac{m-38}{5}$	fd	fd^2
21–25	23	5	−3	−15	45
26–30	28	15	−2	−30	60
31–35	33	28 $\leftarrow f_{m-1}$	−1	−28	28
36–40	38	42 $\leftarrow f_m$	0	0	0
41–45	43	15 $\leftarrow f_{m+1}$	1	15	15
46–50	48	12	2	24	48
51–55	53	3	3	9	27
		N = 120		−25	223

Let assumed mean, A = 38. Then

$$\bar{x} = A + \frac{\Sigma fd}{N} \times h = 38 - \frac{25}{120} \times 5 = 36.96 \text{ rejects per operator}$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{223}{120} - \left(\frac{-25}{120}\right)^2} \times 5 = 6.736 \text{ rejects per operator}$$

By inspection, mode lies in the class 36–40. Thus

$$Mo = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

$$= 36 + \frac{42 - 28}{2 \times 42 - 28 - 15} = 36 + \frac{16}{41} \times 5 = 37.70$$

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{\bar{x} - Mo}{\sigma}$$

$$= \frac{36.96 - 37.70}{6.73} = \frac{-0.74}{6.736} = -0.109$$

Since the coefficient of skewness, $S_k = -0.109$, the distribution is skewed to left (negatively skewed). Thus the concentration of the rejects per operator is more on the lower values of the distribution to the extent of 10.9 per cent.

Example 5.2: The following is the information about the settlement of an industrial dispute in a factory. Comment on the gains and losses from the point of view of workers and that of management:

	Before	*After*
No. of Workers	3000	2900
Mean wages (Rs)	2200	2300
Median wages (Rs)	2500	2400
Standard deviation	300	260

Solution: The comments on gains and losses from both workers and management point of view are as follows:

Total Wages Bill

Before	*After*
$3000 \times 2200 = 66,00,000$	$2900 \times 2300 = 66,70,000$

The total wage bill has increased after the settlement of dispute, workers retained after the settlement are 50 workers less than the previous number.

After the settlement of dispute, the workers as a group are better off in terms of monetary gain. If the workers' efficiency remain same, then it is against the interest of management. But if the workers feel motivated, resulting in increased efficiency, then management can achieve higher productivity. This would be an indirect gain to management also.

Since workers retained after the settlement of dispute are less than the number employed before, it is against the interest of the workers.

Median Wages

The median wage after the settlement of dispute has come down from Rs 2500 to Rs 2400. This indicates that before the settlement 50 per cent of the workers were getting wages above Rs 2500 but after the settlement they will be getting only Rs 2400. It has certainly gone against the interest of the workers.

Uniformity in the Wage Structure

The extent of relative uniformity in the wage structure before and after the settlement can be determined by comparing the coefficient of variation as follows:

	Before	*After*
Coefficient of variation (CV)	$\frac{300}{2200} \times 100 = 13.63$	$\frac{260}{2300} \times 100 = 11.30$

Since CV has decreased after the settlement from 13.63 to 11.30, the distribution of wages is more uniform after the settlement, that is, there is now comparatively less disparity in the wages received by the workers. Such a position is good for both the workers and the management in maintaining a cordial work environment.

Pattern of the Wage Structure

The nature and pattern of the wage structure before and after the settlement can be determined by comparing the coefficients of skewness.

	Before	*After*
Coefficient of skewness Sk_p	$\frac{3(2200-2500)}{300} = -3$	$\frac{3(2300-2400)}{260} = -1.15$

Since coefficient of skewness is negative and has increased after the settlement, therefore it suggests that number of workers getting low wages has increased and that of workers getting high wages has decreased after the settlement.

Example 5.3: From the following data on age of employees, calculate the coefficient of skewness and comment on the result

Age below (years) :	25	30	35	40	45	50	55
Number of employees :	8	20	40	65	80	92	100

[*Delhi Univ., MBA, 1997*]

Solution: The data are given in a cumulative frequency distribution form. So to calculate the coefficient of skewness, convert this data into a simple frequency distribution as shown in Table 5.2.

Table 5.2 Calculations for Coefficient of Skewness

Age (years)	*Mid-value (m)*	*Number of Employees (f)*	*d = (m − A)/h = (m − 37.5)*	*fd*	*fd²*
20–25	22.5	8	−3	−24	72
25–30	27.5	12	−2	−24	48
30–35	32.5	20 ← f_{m-1}	−1	−20	20
35–40	37.5	25 ← f_m	0	0	0
40–45	42.5	1 ← f_{m+1}	1	15	15
45–50	47.5	12	2	24	48
50–55	52.5	8	3	24	72
		N = 100		−5	275

$$\text{Mean, } \bar{x} = A + \frac{\Sigma fd}{N} \times h = 37.5 - \frac{5}{100} \times 5 = 37.25$$

Mode value lies in the class interval 35–40. Thus

$$\text{Mo} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

$$= 35 + \frac{25 - 20}{2 \times 25 - 20 - 15} \times 5 = 35 + \frac{5}{15} \times 5 = 36.67$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h$$

$$= \sqrt{\frac{275}{100} - \left(\frac{-5}{100}\right)^2} \times 5 = \sqrt{2.75 - 0.0025} \times 5 = 8.29$$

Karl Pearson's coefficient of skewness:

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\sigma} = \frac{37.25 - 36.67}{8.29} = \frac{0.58}{8.29} = 0.07$$

The positive value of Sk_p indicates that the distribution is slightly positively skewed.

Example 5.4: (a) The sum of 50 observations is 500, its sum of squares is 6000 and median 12. Find the coefficient of variation and coefficient of skewness.

(b) For a moderately skewed distribution, the arithmetic mean is 100 and coefficient of variation 35, and Pearson's coefficient of skewness is 0.2. Find the mode and the median.

Solution: (a) Given that N = 50, $\Sigma x = 500$, $\Sigma x^2 = 6000$ and Me = 12.

$$\text{Mean, } \bar{x} = \frac{\Sigma x}{N} = \frac{500}{50} = 10$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\Sigma x^2}{N} - (\bar{x})^2} = \sqrt{\frac{600}{50} - (10)^2} = \sqrt{120 - 100} = 4.472$$

$$\text{Coefficient of variation, CV} = \frac{\sigma}{\bar{x}} \times 100 = \frac{4.472}{10} \times 100 = 44.7 \text{ per cent}$$

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean} = 3 \times 12 - 2 \times 10 = 16$$

$$\text{Coefficient of skewness, } Sk_p = \frac{\bar{x} - M_0}{\sigma} = \frac{10 - 16}{4.472} = -1.341$$

(b) Given that $\bar{x} = 100$, CV = 35, $Sk_p = 0.2$.

$$\text{CV} = \frac{\sigma}{\bar{x}} \times 100 \quad \text{or} \quad 35 = \frac{\sigma}{100} \times 100 \text{ or } \sigma = 35$$

$$\text{Also } Sk_p = \frac{\bar{x} - M_0}{\sigma} \text{ or } 0.2 = \frac{100 - M_0}{35} \text{ or } M_0 = 93$$

Mode = 3Me – 2$\bar{x}$ or 93 = 3Me – 2 × 100 or Me = 97.7
Hence Mode is 93 and median is 97.7.

Example 5.5: The data on the profits (in Rs lakh) earned by 60 companies is as follows:

Profits	:	Below 10	10–20	20–30	30–40	40–50	50 and above
No. of Companies	:	5	12	20	16	5	2

(a) Obtain the limits of profits of the central 50 per cent companies
(b) Calculate Bowley's coefficient of skewness.

Solution: (a) Calculations for different quartiles are shown in Table 5.3.

Table 5.3 Computation of Quartiles

Profits (Rs in lakh)	*Frequency (f)*	*Cumulative Frequency (c.f)*
Below 10	5	5
10–20	12	17 ← Q_1 Class
20–30	20	37
30–40	16	53 ← Q_3 Class
40–50	5	58
50 and above	2	60
	N = 60	

Q_1 = size of (N/4)th observation = (60 ÷ 4)th = 15th observation. Thus Q_1 lies in the class 10–20, and

$$Q_1 = l + \left\{\frac{(N/4) - cf}{f}\right\} \times h$$

$$= 10 + \left\{\frac{15 - 5}{12}\right\} \times 10 = 10 + 8.33 = 18.33 \text{ lakh}$$

Q_3 = size of (3N/4)th observation = 45th observation. Thus Q_3 lies in the class 30–40, and

$$Q_3 = l + \left\{\frac{(3N/4) - cf}{f}\right\} \times h$$

$$= 30 + \left\{\frac{45 - 37}{16}\right\} \times 10 = 30 + 5 = 35 \text{ lakh}$$

Hence the profit of central 50 per cent companies lies between Rs 35 lakhs and Rs 10.83 lakh

$$\text{Coefficient of quartile deviation, Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{35 - 18.33}{35 + 18.33} = 0.313$$

(b) Median = size of (N/2)th observation = 30th observation. Thus Median lies in the class 20–30, and

$$\text{Me} = l + \left\{\frac{(N/2) - cf}{f}\right\} \times h = 20 + \left\{\frac{30 - 17}{20}\right\} \times 10 = 20 + 6.5 = 26.5 \text{ lakh}$$

$$\text{Coefficient of skewness, } Sk_b = \frac{Q_3 + Q_1 - 2\text{Me}}{Q_3 - Q_1} = \frac{35 + 18.33 - 2(26.5)}{35 - 18.33} = 0.02$$

The positive value of Sk_b indicates that the distribution is positively skewed and therefore there is a concentration of larger values on the right side of the distribution.

Example 5.6: Apply an appropriate measure of skewness to describe the following frequency distribution.

Age (yrs)	*Number of Employees*	*Age (yrs)*	*Number of Employees*
Below 20	13	35–40	112
20–25	29	40–45	94
25–30	46	45–50	45
30–35	60	50 and above	21

[*Bharthidasan Univ., MBA, 2001*]

Solution: Since given frequency distribution is an open-end distribution, Bowley's method of calculating skewness should be more appropriate. Calculations are shown in Table 5.4.

