5.17 RELATION BETWEEN MEAN, MEDIAN, MODE

For unimodal distributions of moderate skewness (Section 7.7) the following approximate relation has been found to hold:

$$\text{Mean} - \text{Mode} = 3(\text{Mean}-\text{Median}) \quad (5.17.1)$$

Sometimes this relation may be utilised for the calculation of mode. When the distribution is symmetrical, mean, median and mode coincide. In particular, for the "normal distribution" (Section 12.6) mean, median and mode are all equal.

In most frequency distributions, it has been observed that the three measures of central tendency, viz. Mean, Median and Mode, obey the approximate relation (5.17.1), provided the distribution is not very skew. This relation is therefore applied to estimate one of them when the values of the other two are known.

Example 5.62 *Arithmetic Mean = 26.8, Median = 27.9. What is the value of Mode ? Explain the principle involved.* [C.A. Nov. '72]

Solution Using (5.17.1), $26.8 - \text{Mode} = 3(26.8 - 27.9)$

or, $26.8 - \text{Mode} = -3.3$; $\therefore$ $\text{Mode} = 30.1$

5.18 PARTITION VALUES—QUARTILES, DECILES, PERCENTILES

Just as median divides the total number of observations into two equal parts, there are similar other measures which are used to divide, or partition, the observations into a fixed number of parts, say 4, 10 or 100. These are collectively known as *partition values* or *quantiles* or *fractiles*. Some of the important types of partition values are

(a) Median, (b) Quartiles, (c) Deciles, (d) Percentiles.

We know that Median is the middle-most value of a set of observations, i.e. it divides the total number of observations into 2 equal parts. The number of observations smaller than median is the same as the number larger than it. For data of continuous type, exactly one-half of the observations are smaller than median, i.e. median is the value of the variable corresponding to cumulative frequency $N/2$. These ideas are extended to Quartiles, Deciles and Percentiles,

Quartiles are such values which divide the total number of observations into 4 equal parts. Obviously, there are 3 quartiles—

(i) *First quartile* (or *Lower quartile*) : Q_1
(ii) *Second quartile*, (or *Middle quartile*) : Q_2
(iii) *Third quartile* (or *Upper quartile*) : Q_3

The number of observations smaller than Q_1, is the same as the number lying between Q_1 and Q_2, or between Q_2 and Q_3, or larger than Q_3. For data of continuous type, one-quarter of the observations is smaller than Q_1, two-quarters are smaller than Q_2. and three-quarters are smaller than Q_3. This means that Q_1, Q_2, Q_3 are values of the variable corresponding to 'less-than' cumulative frequencies $N/4$, $2N/4$, $3N/4$ respectively. Since, $2N/4 = N/2$, it is evident that the second quartile Q_2 is the same as median.

$$Q_1 < Q_2 < Q_3\,; \quad Q_3 = \text{Median.} \tag{5.18.1}$$

Quartiles are used for measuring central tendency, dispersion and skewness. For instance, the second quartile Q_2 is itself taken as a measure of central tendency, where it is known as *Median*. The lower and the upper quartiles, viz. Q_1 and Q_3, are used to define *Quartile Deviation* (6.3.1), which is a measure of dispersion. In *Bowley's formula* (7.7.3) for skewness all the three quartiles are used.

$$\text{Median} = Q_2$$

$$\textit{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

$$\textit{Skewness} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

[**Note:** (i) Students must not carry a false impression that there is a common difference between the successive quartiles, or that $Q_3 - Q_2 = Q_2 - Q_1$. Only the number of observations' (i e. frequency) between them is the same. In fact,

Less than Q_1, the frequency = $N/4$
Between Q_1 and Q_2, the frequency = $N/4$
Between Q_2 and Q_3, the frequency = $N/4$
Above Q_3, the frequency = $N/4$

(ii) Again, when asked to define quartiles, many students often write $Q_2 = N/4$, $Q_2 = 2N/4$, $Q_2 = 3N/4$. This is incorrect. The correct position is

