Hints and Answers

3.1 x = Rs 146.30

3.2 Rs 42,000 to pay bonus; Average bonus paid per member = 42,000/25 = Rs 1,680

3.3 $\bar{x}$ = Rs 4892.5 tonne; $\bar{x}_w$ = 5032.30

3.4 $\bar{x}$ = Rs 25,000; $\bar{x}_w$ = Rs 19,262.94

3.6 Percentage of technical personnal = 66.67 per cent ; Non-technical = 33.33 per cent

3.7 Mean marks of grils, $\bar{x}_2$ = 65 per cent

3.9 $\bar{x}_1$ (mean height of n_1 = 30 students) = 5'6".

Given $n_2 = 50 - 30 = 20$, $\bar{x}$ (mean height of 50 students) = 68". Thus

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = 5'\,9''$$

3.10 $\bar{x}$ (average wage) = Rs 17,870;

Given n_1 = 25% lowest paid employee 500/4 = 125

$\bar{x}_1$ = Average salary to these 25 per cent employees = Rs 17,870 + 17,870/4 = Rs 22,337.50

n_2 = 20% highest paid employees = 500/10 = 50

$\bar{x}_2$ = Average salary of these 20% employees = 17,870 + 17,870/10 = Rs 19,657.00

$n_3 = 500 - (125 + 50) = 325$

$\bar{x}_3$ = Average salary to these employees = 17,870 + (15 × 17,870)/100 = Rs 20,550.50

$\therefore \quad \bar{x}_{123}$ = Rs 2,090.90.

3.11 Student 1 : $0.30 \times 55 + 0.15 \times 59 + 0.30 \times 64 + 0.15 \times 20 = 16.5 + 8.85 + 19.2 + 3.0 = 47.55$

2 : $0.20 \times 48 + 0.15 \times 54 + 0.30 \times 58 + 0.15 \times 22 = 14.4 + 8.10 + 17.4 + 3.3 = 43.20$

3: $0.30 \times 64 + 0.15 \times 58 + 0.30 \times 63 + 0.15 \times 19 = 19.2 + 8.7 + 18.9 + 2.85 = 49.65$

4: $0.30 \times 52 + 0.15 \times 49 + 0.30 \times 58 + 0.15 \times 23 = 15.6 + 7.35 + 17.4 + 3.45 = 43.80$

5: $0.30 \times 65 + 0.15 \times 60 + 0.30 \times 62 + 0.15 \times 18 = 19.5 + 9.0 + 18.6 + 2.7 = 49.8$

3.12 $\bar{x}_w = \dfrac{\Sigma x_i w_i}{\Sigma w_i}$

$$= \frac{387.6 \times 7.25 + 158.6 \times 8.20 + 115 \times 7.15}{387.6 + 158.6 + 115}$$

$$= \frac{2810.10 + 1300.52 + 822.25}{661.20}$$

$$= \frac{4932.87}{661.20} = 7.46 \text{ per cent}$$

3.6 GEOMETRIC MEAN

In many business and economics problems, we deal with quantities (variables) that change over a period of time. In such cases the aim is to know an average percentage change rather than simple average value to represent the average growth or decline rate in the variable value over a period of time. Thus we need to calculate another measure of central tendency called **geometric mean** (*G.M.*). The specific application of G.M. is found to show multiplicative effects over time in compound interest and inflation calculations.

Geometric mean: A value that represents nth root of the product of a set of n numbers.

Consider, for example, the annual rate of growth of output of a company in the last five years.

Year	*Growth Rate (Per cent)*	*Output at the End of the Year*
1998	5.0	105
1999	7.5	112.87
2000	2.5	115.69
2001	5.0	121.47
2002	10.0	133.61

The simple arithmetic mean of the growth rate is:

$$\bar{x} = \frac{1}{5}(5 + 7.5 + 2.5 + 5 + 10) = 6$$

This value of 'mean' implies that if 6 per cent is the growth rate, then output at the end of 2002 should be 133.81, which is slightly more than the actual value, 133.61. Thus the correct growth rate should be slightly less than 6.

To find the correct growth rate, we apply the formula of geometic mean:

$$\text{G.M.} = \sqrt[n]{\text{Product of all the } n \text{ values}}$$

$$= \sqrt[n]{x_1 \cdot x_2 \ldots x_n} = (x_1 \cdot x_2 \cdot \ldots x_n)^{\frac{1}{n}} \quad (3\text{-}9)$$

In other words, *G.M. of a set of n observations is the* nth *root of their product.*

For the above example, substituting the values of growth rate in the given formula, we have

$$\text{G.M.} = \sqrt[5]{5 \times 7.5 \times 2.5 \times 5 \times 10} = \sqrt[5]{4687.5}$$

$$= 5.9 \text{ per cent average growth.}$$

Calculation of G.M.