Table 5.4 Calculations for Bowley's Coefficient of Skewness

Age (yrs)	Number of Employees (f)	Cumulative frequency (cf)
Below 20	13	13
20–25	29	42
25–30	46	88
30–35	60	148 ← Q_1 class
35–40	112	260
40–45	94	354 ← Q_3 class
45–50	45	399
50 and above	21	420
	N = 420	

Q_1 = size of (N/4)th observation = (420/4) = 105th observation.Thus Q_1 lies in the class 30–35, and

$$Q_1 = l + \frac{(N/4) - cf}{f} \times h = 30 + \frac{105 - 88}{60} \times 5 = 30 + 1.42 = 31.42 \text{ years}$$

Q_3 = size of (3N/4)th observation = (3 × 420/4) = 315th observation. Thus Q_3 lies in the class 40–45, and

$$Q_3 = l + \frac{(3N/4) - cf}{f} \times h = 40 + \frac{315 - 260}{94} \times 5 = 40 + 2.93 = 42.93 \text{ years}$$

Median = size of (N/2)th = (420/2) = 210th observation. Thus median lies in the class 35–40, and

$$\text{Me} = l + \frac{(N/2) - cf}{f} \times h = 35 + \frac{210 - 148}{112} \times 5 = 35 + 2.77 = 37.77 \text{ years}$$

$$\text{Coefficient of skewness, } Sk_b = \frac{Q_3 + Q_1 - 2\text{Me}}{Q_3 - Q_1} = \frac{42.93 + 31.42 - 2 \times 37.77}{42.93 - 31.42}$$

$$= -\frac{1.19}{11.51} = -0.103$$

The negative value of Sk_b indicates that the distribution is negatively skewed.

Conceptual Questions 5A

1. Explain the meaning of skewness using sketches of frequency curves. State the different measures of skewness that are commonly used. How does skewness differ from dispersion?
2. Measures of central tendency, dispersion, and skewness are complementary to each other in describing a frequency distribution. Elucidate.

3. Distinguish between Karl Pearson's and Bowley's measure of skewness. Which one of these would you prefer and why? *[Delhi Univ., MBA, 2000]*
4. Define and discuss the 'quartiles' of a distribution. How are they used for measuring dispersion and skewness and point out the various methods of measuring skewness.
5. Explain brieflty the different methods of measuring skewness. *[Kumaon Univ., MBA, 2000]*
6. Define and discuss the 'quartiles' of a distribution. How are they used for measuring variation and skewness.
7. Distinguish between variation and skewness and point out the various methods of measuring skewness.
8. Briefly mention the tests which can be applied to determine the presence of skewness.
9. Explain the term 'skewness'. What purpose does a measure of skewness serve? Comment on some of the well known measures of skewness.

Self-Practice Problems 5A

5.1 The following data relate to the profits (in thousand rupees) of 1,000 companies:

Profits :	100–120	120–140	140–160	160–180	180–200	200–220	220–240
No. of companies :	17	53	199	194	327	208	2

Calculate the coefficient of skewness and comment on its value. *[MD Univ., MBA, 2001]*

5.2 A survey was conducted by a manufacturing company to find out the maximum price at which people would be willing to buy its product. The following table gives the stated price (in rupees) by 100 persons:

Price :	2.80–2.90	2.90–3.00	3.00–3.10	3.10–3.20	3.20–3.30
No. of persons:	11	29	18	27	15

Calculate the coefficient of skewness and interpret its value.

5.3 Calculate coefficient of variation and Karl Pearson's coefficient of skewness from the data given below:

Marks (less than) :	20	40	60	80	100
No. of students :	18	40	70	90	100

5.4 The following table gives the length of the life (in hours) of 400 TV picture tubes:

Length of Life (in hours)	*No. of Picture Tubes*	*Length of Life (in hours)*	*No. of Picture Tubes*
4000–4199	12	5000–5199	55
4200–4399	30	5200–5399	36
4400–4599	65	5400–5599	25
4600–4799	78	5600–5799	9
4800–4999	90		

Compute the mean, standard deviation, and coefficient of skewness.

5.5 Calculate Karl Pearson's coefficient of skewness from the following data:

Profit (Rs in lakh) : Below	20	40	60	80	100
No. of companies :	8	20	50	64	70

5.6 From the following information, calculate Karl Pearson's coefficient of skewness.

Measure	*Place A*	*Place B*
Mean	256.5	240.8
Median	201.0	201.6
S.D.	215.0	181.0

5.7 From the following data calculate Karl Pearson's coefficient of skewness:

Marks (more than) :	0	10	20	30	40	50	60	70	80
No. of students :	150	140	100	80	80	70	30	14	0

5.8 The following information was collected before and after an industrial dispute:

	Before	*After*
No. of workers employed	515	509
Mean wages (Rs)	4900	5200
Median wages (Rs)	5280	5000
Variance of wages (Rs)	121	144

Comment on the gains or losses from the point of view of workers and that of the management.

5.9 Calculate Bowley's coefficient of skewness from the following data

Sales (Rs in lakh) : Below	50	60	70	80	90
No. of companies :	8	20	40	65	80

[Delhi Univ., MBA, 1998, 2003]

5.10 The following table gives the distribution of weekly wages of 500 workers in a factory:

Weekly Wages (Rs)	*No. of Workers*
Below 200	10
200–250	25
250–300	145
300–350	220
350–400	70
400 and above	30

(a) Obtain the limits of income of the central 50 per cent of the observed workers.
(b) Calculate Bowley's coefficient of skewness.

5.11 Find Bowley's coefficient of skewness for the following frequency distribution

No. of children per family :	0	1	2	3	4
No. of families :	7	10	16	25	18

5.12 In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of the upper and the lower quartiles is 100 and the median is 38, find the value of the upper quartile.

5.13 Calculate Bowley's measure of skewness from the following data:

Payment of Commission	*No. of Salesmen*	*Payment of Commission*	*No. of Salesmen*
100–120	4	200–220	80
120–140	10	220–240	32
140–160	16	240–260	23
160–180	29	260–280	17
180–200	52	280–300	7

5.14 Compute the quartiles, median, and Bowley's coefficient of skewness:

Income (in Rs)	*No. of families*
Below 200	25
200 – 400	40
400 – 600	80
600 – 800	75
800 – 1000	20
1000 and above	16

Hints and Answers

5.1 $\bar{x} = A + \frac{\Sigma fd}{N} \times h = 170 + \frac{393}{1000} \times 20 = 177.86$

$Mo = \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h = 180 + \frac{233}{252} \times 20$

$= 190.55$

$\sigma = h\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = 20\sqrt{\frac{1741}{1000} - \left(\frac{393}{1000}\right)^2}$

$= 25.2$

$Sk_p = \frac{\bar{x} - Mo}{\sigma} = \frac{177.86 - 190.55}{25.2} = -0.5035$

5.2 $\bar{x} = A + \frac{\Sigma fd}{N} \times h = 3.05 + \frac{6}{100} \times 0.1 = 3.056$

$Mo = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$

$= 2.9 + \frac{29 - 11}{2 \times 29 - 11 - 18} \times 0.1 = 2.962$

$\sigma = h\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = 0.1\sqrt{\frac{160}{100} - \left(\frac{6}{100}\right)^2}$

$= 0.1264$

$Sk_p = \frac{\bar{x} - Mo}{\sigma} = \frac{3.056 - 2.962}{0.1264} = 0.744$

5.3 $\bar{x} = A + \frac{\Sigma fd}{N} \times h = 50 - \frac{18}{100} \times 20 = 46.4$

$\sigma = h\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = 20\sqrt{\frac{154}{100} - \left(\frac{-18}{100}\right)^2}$

$= 24.56$

$C.V. = \frac{\sigma}{\bar{x}} \times 100 = \frac{24.56}{46.4} \times 100 = 52.93$

$Mo = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h = 39.52;$

$Sk_p = 0.280$

5.4 $\bar{x} = A + \frac{\Sigma fd}{N} \times h = 4899.5 - \frac{108}{400} \times 200 = 4845.5$

$\sigma = h\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = 200\sqrt{\frac{1368}{400} - \left(\frac{-108}{400}\right)^2}$

$= 365.9$

Mo = Mode lies in the class 4800–4999 ; but real class interval is 4799.5–4999.5

$= l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$

$= 4799.5 + \frac{12}{180 - 78 - 55} \times 200 = 4850.56;$

$Sk_p = -0.014$

5.5 $\bar{x} = A + \frac{\Sigma fd}{N} \times h = 50 - \frac{2}{70} \times 20 = 49.43$

$Mo = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h = 40 + \frac{14}{12 + 14} \times 20$

$= 50.76$

$\sigma = h\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = 20\sqrt{\frac{80}{70} - \left(\frac{-2}{70}\right)^2}$

$= 21.64;\ Sk_p = -0.061$

5.6 $a \equiv \bar{x}$ and $z \equiv M_0$;

Place A : Mode $= 3Me - 2\bar{x} = 90;$

$Sk_p = \frac{266.5 - 90}{215} = 0.823$

Place B : Mode $= 3Me - 2\bar{x} = 123.2;$

$Sk_p = \frac{240.8 - 123.2}{181} = 0.649$

5.7

Marks	*No. of Students*
0–10	10
10–20	40
20–30	20
30–40	0
40–50	10
50–60	40
60–70	16
70–80	14

$\bar{x} = 39.27$; $\sigma = 22.81$;

$$Sk_p = \frac{3(\bar{x} - Me)}{\sigma} = \frac{3(39.27 - 45)}{22.81} = -0.754.$$

5.9 Q_1 lies in the class 50–60; $Q_1 = 60$; Q_3 lies in the class 70–80; $Q_3 = 78$

Median (= Q_2) lies in the class 60–70; Me = 70

$$\text{Bowley's coeff. of } Sk_b = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1} = -0.111$$

5.10 $Q_1 = (80/4) = 20$th observation lies in the class 250–300; $Q_1 = 281.03$; $Q_3 = (3\times80)/4 = 60$th observation lies in the class 300–350; $Q_3 = 344.32$

Median lies in the class 300–350, Me = 315.9

Bowley's coeff. of $Sk_b = -0.111$ (negatively skewed distribution)

5.11 $Q_1 = \text{size of } \left(\frac{n+1}{4}\right)\text{th observation} = 24\text{th observation} = 2$

$$Q_3 = \text{size of } \frac{3(n+1)}{4}\text{th observation} = 72\text{th observation} = 4$$

$$Me = \text{size of } \left(\frac{n+1}{2}\right)\text{th observation} = 48\text{th observation; } Sk_b = \frac{4 + 2 - 2(3)}{4 - 2} = 0$$

5.12 Given $Sk_b = 0.6$; $Q_1 + Q_3 = 100$; Me = 38; $Q_3 = ?$

$$Sk_b = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1} \text{ or } 0.6 = \frac{100 - 2(38)}{Q_3 - (100 - Q_3)} \text{ or } Q_3 = 70$$

5.13 $Q_1 = (n/4)$th observation = 67.5th observation lies in class 180–200; $Q_1 = 183.26$

$Q_3 = \left(\frac{3n}{4}\right)$th observation = 202.5th observation lies in class 220–240; $Q_3 = 227.187$

Me = $(n/2)$th observation = 135th observation lies in class 200–220; Me = 206

$$Sk_b = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1} = -0.035$$

5.3 MOMENTS

According to R. A. Fisher, '*A quantity of data which by its mere bulk may be incapable of entering the mind is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.*'

Among these 'relatively few quantities' are those which are known as **moments**. Two of them, the *mean* and the *variance*, have already been discussed and two other higher moments are in common use. The higher moments are basically used to describe the characteristic of populations rather than samples.