Corresponding to Q_1, the cumulative frequency = N/4
Corresponding to Q_2, the cumulative frequency = 2N/4
Corresponding to Q_3, the cumulative frequency = 3N/4]

Deciles are such values which divide the total number of observations into 10 equal parts. There are 9 deciles $D_1, D_2, .., D_9$ called the *first decile*, the *second decile*, etc. The number of observations smaller than D_1, or between two successive deciles, or larger than D_9 is the same. For data of continuous type, $D_1, D_2, .., D_9$ correspond to cumulative frequencies $N/10$, $2N/10$, .. , $9N/10$ respectively.

$$D_1 < D_2 < \ldots < D_9\,; \qquad D_5 = Q_2 = \textit{Median}. \tag{5.18.2}$$

Percentiles are such values which divide the total number of observations into 100 equal parts. There are 99 percentiles $P_1, P_2, .. P_{99}$, called the *first percentile*, the *second percentile*, and so on. The k-th percentile (P_k) is, therefore, that value of the variable upto which lie exactly $k\%$ of the total number of observations. Hence, P_k corresponds to 'less-than' cumulative frequency $kN/100$. In particular,

$$\begin{gathered} P_{10} = D_1, \quad P_{20} = D_2, \ldots, P_{90} = D_9 \\ P_{25} = Q_1, \quad P_{50} = D_5 = Q_2 = \text{Median}, \;\; P_{75} = Q_3 \\ P_1 < P_2 < \ldots < P_{99} \end{gathered} \tag{5.18.3}$$

5.19 CALCULATION OF PARTITION VALUES

The method of calculation of quartiles, deciles, percentiles is exactly the same as that employed for median, using simple interpolation (page 77) in a cumulative frequency distribution. These may also be obtained graphically from ogive. For example, the 7th decile D_7 is the abscissa of that point on the ogive whose ordinate is $7N/10$.

(a) *From simple series*—The given data are arranged in increasing order of magnitude and a number showing the rank is attached to each observation. The smallest value is given rank 1, the next higher value rank 2, etc. and the largest value is given rank n. The ranks of partition values are as follows:

$$\begin{aligned}
\text{Rank of Median} &= \frac{1}{2}(n+1) \\
\text{Rank of } Q_1 &= \frac{1}{4}(n+1) \\
\text{Rank of } Q_3 &= \frac{3}{4}(n+1) \\
\text{Rank of } D_k &= \frac{k}{10}(n+1) \\
\text{Rank of } P_k &= \frac{k}{100}(n+1)
\end{aligned} \tag{5.19.1}$$

Using simple interpolation, the value of the varible corresponding to the appropriate rank is determined, giving the partition value (see Example 5.63)

$$\begin{aligned}
\textit{Median} &= \text{Value corresponding to rank } \frac{1}{2}(n+1) \\
Q_1 &= \text{Value corresponding to rank } \frac{1}{4}(n+1) \\
Q_3 &= \text{Value corresponding to rank } \frac{3}{4}(n+1) \\
D_k &= \text{Value corresponding to rank } \frac{k}{10}(n+1) \\
P_k &= \text{Value corresponding to rank } \frac{k}{100}(n+1)
\end{aligned} \tag{5.19.2}$$

(b) *From simple frequency distribution*—The cumulative frequency corresponding to each distinct value of the variable is calculated (see Table 5.25, p. 142). If the total frequency be N.

$$\begin{aligned}
\textit{Median} &= \text{Value corresponding to cumulative frequency } \frac{1}{2}(N+1) \\
Q_1 &= \text{Value corresponding to cumulative frequency } \frac{1}{4}(N+1) \\
Q_3 &= \text{Value corresponding to cumulative frequency } \frac{3}{4}(N+1) \\
D_k &= \text{Value corresponding to cumulative frequency } \frac{k}{10}(N+1) \\
P_k &= \text{Value corresponding to cumulative frequency } \frac{k}{100}(N+1)
\end{aligned} \tag{5.19.3}$$