Moments represent a convenient and unifying method for summarizing certain descriptive statistical measures.

The measures of central tendency, variability and skewness which have been discussed to describe a frequency distribution, may be classified into two groups:

(*i*) Percentile system, and

(*ii*) Moment system

The *percentile system* includes measures like median, quartiles, decides, percentiles, and so on. The value of these measures represents a given proportion of frequency distribution.

The *moment system* includes measures like mean, average deviation, standard deviation, and so on. The value of these measures is obtained by taking the deviation of individual observations from a given origin. The term '*moment*' is used in physics and refers to the measure of a force which may generate rotation. The possibility of generating such a force depends upon (i) the amount of force needed and (ii) the distance from the origin of the point at which the force is applied. The term moment used in statistics is analogous to the term used in physics, where (i) size of class intervals represents the 'force' and (ii) deviation of mid-value of each class from an observation represents the distance.

While calculating moments, if deviations are taken from the actual mean, then such moments are denoted by the Greek letter μ (mu). On the other hand, if deviations are taken from some assumed mean (or arbitrary value other than zero), then moments are denoted by Greek letter ν (nu) or μ'.

5.3.1 Moments about Mean

Let $x_1, x_2, \ldots, x_n$ be the n observations in a data set with mean $\bar{x}$. Then the rth moment about the actual mean of a variable both for ungrouped and grouped data is given by:

For ungrouped data: $\mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r$; $r = 1, 2, 3, 4.$

For grouped data: $\mu_r = \frac{1}{n}\Sigma f(x-\bar{x})^r$; $r = 1, 2, 3, 4$; $N = \Sigma f_i$

For different values of $r = 1, 2, 3, 4$, different moments can be obtained as shown below:

Ungrouped data: $\mu_1 = \frac{1}{n}\Sigma(x-\bar{x}) = 0; \mu_2 = \frac{1}{n}\Sigma(x-\bar{x})^2 = \sigma^2$ (variance)

$$\mu_3 = \frac{1}{n}\Sigma(x-\bar{x})^3; \quad \mu_4 = \frac{1}{n}\Sigma(x-\bar{x})^4$$

For grouped data: $\mu_1 = \frac{1}{n}\Sigma f(x-\bar{x}); \quad \mu_2 = \frac{1}{n}\Sigma f(x-\bar{x})^2$

$$\mu_3 = \frac{1}{n}\Sigma f(x-\bar{x})^3; \quad \mu_4 = \frac{1}{n}\Sigma f(x-\bar{x})^4$$

The *first moment* μ_1, about origin gives the *mean* and is a measure of central tendency.

$$\mu_1 = \frac{1}{n}\Sigma(x-0) = \frac{1}{n}\Sigma x \quad \leftarrow \text{A.M.}$$

The *second moment* μ_2 about the mean is known as *variance* and is a measure of dispersion.

$$\mu_2 = \frac{1}{n}\Sigma(x-\bar{x})^2 \quad \leftarrow \text{Variance}$$

The *third moment* μ_3 about the mean indicates the symmetry or asymmetry of the distribution; its value is zero for symmetrical distribution.

$$\mu_3 = \frac{1}{n}\Sigma(x-\bar{x})^3$$

The *fourth moment* μ_4 about the mean is a measure of Kurtosis (or flatness) of the frequency curve.

$$\mu_4 = \frac{1}{n}\Sigma(x-\bar{x})^4 \quad \leftarrow \text{Kurtosis}$$

5.3.2 Moments about Arbitrary Point

When actual mean is in fractions, moments are first calculated about an assumed mean, say A, and then are converted about the actual mean, as shown below:

For grouped data: $\mu'_r = \frac{1}{n}\Sigma f(x-\text{A})^r$; $r = 1, 2, 3, 4$

$$= \frac{1}{n}\Sigma f d^2 \times h^2, \text{ where } d = \frac{x-\text{A}}{h} \text{ or } dh = x - \text{A}$$

For ungrouped data: $\mu'_r = \frac{1}{n}\Sigma(x-\text{A})^r$; $r = 1, 2, 3, 4$

For $r = 1$, we have $\mu'_1 = \frac{1}{n}\Sigma(x-\text{A}) = \frac{1}{n}\Sigma x - \text{A} = \bar{x} - \text{A}$

5.3.3 Moments about Zero or Origin

The moments about zero or origin are obtained as follows:

$$\nu_r = \frac{1}{n}\Sigma f x^r; \quad r = 1, 2, 3, 4$$

The relationship among moments about zero and other moments is as follows:

$$\nu_1 = \text{A} + \mu'_1, \qquad \nu_2 = \mu_2 + (\nu_1)^2$$
$$\nu_2 = \mu_3 + 3\nu_1\nu_2 - 2\nu_1^3, \quad \nu_4 = \mu_4 + 4\nu_1\nu_3 - 6\nu_1^2\nu_2 + 3\nu_1^4$$

5.3.4 Relationship Between Central Moments and Moments about any Arbitrary Point

$$\mu_r = \frac{1}{n}\Sigma(x-\bar{x})^r = \frac{1}{n}\Sigma\{x-\text{A}-(\bar{x}-\text{A})\}^r = \frac{1}{n}\Sigma(x-\text{A}-\mu'_1)^r; \quad \mu_1 = \bar{x} - \text{A}$$

$$= \frac{1}{n}\left[\Sigma (x-\mathbf{A})^r - {}^rC_1\, \mu_1' \Sigma(x-\mathbf{A})^{r-1} + {}^rC_2(\mu_1')^2\Sigma(x-\mathbf{A})^{r-2} + \ldots \right.$$

$$\left. + (-1)^r (\mu_1')^r\right]$$

$$= \mu_r' - {}^rC_1\, \mu_1'\mu_{r-1}' + {}^rC_2(\mu_1')^2\, \mu_{r-2}' + \ldots + (-1)^r (\mu_1')^r \qquad (5\text{-}5)$$

From Eqn. (5-5) for various values of r, we have

$\mu_1 = \mu_1'$; $r = 1$ $\qquad$ $\mu_2 = \mu_2' - (\mu_1')^2$; $r = 2$

$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3$; $r = 3$

$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^3$; $r = 4$

5.3.5 Moments in Standard Units

When expressed in standard units moments about the mean of the population are usually denoted by the Greek letter α (alpha). Thus

$$\alpha_r = \frac{1}{n}\Sigma f z^r = \frac{1}{n}\Sigma f\left\{\frac{x-\mu}{\sigma}\right\}^r \text{ by definite of } z = \frac{x-\mu}{\sigma}$$

$$= \frac{1}{\sigma^r}\frac{1}{n}\Sigma f(x-\mu)^r = \frac{\mu_r}{\sigma^2}$$

Hence, $\alpha_1 = 0$, $\alpha_2 = 1$ $\alpha_3 = \frac{\mu_3}{\sigma^3}$ and $\alpha_4 = \frac{\mu_4}{\sigma^4}$ for $r = 1, 2, 3$, and 4 respectively.

In the notations used by Karl Pearson, we have

(i) β_1 (Beta one) $= \alpha_3^2 = \dfrac{\mu_3^2}{\mu_2^3}$

(ii) β_2 (Beta two) $= \alpha_4 = \dfrac{\mu_4}{\mu_2^2}$

In the notations used by R A Fisher, we have

(i) γ_1 (Gamma one) $= \alpha_3 = \sqrt{\beta_1} = \dfrac{\mu_3}{\mu_2^{3/2}} = \dfrac{\mu_3}{\sigma^3}$

(ii) γ_2 (Gamma two) $= \alpha_4 - 3$ or $\beta_2 - 3 = \dfrac{\mu^4}{\mu_2^2} - 3 = \dfrac{\mu^4 - 3\mu_2^2}{\mu_2^2}$

All these coefficients are pure numbers and are independent of whatever unit the variable x may be expressed in. The values of α_3 and α_4 depend on the shape of the frequency curve and, therefore, can be used to distinguish between different shapes. Thus α_3 or β_1 *is a measure of asymmetry about the mean or skewness and is* $\beta_1 = 0$ *for a symmetrical distribution.* A curve with $\alpha_3 > 0$ is said have positive skewness and one with $\alpha_3 < 0$, negative skewness. For most distributions α_3 lies between -3 and 3.

The coefficient of skewness in terms of moments is given by

$$\text{Sk} = \frac{(\beta_2 + 3)\sqrt{\beta_1}}{2(5\beta_2 - 6\beta_1 - 9)}$$

If $\beta_1 = 0$ or $\beta_2 = -3$, then skewness is zero. But $\beta_2 = \frac{\mu_4}{\mu_2^2}$ can not be zero and hence the only condition for skewness to be zero is $\beta_1 = 0$ and the coefficient has no sign.

5.3.6 Sheppard's Corrections for Moments

While calculating higher moments of grouped frequency distributions, it is assumed that frequencies are concentrated at the mid-values of class-intervals. However, it causes certain errors while calculating moments. W E Sheppard proved that, if

(i) the frequency curve of the distribution is continuous,
(ii) the frequency tapers off to zero at both ends, and
(iii) the member of classes are not too large,

then the above assumption that frequencies are concentrated at the mid-value of the class intervals is corrected using Shappard's corrections as follows:

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12}$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h_4$$

where h is the width of the class interval, μ_1 and μ_3 need no correction.

Example 5.7: The first four moments of a distribution about the origin are 1, 4, 10, and 46 respectively. Obtain the various characteristics of the distribution on the basis of the information given. Comment upon the nature of the distribution.

Solution: In the usual notations, we have

$$A = 0,\ \mu_1' = 1,\ \mu_2' = 4,\ \mu_3' = 10 \text{ and } \mu_4' = 46$$

$$\bar{x} = \text{first moment about origin} = \mu_1' = 1$$

$$\text{Variance } (\sigma^2) = \mu_2 = \mu_2' - (\mu_1')^2 = 4 - 1 = 3$$

$$\text{S.D. } (\sigma) = \sqrt{\sigma^2} = \sqrt{3} = 1.732$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 10 - 3(4)(1) + 2(1)^3 = 0.$$

Karl Pearson's coefficient of skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad \text{or} \quad \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Substituting values in the above formula, we get $\gamma_1 = 0$ ($\beta_1 = 0$). This shows that the given distribution is symmetrical, and hence Mean = Median = Mode for the given distribution.

Example 5.8: The first three moments of a distribution about the value 1 of the variable are 2, 25 and 80. Find the mean, standard deviation and the moment-measure of skewness.

Solution: From the data of the problem, we have

$$\mu_1' = 2,\ \mu_2' = 25,\ \mu_3' = 80 \text{ and } A = 1.$$

The moments about the arbitrary point A = 1 are calculated as follows;

$$\text{Mean, } \bar{x} = \mu_1' + A = 2 + 1 = 3$$

$$\text{Variance, } \mu_2 = \mu_2' - (\mu_1')^2 = 25 - (2)^2 = 21$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = 80 - 3(2)(25) + 2(2)^2 = -54$$

$$\text{Standard deviation, } \sigma = \sqrt{\mu_2} = \sqrt{21} = 4.58$$

$$\text{Moment-measure of skewness, } \gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-54}{(21)^{3/2}} = \frac{-54}{96.234} = -0.561$$

Example 5.9: Following is the data on daily earnings (in Rs) of employees in a company:

Earnings	:	50–70	70–90	90–110	110–130	130–150	150–170	170–190
No. of workers	:	4	8	12	20	6	7	3

Calculate the first four moments about the point 120. Convert the results into moments about the mean. Compute the value of γ_1 and γ_2 and comment on the result.

[Delhi Univ., MBA, 1990, 2002]

Solution: Calculations for first four moments are shown in Table 5.5:

Table 5.5 Computation of First Four Moments

Class	Mid-value	Frequency	$d = \frac{m-120}{20}$	fd	fd^2	fd^3	fd^4
50–70	60	4	−3	−12	36	−108	324
70–90	80	8	−2	−16	32	−64	128
90–110	100	12	−1	−12	12	−12	12
110–130	120	20	0	0	0	0	0
130–150	140	6	1	6	6	6	6
150–170	160	7	2	14	28	56	112
170–190	180	3	3	9	27	81	243
		60		−11	141	−41	825

The moments about some arbitrary origin or point (A = 120) is given by

$$\mu_r' = \left(\frac{1}{n}\right)\Sigma f(x - \mathrm{A})^r \text{ (for grouped data)}$$

$$= \frac{1}{n}\left(\Sigma f d^r\right)h^r; \quad d = \frac{m - \mathrm{A}}{h} \text{ or } m - \mathrm{A} = hd$$

For A = 120 and $x = m$, we get

$$\mu_1' = \frac{1}{n}\Sigma fd \times h = \frac{1}{60}(-11) \times 20 = -3.66$$

$$\mu_2' = \frac{1}{n}\Sigma fd^2 \times h^2 = \frac{1}{60}(141) \times (20)^2 = 940$$

$$\mu_3' = \frac{1}{n}\Sigma fd^3 \times h^3 = \frac{1}{60}(-41)\,(20)^3 = -5{,}466.66$$

$$\mu_4' = \frac{1}{n}\Sigma fd^4 \times h^4 = \frac{1}{60}(825)\,(20)^4 = 22{,}00{,}000$$

The moments about actual mean ($\mu_2' = 940$) is given by

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 940 - (-3.66)^2 = 926.55$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = -5{,}466.66 - 3(940)(-3.66) + 2(-3.66)^3$$

$$= -5{,}466.66 + 10{,}340.094 - 98.59 = 4{,}774.83$$

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

$$= 22{,}00{,}000 - 4(-5{,}466.66)(-3.66) + 6(940)(-3.66)^2 - 3(-3.66)^4$$

$$= 22{,}00{,}000 - 80{,}178.50 + 7{,}582.03 - 542.27 = 21{,}95{,}107.20$$

Since μ_3 is positive, therefore the given distribution is positively skewed. The relative measure of skewness is given by

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\mu_2\sqrt{\mu_2}} = \frac{4774.83}{926.55\sqrt{926.55}} = 0.169$$

Thus, $\beta_1 = \gamma_1^2 = 0.0285$. This implies that distribution is positively skewed to the right.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{21{,}95{,}107.20}{(926.56)^2} = 2.56$$

$$\gamma_2 = \beta_2 - 3 = 2.56 - 3 = -0.44$$

Since γ_2 is negative, the distribution is plagtikurtic.

5.4 KURTOSIS

The measure of kurtosis, describes the degree of concentration of frequencies (observations) in a given distribution. That is, whether the observed values are concentrated more around the mode (a peaked curve) or away from the mode towards both tails of the frequency curve.

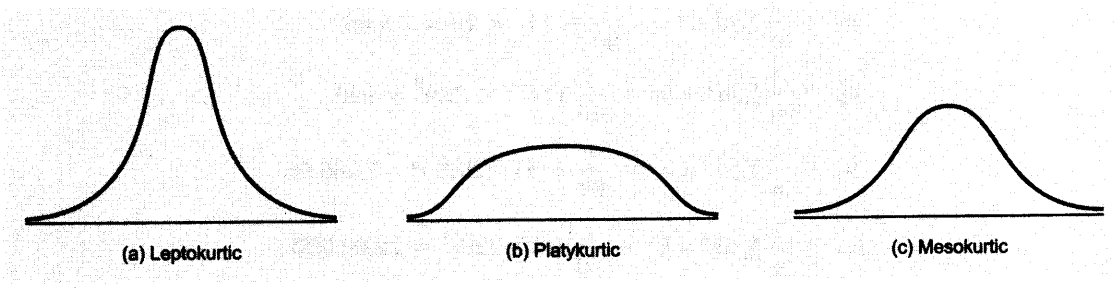
Kurtosis refers to the degree of flatness or peakedness in the region around the mode of a frequency curve.

The word '**kurtosis**' comes from a Greek word meaning 'humped'. In statistics, it refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. A few definitions of kurtosis are as follows:

- *The degree of kurtosis of a distribution is measured relative to the peakedness of a normal curve.* —Simpson and Kafka
- *A measure of kurtosis indicates the degree to which a curve of a frequency distribution is peaked or flat-topped.* —Croxten and Cowden
- *Kurtosis refers to the degree of peakedness of hump of the distribution.* —C. H. Meyers

Two or more distributions may have identical average, variation, and skewness, but they may show different degrees of concentration of values of observations around the mode, and hence may show different degrees of peakedness of the hump of the distributions as shown in Fig. 5.2.

Figure 5.2
Shape of Three Different Curves Introduced by Karl Pearson



5.4.1 Measures of Kurtosis

Leptokurtic refers to a frequency curve that is more peaked than the normal curve.

Platykurtic refers to a frequency curve that is flat-topped than the normal curve.

Mesokurtic refers to a frequency curve that is a normal (symmetrical) curve.

The fourth standardized moment α_4 (or β_2) is a measure of *flatness or peakedness* of a single *humped* distribution (also called *Kurtosis*). For a normal distribution $\alpha_4 = \beta_2 = 3$ so that $\gamma_2 = 0$ and hence any distribution having $\beta_2 > 3$ will be peaked more sharply than the normal curve known as ***leptokurtic*** (narrow) while if $\beta_2 < 3$, the distribution is termed as ***platykurtic*** (broad).

The value of β_2 is helpful in selecting an appropriate measure of central tendency and variation to describe a frequency distribution. For example, if $\beta_2 = 3$, mean is preferred; if $\beta_2 > 3$ (leptokurtic distribution), median is preferred; while for $\beta_2 < 3$ (platykurtic distribution), quartile range is suitable.

Remark: W. S. Gosset, explained different shapes of frequency curves as: Platykurtic curves, like the platypas, are squat with short tails; leptokurtic curves are high with long tails like the Kangaroos noted for leaping.

Example 5.10: The first four moments of a distribution about the value 5 of the variable are 2, 20, 40, and 50. Show that the mean is 7. Also find the other moments, β_1 and β_2, and comment upon the nature of the distribution.

Solution: From the data of the problem, we have

$$\mu_1' = 2,\ \mu_2' = 20,\ \mu_3' = 40,\ \mu_4' = 40 \text{ and } A = 5$$

Now the moments about the arbitrary point 5 are calculated as follows:

$$\text{Mean, } \bar{x} = \mu_1' + A = 2 + 5 = 7$$

$$\text{Variance, } \mu_2 = \mu_2' - (\mu_1')^2 = 20 - (2)^2 = 16$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 = 40 - 3(2)(20) + 2(2)^3 = -64$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$$
$$= 50 - 4(2)(40) + 6(20)(2)^2 - 3(2)^4 = 162$$

The two constants, β_1 and β_2, calculated from central moments are as follows:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-64)^2}{(16)^3} = \frac{4096}{4096} = 1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = \frac{162}{256} = 0.63$$

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{-64}{(16)^{3/2}} = -1\ (< 0),$$ distribution is negatively skewed.

$\gamma_2 = \beta_2 - 3 = 0.63 - 3 = -2.37\ (< 0)$, distribution is platykurtic.

Example 5.11: Find the standard deviation and kurtosis of the following set of data pertaining to kilowatt hours (kwh) of electricity consumed by 100 persons in a city.

Consumption (in kwh) :	0–10	10–20	20–30	30–40	40–50
Number of users :	10	20	40	20	10

Solution: The calculations for standard deviation and kurtosis are shown in Table 5.6.