(c) *From grouped frequency distribution*—(i) *By appliction of simple interpolation* —A cumulative frequency distribution is constructed showing the class boundaries and the corresponding cumulative frequencies ('less-than'type). Using simple interpolation, we now find (see Examples 5.64, 5.65, 5.66)

$$\begin{aligned}
\textit{Median} &= \text{Value corresponding to cumulative frequency } \frac{1}{2}N \\
Q_1 &= \text{Value corresponding to cumulative frequency } \frac{1}{4}N \\
Q_3 &= \text{Value corresponding to cumulative frequency } \frac{3}{4}N \qquad (5.19.4) \\
D_k &= \text{Value corresponding to cumulative frequency } \frac{k}{10}N \\
P_k &= \text{Value corresponding to cumulative frequency } \frac{k}{100}N
\end{aligned}$$

If the cumulative frequencies are viewed as ranks, some authors define median as the value corresponding to rank $N/2$, $Q1$ corresponding to rank $N/4$, etc. (see Example 5.68).

(ii) *Graphical method*—An ogive ('less-than' type) is drawn. From this ogive (see Example 5.66)

$$\begin{aligned}
\textit{Median} &= \text{Abscissa corresponding to ordinate } \frac{1}{2}N \\
Q_1 &= \text{Abscissa corresponding to ordinate } \frac{1}{4}N \\
Q_3 &= \text{Abscissa corresponding to ordinate } \frac{3}{4}N \qquad (5.19.5) \\
D_k &= \text{Abscissa corresponding to ordinate } \frac{k}{10}N \\
P_k &= \text{Abscissa corresponding to ordinate } \frac{k}{100}N
\end{aligned}$$

Example 5.63 *Obtain the values of Median and the two Quartiles: 391 384 591 407 672 522 777 773 2488 1490*

[C.A. May '69]

Solution The given values are arranged in increasing order of magnitude and ranked (Table 5.31). Here, $n = 10$; so using (5.19.1)

$$\text{Rank of Median} = \frac{1}{2}(10 + 1) = 5.5$$

$$\text{Rank of } Q_1 = \frac{1}{4}(10 + 1) = 2.75$$

$$\text{Rank of } Q_3 = \frac{3}{4}(10 + 1) = 8.25$$

Table 5.31 Ordered Values and Their Ranks

Rank	Value
1	384
2	391
2.75 →	← Q_1
3	407
4	522
5	591
5.5 →	←Median
6	672
7	733
8	777
8.25 →	← Q_3
9	1490
10	2488

Using simple interpolation, we have

$$\frac{5.5-5}{6-5} = \frac{\text{Median} - 591}{672-591}; \quad \frac{2.75-2}{3-2} = \frac{Q_1 - 391}{407-391}$$

$$\frac{8.25-8}{9-8} = \frac{Q_3 - 777}{1490-777}$$

Solving the three equations,

$$\text{Median} = 631.5; \; Q_1 = 403; \; Q_3 = 955.25$$

[**Note:** Since the second quartile Q_2 is the same as Median, for practical purposes we speak of two quartiles only, viz. the lower quartile Q_1, and the upper quartile Q_3].

Example 5.64 *Determine Q_1, Q_3 and the median (Q_2) from the following frequency table:*

Marks in English	10–19	20–29	30–39	40–49	50–59	60–69	Total
Frequency	8	11	15	17	12	7	70

[B.U.,B.A. (Econ) '73]

Solution In a grouped frequency distribution, Q_1 Q_2 and Q_3 are values of the variable corresponding to cumulative frequencies $N/4$, $2N/4$ and $3N/4$ respectively.