Table 5.6 Calculations of Standard Deviation and Kurtosis

Consumption (in kwh)	*Number of Users (f)*	*Mid-Value (m)*	*d = (m − A)/10 = (m − 25)/10*	*−fd*	*fd²*
0–10	10	5	−2	−20	40
10–20	20	15	−1	−20	20
20–30	40	(25) ← A	0	0	0
30–40	20	35	1	20	20
40–50	10	45	2	20	40
	100			0	120

$$\bar{x} = A + \frac{\sum fd}{N} \times h = 25 + \frac{0}{100} \times 10 = 25$$

Since $\bar{x} = 25$ is an integer value, therefore we may calculate moments about the actual mean

$$\mu_r = \frac{1}{n}\sum f(x - \bar{x})^r = \frac{1}{n}\sum f(m - \bar{x})^r$$

Let $d = \frac{m - \bar{x}}{h}$ or $(m - \bar{x}) = hd$. Therefore

$$\mu_r = h^r \frac{1}{n}\sum fd^r\ ;\quad h = \text{width of class intervals}$$

The calculations for moments are shown in Table 5.7.

Table 5.7 Calculations for Moments

Mid-value (m)	*Frequency (f)*	$d = \frac{m-25}{10}$	*fd*	*fd²*	*fd³*	*fd⁴*
5	10	−2	−20	40	−80	160
15	20	−1	−20	20	−20	20
(25) ← A	40	0	0	0	0	0
35	20	1	20	20	20	20
45	10	2	20	40	80	160
	100		0	120	0	360

Moments about the origin A = 25 are:

$$\mu_1 = h\frac{1}{N}\Sigma fd = 10 \times \frac{1}{100} = 0$$

$$\mu_2 = h^2\frac{1}{N}\Sigma fd^2 = (10)^2\frac{1}{100}\times 120 = 120$$

$$\mu_3 = h^3\frac{1}{N}\Sigma fd^3 = (10)^3\frac{1}{100}\times 0 = 0$$

$$\mu_4 = h^4\frac{1}{N}\Sigma fd^4 = (10)^4\frac{1}{100}\times 360 = 36{,}000$$

$$\text{S.D. } (\sigma) = \sqrt{\mu_2} = \sqrt{120} = 10.95$$

Karl Pearson's measure of kurtosis is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{36{,}000}{(120)^2} = 2.5$$

and therefore $\gamma_2 = \beta_2 - 3 = 2.5 - 3 = -0.50$

Since $\beta_2 < 3$ (or $\gamma_2 < 0$), distribution curve is platykurtic.

Example 5.12: Calculate the value of γ_1 and γ_2 from the following data and interpret them.

Profit (Rs in lakh) :	10–20	20–30	30–40	40–50	50–60
Number of companies :	18	20	30	22	10

Comment on the skewness and kurtosis of the distribution. [*Kumaon Univ., MBA, 1999*]

Solution: Calculations for moments about an arbitrary constant value are shown in Table 5.8.

Table 5.8 Calculations of Moments

Profit (Rs lakh)	Mid-value (m)	Number of Companies (f)	$d = (m-35)/10$	fd	fd^2	fd^3	fd^4
10–20	15	18	−2	−36	72	−144	288
20–30	25	20	−1	−20	20	−20	20
30–40	(35) ← A	30	0	0	0	0	0
40–50	45	22	1	22	22	22	22
50–60	55	10	2	20	40	80	160
		100		−14	154	−62	490

$$\mu_1' = \frac{\Sigma fd}{N}\times h = \frac{-14}{100} \times 10 = -1.4;$$

$$\mu_2' = \frac{\Sigma fd^2}{N} \times h^2 = \frac{154}{100} \times 100 = 154$$

$$\mu_3' = \frac{\Sigma fd^2}{N} \times h^3 = \frac{-62}{100} \times 1000 = -620;$$

$$\mu_4' = \frac{\Sigma fd^4}{N} \times h^4 = \frac{490}{100} \times 10{,}000 = 49{,}000$$

The central moments are as follows:

$$\mu_2 = \mu_2' - (\mu_1')^2 = 154 - (-1.4)^2 = 152.04$$

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

$$= -620 - 3(-1.4)(154) + 2(-1.4)^3 = 21.312$$

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

$$= 49{,}000 - 4(-1.4)(-620) + 6(154)(-1.4)^2 - 3(-1.4)^3 = 47{,}327.51$$

Karl Pearson's relative measure of skewness and kurtosis are as follows:

$$\text{Measure of skewness, } \gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{21.312}{(152.04)^{3/2}} = \frac{21.312}{1874.714} = 0.0114$$

$$\text{Measure of kurtosis, } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{47{,}327.51}{(152.04)^2} = 2.047$$

$$\gamma_2 = \beta_2 - 3 = 2.047 - 3 = -0.953$$

The value of $\gamma_1 = 0.0114$ suggests that the distribution is almost symmetrical and $\gamma_2 = -0.953$ (< 0) indicates a platykurtic frequency curve.

Conceptual Questions 5B

10. What do you understand by the terms skewness and kurtosis? Point out their role in analysing a frequency distribution. *[Delhi Univ., MBA, 1994]*

11. Averages, dispersion, skewness, and kurtosis are complementary to one another in understanding a frequency distribution? Elucidate.

12. Explain how the measure of skewness and kurtosis can be used in deseribing a frequency distribution. *[Delhi Univ., MBA, 1991]*

13. Define moments. Establish the relationship between the moments about mean and moments about any arbitrary point.

14. Explain the terms 'skewness' and 'kurtosis' used in connection with the frequency distribution of a continuous variable. Give the different measures of skewness (any two of the measures to be given) and kurtosis.

15. What do you mean by 'kurtosis' in statistics? Explain one of the methods of measuring it.

16. What is meant by 'moments' of a frequency distribution? Show how moments are used to describe the characteristics of a distribution, that is, central tendency, dispersion, skewness, and kurtosis. *[Delhi Univ., MBA, 1997]*

17. How do measures of central tendency, dispersion, skewness, and kurtosis help in analysing a frequency distribution? Explain with the help of an example. *[Sukhadia Univ., MBA, 1999]*

18. In what way measures of central tendency, variation, skewness and kurtosis are complementory to one another in understanding a frequency distribution? Elucidate. *[Osmania Univ., MBA, 1995]*

19. A frequency distribution can be described almost completely by the first four moments and two measures based on moments. Examine.

Self-Practice Problems 5B

5.15 The first two moments of a distribution about the value 5 of the variable are 2 and 20. Find the mean and the variance.

5.16 In a certain distribution the first four moments about the point 4 are 15, 17, – 30, and 108 respectively. Find the kurtosis of the frequency curve and comment on its shape.

5.17 Find the first four moments about the mean for the set of numbers 2, 4, 6, and 8.

5.18 Explain whether the following results of a piece of computation for obtaining the second central moment are consistent or not; $n = 120$, $\Sigma fx = -125$, $\Sigma fx^2 = 128$.

5.19 The first four central moments are 0, 4, 8, and 144. Examine the skewness and kurtosis.

5.20 The central moments of a distribution are given by $\mu_2 = 140$, $\mu_3 = 148$, $\mu_4 = 6030$. Calculate the moment measures of skewness and kurtosis and comment on the shape of the distribution.

5.21 Calculate β_1 and β_2 (measure of skewness and kurtosis) for the following frequency distribution and hence comment on the type of the frequency distribution:

x :	2	3	4	5	6
f :	1	3	7	2	1

5.22 Compute the first four moments about the mean from the following data:

Mid-value of variate	5	10	15	20	25	30	35
Frequency	8	15	20	32	23	17	5

Comment upon the nature of the distribution.

5.23 A record was kept over a period of 6 months by a sales manager to determine the average number of calls made per day by his six salesmen. The results are shown below:

Salesmen	A	B	C	D	E	F
Average number of calls per day	8	10	12	15	7	5

(a) Compute a measure of skewness. Is the distribution symmetrical?

(b) Compute a measure of kurtosis. What does this measure mean?

5.24 Find the second, third, and fourth central moments of the frequency distribution given below. Hence find the measure of skewness and a measure of kurtosis of the following distribution:

Class limits	Frequency
100–104.9	7
105–109.9	13
110–114.9	25
115–119.9	25
120–124.9	30

5.25 Find the first four moments about the mean for the following distribution:

Class Interval	60–62	63–65	66–68	69–71	72–74
Frequency	5	18	42	27	8

5.26 Find the variance, skewness, and kurtosis of the following frequency distribution by the method of moments:

Class interval	0–10	10–20	20–30	30–40
Frequency	1	4	3	2

5.27 Find the kurtosis for the following distribution

Class interval	0–10	10–20	20–30	30–40
Frequency	1	3	4	2

Comment on the nature of the distribution.

Hints and Answers

5.15 Given $\mu_1' = 2$, $\mu_2' = 20$, A = 5; $\bar{x} = \mu_1' + A = 7$; $\mu_2 (= \sigma^2) = \mu_2' - (\mu_1')^2 = 16$

5.16 $\mu_1' = 1.5$, $\mu_2' = 17$, $\mu_3' = -30$ and $\mu_4' = 108$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4}{\{\mu_2' - (\mu_1')^2\}^2} = 2.308;$$

distribution is platykurtic.

5.17 $\mu_1 = 0, \mu_2 = 5, \mu_3 = 0$ and $\mu_4 = 41$

5.18 $\mu_2 = \Sigma fx^2/N - (\Sigma fx/N)^2 = 128/120 - (-125/120)^2 = -0.0146$

since σ^2 cannot be negative, therefore the data is inconsistent.

5.19 $\beta_1 = \mu_3^2/\mu_2^3 = 2/(4)^3 = 1$; $\gamma_1 = +\sqrt{\beta_1} = 1$

$\beta_2 = \mu_4/\mu_2^2 = 144/(4)^2 = 9$; $\gamma_2 = \beta_2 - 3 = 6$

5.20 $\beta_1 = \mu_3^2/\mu_2^3 = (148)^2/(140)^3$; $\gamma_1 = +\sqrt{\beta_1} = 0.089$

(Approximately symmetrical and platykurtic)

$\beta_2 = \mu_4/\mu_2^2 = 6030/(140)^2 = 0.3076$

5.21 Let A = 4; $\mu_1' = \Sigma fd/\Sigma f = -0.07$

$\mu_2' = \Sigma fd^2/\Sigma f = 0.92$;

$\mu_3' = \Sigma fd^3/\Sigma f = -0.07$; $\mu_4' = \Sigma fd^4/\Sigma f = 2.64$

$\mu_2 = \mu_2' - (\mu_1')^2 = 0.924$;

$\mu_3 = \mu_3' - 2\mu_2'\mu_1' + 2\mu_1'^3 = 0.123$

$\mu_4' = \mu_4' - 4\mu_3'\mu_1' + \mu_2'\mu_1'^2 - 3\mu_1^4 = 2.691$

$\beta_1 = \mu_3^2/\mu_2^3 = 0.019$; $\beta_2 = \mu_4/\mu_2^2 = 3$ (approx.)