Table 5.32 Cumulative Frequency Distribution

Marks (Class boundary)	Cumulative Frequency (less-than)
9.5	0
19.5	8
Q_1→	←$N/4 = 17.5$
29.5	19
39.5	34
Q_2→	←$2N/4 = 35$
49.5	51
Q_3→	← $3N/4 = 52.5$
59.5	63
69.5	$70 = N$

[**Note:** The values of $N/4$, $2N/4$, $3N/4$ are first obtained numerically, and then their positions are indicated by arrow marks in the column of cumulative, frequencies. Q_1, Q_2, Q_3 are then shown in the corresponding spaces under the first column. Here, since $N/4 = 70/4 = 17.5$ exceeds cumulative frequency 8, but is smaller than the next cumulative frequency 19, . $N/4$ is shown in the space between 8 and 19. Q_1 therefore must lie in between the corresponding values of the variable, viz. 19.5 and 29.5 Similarly, $2N/4 = 35$ lies between the cumulative frequencies 34 and 51, and hence the position of Q_2, is shown between the corresponding values 39.5 and 49.5. Similarly for Q_3]

Using simple interpolation (3.16.1), we have

$$\frac{Q_1 - 19.5}{29.5 - 19.5} = \frac{17.5 - 8}{19 - 8}; \quad \frac{Q_2 - 39.5}{49.5 - 39.5} = \frac{35 - 34}{51 - 34}$$

$$\frac{Q_3 - 49.5}{59.5 - 49.5} = \frac{52.5 - 51}{63 - 51}$$

Solving these equations, we get $Q_1 = 28$, $Q_2 = 40$, $Q_3 = 51$

Example 5.65 *From the following data calculate the first quartile and 4th decile :*

X	*0–5*	*5–10*	*10–15*	*15–25*	*25–35*	*35–60*	*60–80*
Freq.	*12*	*30*	*51*	*84*	*66*	*50*	*7*

[C.U., B.Com. (Hons) '69]

Solution (Note that here all classes do not have the same width ; but this does not cause any difficulty in the calculation of partition values). The total frequency is $N = 300$. Hence, using (5.19.4), the first quartile Q_1 and the fourth decile D_4 are those values of X which correspond to cumulative frequencies $\frac{1}{4} \times 300 = 75$ and $\frac{4}{10} \times 300 = 120$ respectively.

Table 5.33 Cumulative Frequency Distribution

Class Boundary	*Cumulative Frequency (less-than)*
0	0
5	12
10	42
$Q_1 \rightarrow$	$\leftarrow N/4 = 75$
15	93
$D_4 \rightarrow$	$\leftarrow 4/N\ 10 = 120$
25	177
35	243
60	293
80	$300 = N$

Using simple interpolation,

$$\frac{Q_1 - 10}{15 - 10} = \frac{75 - 42}{93 - 42}, \text{ hence we get } Q_1 = 13.24$$

$$\frac{D_4 - 15}{25 - 15} = \frac{120 - 93}{177 - 93}, \text{ hence we get } Q_4 = 18.21 \qquad Ans.\ Q_1 = 13.24, D_4 = 18.21$$

7.7 SKEWNESS

A frequency distribution is said to be 'symmetrical', if the frequencies are symmetrically distributed about mean, i.e. when values of the variable equidistant from mean have equal frequencies.

Illustration 1. Symmetrical distributions:

(i)	x :	10	15	20	25	30
	f :	3	7	16	7	3

(ii)	x :	10	15	20	25	30	35
	f :	3	7	16	16	7	3
(iii)	x :	5–8	9–12	13–16	17–20	21–24	
	f :	7	18	23	18	7	
(iv)	x :	5–8	9–12	13–16	17–20	21–24	25–28
	f :	7	18	23	23	18	7

Note that in the above distributions, the means are respectively 20, 22.5, 14.5 and 16.5. The median and the mode for each also have the same value. In fact, for any symmetrical distribution mean, median and mode are equal.

In general, however, frequency distributions are not symmetrical; some are slightly asymmetrical and some others may be highly asymmetrical.

Illustration 2. Asymmetrical (or Skew) distributions:

(i)	x :	5–8	9–12	13–16	17–20	21–24
	f :	7	18	23	16	7
(ii)	x :	5–8	9–12	13–16	17–20	21–24
	f :	7	18	23	10	3

Here the frequencies are not symmetrically distributed; in distribution (i) the extent of asymmetry is small, while in (ii) it is comparatively larger.

The word "*skewness*" is used to denote the 'extent of asymmetry' in the data. When the frequency distribution is not symmetrical, it is said to be '*skew*'. The word 'skewness' literally denotes 'asymmetry, or 'lack of symmetry', and 'skew' denotes 'asymmetrical'. A symmetrical distribution has therefore zero skewness. Skewness may also be positive or negative.

Skewness is measured by the following formulae:

(1) *Pearsons' first measure–*

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} \tag{7.7.1}$$

(2) *Pearson's second measure–*

$$\text{Skewness} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}} \tag{7.7.2}$$

(3) *Bowley's measure–*

$$\text{Skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

$$= \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \tag{7.7.3}$$

where Q_1, Q_2, Q_3 denote the first, second and third quartiles of the distribution.

(4) *Moment measure–*

$$\text{Skewness } (\gamma_1) = \frac{m_3}{\sigma^3} = \frac{m_3}{\left(\sqrt{m_2}\right)^3} \tag{7.7.4}$$

where m_2 and m_3 are the second and third central moments, and σ denotes the S.D. [Note that this is the third moment α_3 of the standardized variable z, i.e. $\alpha_3 = m_3/\sigma^3 = \sqrt{\beta_1} = \gamma_1$; see (7.3.3) and (7.4.3)].

It should be noted that all the measures of skewness are pure numbers (which do not depend upon the units of measurement), and have the value zero when the distribution is symmetrical.

(**Note:** The measures of skewness (7.7.1) to (7.7.4) are also sometimes known as 'coefficients of skewness')