The distribution is approximately normal.

5.23 $\beta_1 = 0.11$; $\beta_2 = 1.97$

5.24 $\mu_2 = 54$; $\mu_3 = 100.5$, $\mu_4 = 7827$;

$\gamma_1 = +\sqrt{\beta_1} = 0.2533$; $\gamma_2 = \beta_2 - 3 = -0.3158$

5.25 $\mu_1 = 0$, $\mu_2 = 8.527$, $\mu_3 = -2.693$, $\mu_4 = 199.375$

5.26 $\sigma^2 = \mu_2 = 84$, $\gamma_1 = +\sqrt{\beta_1} = 0.0935$; $\beta_2 = 2.102$

5.27 $\mu_2 = 81$, $\mu_2 = 14817$; $\beta_2 = \mu_4/\mu_2^2 = 2.26$

Formulae Used

1. **Absolute measure of skewness**

$$\text{Sk} = \bar{x} - \text{Mode or } Q_3 + Q_1 - 2\,\text{Med}$$

2. **Coefficient of skewness**

Karl Pearson's

$$\text{Sk}_p = \frac{\bar{x} - \text{Mode}}{\sigma} \quad \text{or} \quad \frac{3(\bar{x} - \text{Med})}{\sigma}$$

$$\text{Bowley's, Sk}_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$

$$\text{Kelly's, Sk}_k = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}} \quad \text{or} \quad \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$$

3. **Coefficient of skewness based on moments**

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}; \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

4. Moments

About the mean (origin)

$$\mu_r = \frac{1}{n}\sum (x - \bar{x})^r \ , r = 1, 2, 3, 4$$

About an arbitrary point, A

$$\mu'_r = \frac{1}{n}\sum (x - A)^r \ , r = 1, 2, 3, 4$$

5. Kurtosis $\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}}$

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3$$

6. For a normal curve, $\beta_2 = 3$ or $\gamma_2 = 0$; for a leptokurtic curve, $\beta_2 > 3$ or $\gamma_2 > 0$ and for a platykurtic curve, $\beta_2 < 3$ or $\gamma_2 < 0$.

Review Self-Practice Problems

5.28 Calculate the first four moments about the mean and also the value of β_1 and β_2 from the following data:

Marks :

0–10	10–20	20–30	30–40	40–50	50–60	60–70
Number of students :						
8	12	20	30	15	10	5

[Kumaon Univ., MBA, 1998]

5.29 From the following data calculate moments about (a) assumed mean, 25 (b) actual mean, and (c) moments about zero from the following data:

Variable	:	0–10	10–20	20–30	30–40
Frequency	:	1	3	4	2

[MD Univ., BCom, 1997]

5.30 The first four moments of a distribution about $x = 2$ are 1, 2.5, 5.5, and 16. Calculate the four moments about $\bar{x}$ and about zero.

[Delhi Univ., MCom; MD Univ., MCom, 1999]

5.31 The first four central moments of distribution are 0, 2.5, 0.7, and 18.75. Comment on the skewness and kurtosis of the distribution. *[Kanpur Univ., MCom, 1998]*

5.32 Using moments, calculate a measure of relative skewness and a measure of relative kurtosis for the following distribution and comment on the result obtained:

Daily Wages (in Rs)	*No. of Workers*	*Daily Wages (in Rs)*	*No. of Workers*
70 but below 90	8	130 but below 150	9
90 but below 110	11	150 but below 170	4
110 but below 130	18		

[Kerala Univ., BCom, 1998]

5.33 Find the coefficient of skewness from the following information:

Difference of two quartiles = 8; Mode = 1;

Sum of two quartiles = 22; Mean = 8.

[Delhi Univ., BCom (H), 1997]

5.34 From the data given below calculate the coefficient of variation:

Karl pearson's coefficient of skewness = 0.42

Arithmetic mean = 86

Median = 80

[Osmania Univ., BCom, 1998]

5.35 From the following data of the wages of 50 workers of a factory, compute the first four moments about mean and also the value of β_1 and β_2. Comment on the results

Weekly Wages (Rs)	*Number of Workers*	*Weekly Wages (Rs)*	*Number of Workers*
100–120	1	180–200	12
120–140	3	200–220	4
140–160	7	220–240	3
160–180	20		

[Kurukshetra Univ., BCom, 1996]

5.36 In a frequency distribution, the coefficient of skewness based on quartiles is 0.6. If the sum of upper and lower quartiles is 100 and the median is 38, find the value of the upper quartile.

5.37 The following data are given to an economist for the purpose of economic analysis. The data refer to the length of a certain type of battery:

$$n = 100, \quad \sum fd = 50, \quad \sum fd^2 = 1970,$$

$$\sum fd^3 = 2948, \quad \sum fd^4 = 86{,}752$$

where $d = (x - 48)$. Do you think that the distribution is platykurtic? *[Delhi Univ., B.Com (H) 1998]*

5.38 The daily expenditure (in Rs) of 100 families is given below

Daily expenditure :

0–20	20–40	40–60	60–80	80–100
Number of families :				
13	f_2	27	f_4	16

If mode of the distribution is 44, calculate Karl Pearson's coefficient of skewness.

5.39 Pearson's coefficient of skewness for a distribution is 0.4 and coefficient of variance is 30 per cent. Its mode is 88. Find the mean and median.

5.40 Calculate β_1 and β_2 from the frequency distribution and interpret the results.

Age	*Frequency*	*Age*	*Frequency*
25–30	2	45–50	25
30–35	8	50–55	16
35–40	18	55–60	7
40–45	27	60–65	2

[Kumaon Univ., MBA, 2003]

5.41 The following table gives the distribution of monthly wages of 500 workers in a factory:

Monthly Wages (Rs hundred)	*Number of Workers*	*Monthly Wages (Rs hundred)*	*Number of Workers*
15–20	10	30–35	220
20–25	25	35–40	70
25–30	145	40–45	30

Compute Karl Pearson's and Bowley's coefficient of skewness. Interpret your answer.

[Delhi Univ., MBA 2002]

5.42 The first two moments of a distribution about the value 5 of the variable are 2 and 20. Find the mean and variance.

Hints and Answers

5.28 $\mu_1' = -1.8$, $\mu_2' = 240$, $\mu_3' = -1020$,
$\mu_4' = 1{,}44{,}000$
$\mu_2 = 236.76$, $\mu_3 = 264.336$, $\mu_4 = 1{,}41{,}290.11$
$\beta_1 = 0.005$ and $\beta_2 = 2.521$

5.29 (a) Moments about assumed mean, A = 25
$\mu_1' = -3$, $\mu_2' = 90$, $\mu_3' = -900$, $\mu_4' = 21{,}000$.

(b) Moments about actual mean
$\mu_1 = 0$, $\mu_2 = 81$, $\mu_3 = -144$, $\mu_4 = 14{,}817$

(c) Moments about zero:
$\nu_1 = A + \mu_1' = 25 - 3 = 22$ (mean value)
$\nu_2 = \mu_2 + (\nu_1)^2 = 565$; $\nu_3 = \mu_3 + 3\nu_1^2\nu_2 - 2\nu_1^3$
$= 15{,}850$
$\nu_4 = \mu_4 + 4\nu_1\nu_3 - \nu_2\, 6\nu_1^2 + 3\nu_1^4 = 4{,}71{,}625$

5.30 Given $\mu_1' = 1$, $\mu_2' = 2.5$, $\mu_3' = 5.5$ and $\mu_4' = 16$ and A = 2
Moments about mean:
$\mu_2 = \mu_2' - (\mu_1')^2 = 1.5$;
$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 3(\mu_1')^2 = 0$
$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4 = 6$
Moments about zero:
$\nu_1 = A + \mu_1' = 2 + 1 = 3$; $\nu_2 = 10.5$;
$\nu_3 = 40.5$ and $\nu_4 = 168$

5.31 Given $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Measure of skewness, $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = 0.031$ (> 0), the distribution is slightly positively skewed.

Measure of kurtosis, $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = 3$, the distribution is mesokurtic.

5.32 Moments about arbitrary point
$\mu_1' = -2$; $\mu_2' = 136$; $\mu_3' = -680$ and
$\mu_4' = 42{,}400$
Moments about mean:
$\mu_2 = 132$, $\mu_2 = 120$ and $\mu_4 = 40{,}176$
Measure of skewness, $\beta_1 = 0.006$ and measure of kurtosis, $\beta_2 = 2.306$

5.33 Mode = 3 Median – 2 Mean or 11 = 3 Med – 2 × 8 or Med = 9
$Q_3 + Q_1 = 22$ and $Q_3 - Q_1 = 8$, i.e., $Q_3 = 15$, $Q_1 = 7$

$$\text{Coefficient of skewness} = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1}$$
$$= \frac{15 + 7 - 2(9)}{8} = 0.5$$

5.34 Mode = 3 Median – 2 Mean = 3 (80) – 2 (86) = 68

$$\text{Coefficient of skewness} = \frac{\bar{x} - \text{Mode}}{\sigma}$$

or $$0.42 = \frac{86 - 68}{\sigma} \quad \text{or} \quad \sigma = 42.86$$

$$\text{Coefficient of variation (CV)} = \frac{\sigma}{\bar{x}} \times 100$$
$$= \frac{42.86}{68} \times 100$$
$$= 49.84 \text{ per cent.}$$

5.35 Moments about arbitrary mean

$$\mu_1' = \frac{\Sigma fd}{n} \times h = 2.6; \quad \mu_3' = \frac{\Sigma fd^3}{n} \times h^3 = 1340$$

$$\mu_2' = \frac{\Sigma fd^2}{n} \times h^2 = 166;$$

$$\mu_4' = \frac{\Sigma fd^4}{n} \times h^4 = 91{,}000.$$

Moments about mean
$\mu_1 = 0$; $\mu_2 = 159.24$; $\mu_3 = 80.352$,
$\mu_4 = 83{,}659.87$
$\beta_1 = \mu_3^2/\mu_2^3 = 0.0016$
(distribution is almost symmetrical)
$\beta_2 = \mu_4/\mu_2^2 = 3.3$ (> 3), distribution is platykurtic.