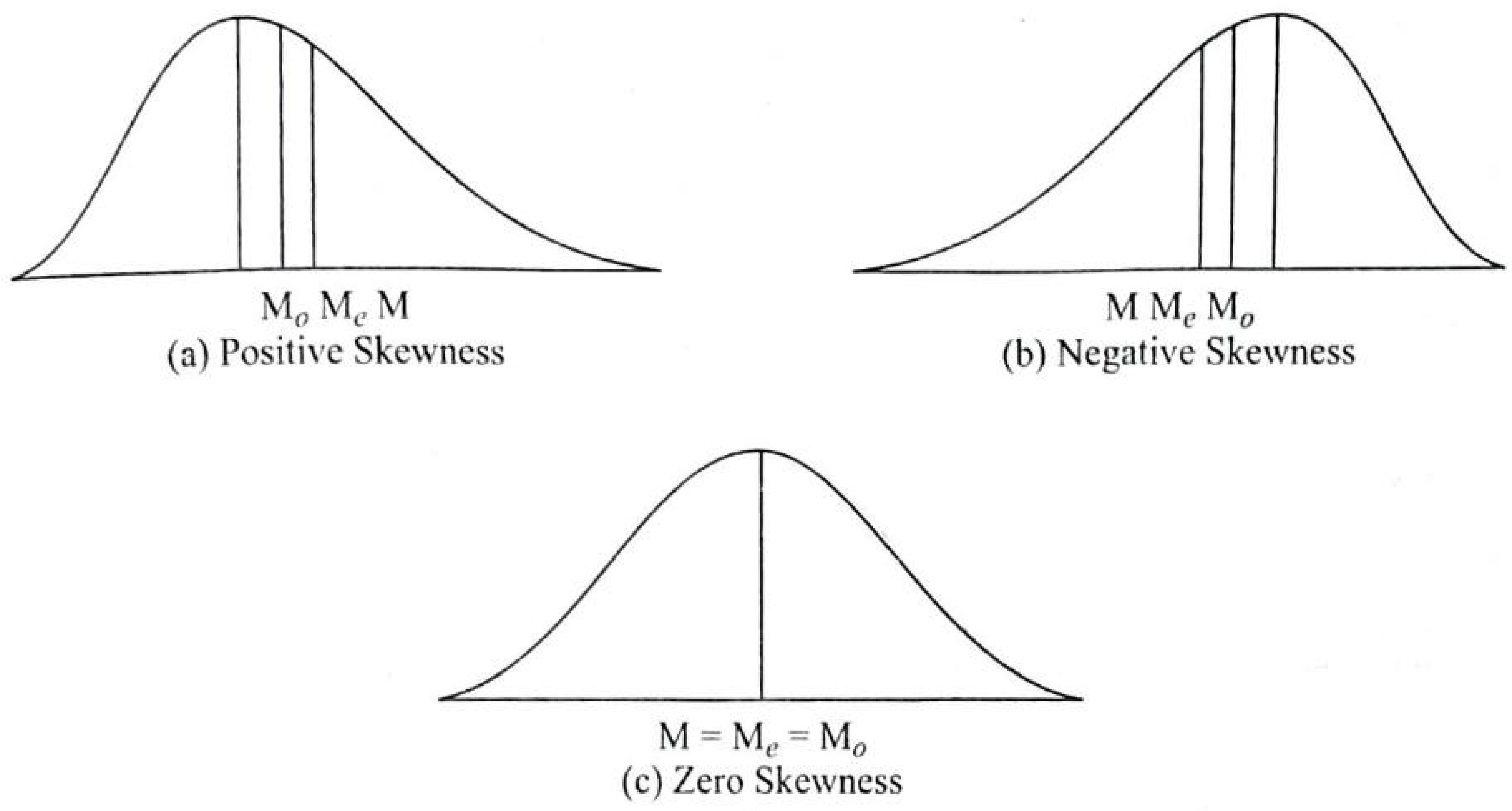

Fig. 7.1 Positions of Mean (M), Median (M_e), Mode(M_o) for Different Types of Skewness

Discussion on different measures—For a symmetrical distribution Mean, Median and Mode are equal. The more the asymmetry in the data, the larger is the discrepancy between them. So the difference between Mean and Mode may be taken as a measure of skewness. This difference, when judged relative to S.D., gives Pearson's measure (7.7.1). If Mode is not known accurately, the approximate relation *Mean – Mode* = 3 (*Mean – Median*) is utilised in (7.7.2). Note that skewness measured by (7.7.1) and (7.7.2) is positive when Mean is larger than Median and Mode, and negative when Mean is less.

Bowley's measure of skewness (7.7.3) has been developed from the following viewpoint. For a symmetrical distribution, Q_2 (i.e. Median) lies exactly midway between Q_1 and Q_3; for a positively skew distribution (i.e. when the longer "tail" of the frequency curve lies towards the right, Fig. 7.1a) Q_3 will be wider away from Q_2 than Q_1, and for a negatively skew distribution (Fig. 7.1b) the reverse will be the case. The difference of the midpoint of Q_1 and Q_3 from the Median Q_2, taken relative to Quartile Deviation, gives (7.7.3).

$$\text{Skewness} = \frac{\frac{1}{2}(Q_1 + Q_3) - Q_2}{\frac{1}{2}(Q_3 - Q_1)} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

As regards the moment measure of skewness (7.7.4), it may be noted that in a symmetrical distribution, for each positive value of $(x_i - \bar{x})$ there is a corresponding negative value. When these deviation $(x_i - \bar{x})$ are cubed, positive values retain their positive sign and negative values the negative sign, so that $m_3 = \Sigma f_i(x_i - \bar{x})^3/N$ will be zero. For positively skew distributions (Fig. 7.1a), large positive values of $(x_i - \bar{x})$ are magnified considerably when cubed, and ultimately the sum of positive cubed deviations outweigh the negative cubes, making m_3 positive. In a similar manner, for negatively skew distributions (Fig. 7.1b), m_3 becomes negative.

Limits of different measures—Bowley's measure (7.7.3) lies between – 1 and + 1. There are no theoretical limits to Pearson's measure (7.7.1); but in practice the value is rarely very high. The measure of skewness (7.7.2) lies between – 3 and + 3, although these limits are very seldom attained. No theoretical limits can be set for the moment measure of skewness (7.7.4).

Example 7.11 *Find a suitable measure of skewness for the following distribution:*

Annual Sales (Rs '000)	*0–20*	*20–50*	*50–100*	*100–250*	*250–500*	*500–1000*
No. of Firms	*20*	*50*	*69*	*30*	*25*	*19*

[CU., M.Com. '72]

Solution Since the class-intervals are of unequal width, Bowley's measure of skewness (7.7.3), which is based on quartiles, will be appropriate here (Note that unequal widths of class intervals do not cause any difficulty in the calculation of partition values; however, other measures of skewness will involve laborious calculations for mean, s.d., moments, etc.). Using simple interpolation in a cumulative frequency distribution (as in Example 5.64), we find that (in Rs '000) $Q_1 = 39.95$, $Q_2 = 76.45$, $Q_3 = 203.75$ (Calculations are not shown here). Putting these values in formula (7.7.3):

$$\text{Skewness} = \frac{203.75 - 2 \times 76.45 + 39.95}{203.75 - 39.95} = \frac{90.80}{163.80} = +0.55$$

Example 7.12 *Calculate the coefficient of skewness based on quartiles from the following:*

More than	0	5474	More than	60	2718
More than	10	5426	More than	70	1406
More than	20	5259	More than	80	764
More than	30	5023	More than	90	370
More than	40	4475	More than	100	160
More than	50	3712	More than	110	39

[C.A., Nov. '74]

Solution The measure of skewness based on quartiles is given by Bowley's formula (7.7.3). We have to find the quartiles Q_1, Q_2, Q_3, by using simple interpolation (page 64) in the following table showing the cumulative frequencies of more-than type.

Table 7.2 Cumulative Frequencies

x	*Cum. Freq. (more-than)*
0	$5474 = N$
10	5426
20	5259
30	5023
40	4475
$Q_1 \rightarrow$	$\leftarrow 3N/4 = 4105.5$
50	3712
$Q_2 \rightarrow$	$\leftarrow 2N/4 = 2737$
60	2718
70	1406
$Q_3 \rightarrow$	$\leftarrow N/4 = 1368.5$
80	764
[illegible]	370
100	160
110	39

Since Q_1, Q_2, Q_3 are values of x corresponding to (less-than) cumulative frequencies $N/4$, $2N/4$, $3N/4$ respectively, they also correspond to more-than cumulative frequencies $3N/4$, $2N/4$, $N/4$.

Using simple interpolation,

$$\frac{Q_1 - 40}{50 - 40} = \frac{4105.5 - 4475}{3712 - 4475}$$

$$\frac{Q_2 - 50}{60 - 50} = \frac{2737 - 3712}{2718 - 3712}$$

$$\frac{Q_3 - 70}{80 - 70} = \frac{1368.5 - 1406}{764 - 1406}$$

Solving, we get $Q_1 = 44.84$, $Q_2 = 59.81$, $Q_3 = 70.58$

Substituting these values in the formula

$$\text{Skewness} = \frac{70.58 - 2 \times 59.81 + 44.84}{70.58 - 44.84} = \frac{-4.20}{25.74} = -0.16$$

Example 7.13 *A frequency distribution gives the following results: (i) Coefficient of variation = 5, (ii) Variance = 4, (iii) Karl Pearson's coefficient of skewness = 0.5. Find the mean and the mode of the distribution.* [I.C.W.A., Dec. '75 & '76-old]

Solution From (ii), S.D. $= \sqrt{(\text{Variance})} = \sqrt{4} = 2$.

Substituting the values of S.D. and C.V. in the formula

$$\text{C.V.} = \frac{\text{S.D.}}{\text{Mean}} \times 100, \quad \text{we get } 5 = \frac{2}{\text{Mean}} \times 100$$

Solving this, Mean = 40. Now, putting the values of Mean and Skewness in formula (7.7.1), viz.