5.36 Given Sk = 0.6, $Q_1 + Q_3 = 100$, Med = 38

$$\text{Sk}_b = \frac{Q_3 + Q_1 - 2\text{Med}}{Q_3 - Q_1} \quad \text{or} \quad 0.6 = \frac{100 - 2\times 38}{Q_3 - Q_1}$$
$$= \frac{100 - 76}{Q_3 - (100 - Q_3)} \quad \text{or } Q_3 = 70$$

5.37 $$\mu_2 = \mu_2' - (\mu_1')^2 = \frac{\Sigma fd^2}{n} - \left[\frac{\Sigma fd}{n}\right]^2 = 19.7 - (0.5)^2$$
$$= 19.45$$

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

$$= \frac{\Sigma fd^4}{n} - \frac{4\Sigma fd}{n} \times \frac{\Sigma fd^3}{n} + \left(\frac{\Sigma fd}{n}\right)^2 \frac{\Sigma fd^2}{n} - 3\left(\frac{\Sigma fd}{n}\right)^4$$

$$= 867.52 - 4(0.5)(29.48) + 6(19.7)(0.5) - 3(0.5)^4$$

$$= 837.92$$

$$\therefore \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{837.92}{(19.45)^2} = 2.214 (< 3), \text{ distribution is platykurtic.}$$

5.38 Let the frequency for the class 20–40 be f_2. Then frequency for the class 60–80 will be

$$f_4 = 100 - (13 + f_2 + 27 + 16) = 44 - f_2$$

Expenditure	*Number of Families* (f)	*Cumulative Frequency* (cf)
0– 20	13	13
20– 40	f_2	$13 - f_2$
40– 60	27	$40 - f_2$
60– 80	$44 - f_2$	84
80–100	16	100

$$\text{Mode} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

$$= 40 + \frac{27 - f_2}{54 - f_2 - 44 + f_2} \times 20 \text{ or } f_2 = 25$$

Thus frequency for the class 20–40 is 25 and for the class 60–80 is $44 - 25 = 19$

Apply the formula, $\text{Sk} = \frac{\bar{x} - \text{Mo}}{\sigma} = \frac{50 - 44}{25.3} = 0.237$

5.39 Given Sk = 0.4, CV = 0.30, Mode = 88

$$\text{Sk} = \frac{\bar{x} - \text{Mo}}{\sigma} = \frac{1 - (\text{M}_0/\bar{x})}{(\sigma/\bar{x})} = \frac{1 - (88/\bar{x})}{0.30};$$

$\text{CV} = \sigma/\bar{x}$ or $0.30 = \sigma/\bar{x}$

$$\frac{88}{\bar{x}} = 1 - 0.4 \times 0.3 = 0.88 \quad \text{or} \quad \bar{x} = 100$$

Also, Mode = 3 Med – 2 $\bar{x}$ or $88 = 3$ Med $- 2(100)$ or Med = 96

5.40 $\beta_1 = \mu_3^2/\mu_2^3 = (0.1955)^2/(2.238)^3 = 0.0034;$

$\beta_2 = \mu_4/\mu_2^2 = 12.966/(2.238)^2 = 2.59$

5.42 Given A = 5, $\mu'_1 = 2$, and $\mu'_2 = 20$.

Mean = A + $\mu'_1 = 7$ and variance, $\mu_2 = \mu'_2 - (\mu'_1)^2 = 16$

Chapter 6

. . . is touched by that dark miracle of chance which makes new magic in a dusty world.

—Thomas Wolfe

Fundamentals of Probability

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- help yourself to understand the amount of uncertainty that is involved before making important decisions.
- understand fundamentals of probability and various probability rules that help to you measure uncertainty involving uncertainty.
- perform several analyses with respect to business decision involving uncertainty.

6.1 INTRODUCTION

So far we discussed several methods of summarizing sample data to gain knowledge about the entire population or process. Making inferences from sample data, however, involve *uncertainties*. Thus decision-makers always face some degree of risk while selecting a particular course of action or strategy to solve a decision problem involving uncertainty. It is because each strategy can lead to a number of different possible outcomes (or results). Thus it is necessary for the decision-makers to enhance their capability of grasping the probabilistic situation so as to gain a deeper understanding of the decision problem and base their decisions on rational considerations. The knowledge of the concepts of probability, probability distributions, and various related statistical techniques is therefore needed. The knowledge of probability and its various types of distributions helps in the development of probabilistic decision models. The material in this chapter is designed to

(i) explain the fundamentals of probability and related concepts, and
(ii) illustrate the application of these concepts to decision problems.

6.2 CONCEPTS OF PROBABILITY

In order to obtain a deeper understanding of probability, it is necessary to use certain terms and definitions more precisely. A special type of phenomenon known as *randomness* or *random variation* is of fundamental importance in probability theory. Based upon situations where randomness is present, we can define particular types of occurrences or *events*.

6.2.1 Random Experiment

Random experiment: A process of obtaining information through observation or measurement of a phenomenon whose outcome is subject to chance.

Random experiment (also called *act, trial, operation* or *process*) is an activity that leads to the occurrence of one and only one of several possible outcomes which is not likely to be known until its completion, that is, the outcome is not perfectly predictable. This process has the properties that (i) all possible outcomes can be specified in advance, (ii) it can be repeated, and (iii) the same outcome may not occur on various repetitions so that the actual outcome is not known in advance. The variation among experimental outcomes caused by the effects of uncontrolled factors is called *random variation*. It is assumed that these effects vary randomly and unpredictably from one repetition of an experiment to the next.

The outcome (observation or measurement) generated by an experiment may or may not produce a numerical value. Few examples of experiments are as follows:

(i) Measuring blood pressure of a group of individuals,
(ii) checking an automobile's petrol mileage,
(iii) Tossing a coin and observing the face that appears.
(iv) Testing a product to determine whether it is defective or an acceptable product.
(v) Measuring daily rainfall, and so on.

A simple event: The basic possible outcome of an experiment, it cannot be broken down into simple outcomes.

In all such cases, there is uncertainty surrounding the outcome until an outcome is observed. For example, if we toss a coin, the outcome will not be known with certainty until either the head or the tail is observed. The number of outcomes may be finite or infinite depending on the nature of the experiment. For example, in the experiment of tossing a coin, the outcomes are finite and are represented by the head and tail, whereas in the experiment of measuring the time between successive failures of an electronic device, the outcomes are infinite and are represented by the time of failure.

The outcome of an experiment may be expressed in numerical or non-numerical value. For example,

(i) counting the number of arrivals at a service window (numerical outcome), and
(ii) payment made by cash, cheque, or credit card (non-numerical outcome).

Although an individual outcome associated with a random experiment cannot be predicted exactly, the frequency of occurrence of such an outcome can be noted in a large number of repetitions and thus becomes the basis for resolving problems dealing with uncertainty.

Each experiment may result in one or more outcomes, which are called **events** and denoted by capital letters.

6.2.2 Sample Space

Sample space: The set of all possible outcomes or simple events of an experiment.

The set of all possible distinct outcomes (events) for a random experiment is called the **sample space** (or *event space*) provided.

(i) no two or more of these outcomes can occur simultaneously;
(ii) exactly one of the outcomes must occur, whenever the experiment is performed.

Sample space is denoted by the capital letter S.

Illustrations

1. Consider the experiment of recording a person's blood type. The four possible outcomes are the following simple events:

   E_1 : Blood type A E_2 : Blood type B
   E_3 : Blood type AB E_4 : Blood type O

   The sample space is $S = \{E_1, E_2, E_3, E_4\}$.

   Some experiments can be generated in stages and sample space can be displayed in a *tree diagram*. Each successive level of branching on the tree corresponds to a step required to generate the final outcome as shown in Fig. 6.1. The sample events in the tree diagram form the sample space $S = \{E_1, E_2, E_3, \ldots, E_8\}$.

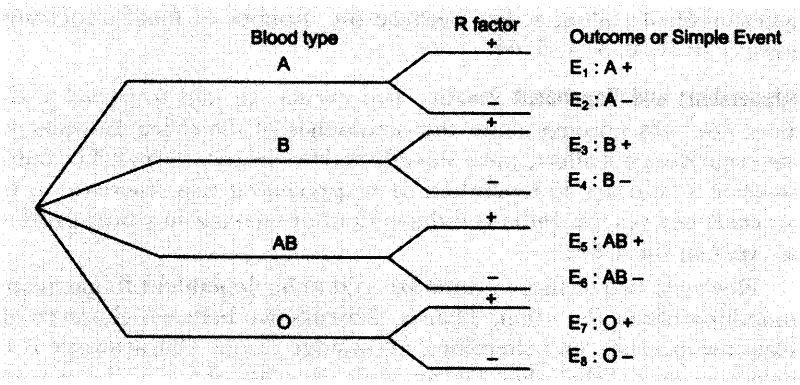


Figure 6.1
Tree Diagram

2. Consider the experiment of tossing two coins. The four possible outcomes are the following sample events

$$E_1 : HH \qquad E_2 = HT \qquad E_3 : TH \qquad E_4 : T\,T$$

The sample space is $S = \{E_1, E_2, E_3, E_4\}$. The sample events can be displayed in a tree diagram shown in Fig. 6.2.

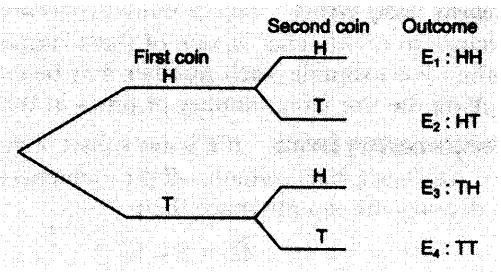


Figure 6.2
Tree Diagram

6.2.3 Event Types

A single possible outcome (or result) of an experiment is called a simple (or elementary) event. An **event** is the set (or collection) of one or more simple events of an experiment in the sample space and having a specific common characteristic. For example, for the above-defined sample space S, the collection (H, T), (T, H) is the event containing simple event as: H or T. Other examples of events are:

Event: Any subset of outcomes of an experiment.

- More than 5 customers at a service facility in one hour
- Telephone calls lasting no more than 10 minutes
- 75 per cent marks or better in an examination
- Sales volume of a retail store more than Rs 2,000 on a given day

Mutually Exclusive Events If two or more events cannot occur simultaneously in a single trial of an experiment, then such events are called mutually exclusive events or disjoint events. In other words, two events are mutually exclusive if the occurrence of one of them prevents or rules out the occurrence of the other. For example, the numbers 2 and 3 cannot occur simultaneously on the roll of a dice.

Mutually exclusive events: Events which cannot occur together or simultaneously.

Symbolically, a set of events $\{A_1, A_2, \ldots, A_n\}$ is mutually exclusive if $A_i \cap A_j = \varnothing\ (i \neq j)$. This means the intersection of two events is a null set ($\varnothing$); it is impossible to observe an event that is common in both A_i and A_j.

Collectively Exhaustive Events A list of events is said to be collectively exhaustive when all possible events that can occur from an experiment includes every possible outcome. That is, two or more events are said to be collectively exhaustive if one of the events must occur. Symbolically, a set of events $\{A_1, A_2, \ldots, A_n\}$ is collectively exhaustive if the union of these events is identical with the sample space S. That is,

Collectively exhaustive events: The list of events that represents all possible experimental outcomes.

$$S = \{A_1 \cup A_2 \cup \ldots \cup A_n\}$$

For example, being a male and female are mutually exclusive and collectively exhaustive events. Similarly, the number 7 cannot come upon the uppermost face during the

experiment of rolling a dice because the number of faces uppermost has the sample space S = {1, 2, 3, 4, 5, 6}.

Independent and Dependent Events Two events are said to be *independent* if information about one tells nothing about the occurrence of the other. In other words, outcome of one event does not affect, and is not affected by, the other event. The outcomes of successive tosses of a coin are independent of its preceding toss. Increase in the population (in per cent) per year in India is independent of increase in wheat production (in per cent) per year in the USA.

However, two or more events are said to be dependent if information about one tells something about the other. That is, dependence between characteristics implies that a relationship exists, and therefore, knowledge of one characteristic is useful in assessing the occurrence of the other. For example, drawing of a card (say a queen) from a pack of playing cards without replacement reduces the chances of drawing a queen in the subsequent draws.

Compound Events When two or more events occur in connection with each other, then their simultaneous occurrence is called a compound event. These event may be (i) independent, or (ii) dependent.

Equally Likely Events Two or more events are said to be equally likely if each has an equal chance to occur. That is, one of them cannot be expected to occur in preference to the other. For example, each number may be expected to occur on the uppermost face of a rolling die the same number of times in the long run.

Complementary Events If E is any subset of the sample space, then its complement denoted by (read as E-bar) contains all the elements of the sample space that are not part of E. If S denotes the sample space then

$$\begin{aligned}\overline{E} &= S - E \\ &= \{\text{All sample elements not in E}\}\end{aligned}$$

For example, if E represents companies with sales less than or equal to Rs 25 lakh, written as $E = \{x : x \leq 25\}$, then this set is a complement of the set, $\overline{E} = \{x : x > 25\}$. Obviously such events must be mutually exclusive and collective exhaustive.

6.3 DEFINITION OF PROBABILITY

Probability: A numerical measure of the likelihood of occurrence of an uncertain event.

A general definition of probability states that **probability** is a numerical measure (between 0 and 1 inclusively) of the likelihood or chance of occurrence of an uncertain event. However, it does not tell us how to compute the probability. In this section, we shall discuss different conceptual approaches of calculating the probability of an event.

6.3.1 Classical Approach

This approach of defining the probability is based on the assumption that all the possible outcomes (finite in number) of an experiment are mutually exclusive and equally likely. It states that, during a random experiment, if there are 'a' possible outcomes where the favouable event A occurs and 'b' possible outcomes where the event A does not occur, and all these possible outcomes are mutually exclusive, exhaustive, and equiprobable, then the probability that event A will occur is defined as

$$P(A) = \frac{a}{a+b} = \frac{\text{Number of favourable outcomes}}{\text{Total number of possible outcomes}} = \frac{c(A)}{c(S)}$$

For example, if a fair die is rolled, then on any trial each event (face or number) is equally likely to occur since there are six equally likely exhaustive events, each will occur 1/6 of the time, and therefore the probability of any one event occurring is 1/6. Similarly for the process of selecting a card at random, each event or card is mutually exclusive, exhaustive, and equiprobable. The probability of selecting any one card on a trial is equal to 1/52, since there are 52 cards. Hence, in general, for a random experiment with

n mutually exclusive, exhaustive, equiprobable events, the probability of any of the events is equal to $1/n$.

Since the probability of occurrence of an event is based on prior knowledge of the process involved, therefore this approach is often called *a priori classical probability approach*. This means, we do not have to perform random experiments to find the probability of occurrence of an event. This also implies that no experimental data are required for computation of probability. Since the assumption of equally likely simple events can rarely be verified with certainty, therefore this approach is not used often other than in games of chance.

The assumption that all possible outcomes are equally likely may lead to a wrong calculation of probability in case some outcomes are more or less frequent in occurrence. For example, if we classify two children in a family according to their sex, then the possible outcomes in terms of number of boys in the family are 0, 1, 2. Thus according to the **classical approach**, the probability for each of the outcomes should be 1/3. However, it has been calculated that the probabilities are approximately 1/4, 1/2, and 1/4 for 0, 1, 2 boys respectively. Similarly, we cannot apply this approach to find the probability of a defective unit being produced by a stable manufacturing process as there are only two possible outcomes, defective or non-defective.

Classical approach: The probability of an event A is the ratio of the number of outcomes in favour of A to the number of all possible outcomes, provided experimental outcomes are equally likely to occur.

6.3.2 Relative Frequency Approach

In situations where the outcomes of a random experiment are not all equally likely or when it is not known whether outcomes are equally likely, application of the classical approach is not desirable to quantify the possible occurrence of a random event. For example, it is not possible to state in advance, without repetitive trials of the experiment, the probabilities in cases like (i) whether a number greater than 3 will appear when die is rolled or (ii) if a lot of 100 items will include 10 defective items.

This approach of computing probability is based on the assumption that a random experiment can be repeated a large number of times under identical conditions where trials are independent to each other. While conducting a random experiment, we may or may not observe the desired event. But as the experiment is repeated many times, that event may occur some proportion of time. Thus, the approach calculates *the proportion of the time (i.e. the **relative frequency**) with which the event occurs over an infinite number of repetitions of the experiment under identical conditions*. Since no experiment can be repeated an infinite number of times, therefore a probability can never be exactly determined. However, we can approximate the probability of an event by recording the relative frequency with which the event has occurred over a finite number of repetitions of the experiment under identical conditions. For example, if a die is tossed n times and s denotes the number of times the event A (i.e., number 4, 5, or 6) occurs, then the ratio $P(A) = c(s)/n$ gives the proportions of times the event A occurs in n trials, and are also called relative frequencies of the event in n trials. Although our estimate about P(A) may change after every trial, yet we will find that the proportion $c(s)/n$ tends to cluster around a unique central value as the number of trials n becomes even larger. This unique central value (also called probability of event A) is defined as:

Relative frequency approach: The probability of an event A is the ratio of the number of times that A has occurred in n trails of an experiment.

$$P(A) = \lim_{n \to \infty} \left\{ \frac{c(s)}{n} \right\}$$

where $c(s)$ represents the number of times that an event s occurs in n trials of an experiment.

Since the probability of an event is determined objectively by repetitive empirical observations of experimental outcomes, it is also known as *empirical probability*. Few situations to which this approach can be applied are follows:

(i) Buying lottery tickets regularly and observing how often you win
(ii) Commuting to work daily and observing whether or not a certain traffic signal is red when cross it.

(iii) Observing births and noting how often the baby is a female
(iv) Surveying many adults and determing what proportion smokes.

Subjective approach: The probability of an event based on the personal beliefs of an individual.

6.3.3 Subjective Approach

The **subjective approach** of calculating probability is always based on the degree of beliefs, convictions, and experience concerning the likelihood of occurrence of a random event. It is thus a way to quantify an individual's beliefs, assessment, and judgment about a random phenomenon. Probability assigned for the occurrence of an event may be based on just guess or on having some idea about the relative frequency of past occurrences of the event. This approach must be used when either sufficient data are not available or sources of information giving different results are not known.

6.3.4 Fundamental Rules of Probability

No matter which approach is used to define probability, the following fundamental rules must be satisfied. Let S be the sample space of an experiment that is partitioned into mutually exclusive and exhaustive events A_1, A_2,..., A_n which may be elementary or compound. The probability of any event A in S is governed by the following rules:

(i) Each probability should fall between 0 and 1, i.e. $0 \leq P(A_i) \leq 1$, for all i, where $P(A_i)$ is read as: 'probability of event A_i'. In otherwords, the probability of an event is restricted to the range *zero* to *one* inclusive, where zero represents an impossible event and one represents a certain event.

For example, probability of the number seven occurring, on rolling a dice, $P(7) = 0$, because this number is an impossible event for this experiment.

(ii) $P(S) = P(A_1) + P(A_2) + \ldots + P(A_n) = 1$, where P(S) is read as: 'probability of the certain event'. This rule states that the sum of probabilities of all simple events constituting the sample space is equal to one. This also implies that if a random experiment is conducted, one of its outcomes in its sample space is certain to occur.

Similarly, the probability of an impossible event or an empty set is zero. That is $P(\Phi) = 0$.

(iii) If events A_1 and A_2 are two elements in S and if occurrence of A_1 implies that A_2 occurs, that is, if A_1 is a subset of A_2, then the probability of A_1 is less than or equal to the probability of A_2. That is, $P(A_1) \leq P(A_2)$.

(iv) $P(\overline{A}) = 1 - P(A)$, that is, the probability of an event that does not occur is equal to one minus the probability of the event that does occur (the probability rule for complementary events).

6.3.5 Glossary of Probability Terms

If A and B are two events, then

$A \cup B$ = an event which represents the occurrence of either A or B or both.
$A \cap B$ = an event which represents the simultaneous occurrence of A and B.
$\overline{A}$ = complement of event A and represents non-occurrence of A.
$\overline{A} \cap \overline{B}$ = both A and B do not occur.
$\overline{A} \cap B$ = event A does not occur but event B occurs.
$A \cap \overline{B}$ = event A occurs but event B does not occur.
$(A \cap \overline{B}) \cup (\overline{A} \cap B)$ = exactly one of the two events A and B occurs.

6.4 COUNTING RULES FOR DETERMINING THE NUMBER OF OUTCOMES

In order to assign probabilities to experimental outcomes it is first necessary to identify and then count them. Following are three important rules for counting the experimental outcomes.