When the number of observations are more than three, the G.M. can be calculated by taking logarithm on both sides of the equation. The formula (3-9) for G.M. for ungrouped data can be expressed in terms of logarithm as shown below:

$$\text{Log (G.M.)} = \frac{1}{n} \log (x_1 \cdot x_2 \ldots \cdot x_n)$$

$$= \frac{1}{n} \{\log x_1 + \log x_2 + \ldots + \log x_n\} = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$

and therefore

$$\text{G.M.} = \text{antilog} \left\{ \frac{1}{n} \Sigma \log x_i \right\} \quad (3\text{-}10)$$

If the observations $x_1, x_2, \ldots, x_n$ occur with frequencies $f_1, f_2, \ldots, f_n$, respectively, and the total of frequencies is $n = \Sigma f_i$, then the G.M. for such data is given by

$$\text{G.M.} = \left(x_1^{f_1} \cdot x_2^{f_2} \cdot \ldots \cdot x_n^{f_n}\right)^{1/n}$$

or

$$\log \text{(G.M.)} = \frac{1}{n} \{f_1 \log x_1 + f_2 \log x_2 + \ldots + f_n \log x_n\}$$

$$= \frac{1}{n} \sum_{i=1}^{n} f_i \log x_i$$

or

$$\text{G.M.} = \text{Antilog} \left\{ \frac{1}{n} \Sigma f_i \log x_i \right\} \quad (3\text{-}11)$$

Example 3.24: The rate of increase in population of a country during the last three decades is 5 per cent, 8 per cent, and 12 per cent. Find the average rate of growth during the last three decades.

Solution: Since the data is given in terms of percentage, therefore geometric mean is a more appropriate measure. The calculations of geometric mean are shown in Table 3.18:

Table 3.18 Calculations of G.M.

Decade	*Rate of Increase in Population (%)*	*Population at the End of Decade (x) Taking Proceeding Decade as 100*	$\log_{10} x$
1	5	105	2.0212
2	8	108	2.0334
3	12	112	2.0492

Using the formula (3-10), we have

$$\text{G.M.} = \text{Antilog}\left\{\frac{1}{n}\Sigma \log x\right\} = \text{Antilog}\left\{\frac{1}{3}(6.1038)\right\}$$

$$= \text{Antilog}\ (2.0346) = 108.2$$

Hence the average rate of increase in population over the last three decades is 108.2 – 100 = 8.2 per cent.

Example 3.25: A given machine is assumed to depreciate 40 per cent in value in the first year, 25 per cent in the second year, and 10 per cent per year for the next three years, each percentage being calculated on the diminishing value. What is the average depreciation recorded on the diminishing value for the period of five years?
Solution: The calculations of geometric mean is shown in Table 3.19.

Table 3.19 Calculations of G.M.

Rate of Depreciation (x_i) (in percentage)	*Number of Years (f_i)*	$\log_{10} x_i$	$f_i \log_{10} x_i$
40	1	1.6021	1.6021
25	1	1.3979	1.3979
10	3	1.0000	3.0000
			6.0000

Using formula (3-11), we have

$$\text{G.M.} = \text{Antilog}\left\{\frac{1}{n}\Sigma f \log x\right\} = \text{Antilog}\left\{\frac{1}{5}(6.0000)\right\}$$

$$= \text{Antilog}\ (1.2) = 15.85$$

Hence, the average rate of depreciation for first five years is 15.85 per cent.

3.6.1 Combined Geometric Mean

The combined geometric mean of observations formed by pooling the geometric means of different sets of data is defined as:

$$\log \text{G.M.} = \frac{\sum_{i=1}^{n} n_i \log G_i}{\sum_{i=1}^{n} n_i} \tag{3-12}$$

where G_i is the geometric mean of the ith data set having n_i number of observations.

3.6.2 Weighted Geometric Mean

If different observations x_i $(i = 1, 2, \ldots, n)$ are given different weights (importance), say w_i $(i = 1, 2, \ldots, n)$ respectively, then their weighted geometric mean is defined as:

$$\text{G.M.}\ (w) = \text{Antilog}\left[\left(\frac{1}{n}\right)\Sigma\, w \log x\right]$$

$$= \text{Antilog}\left[\left(\frac{1}{\Sigma w}\right)\Sigma\, w \log x\right] \tag{3-13}$$

Example 3.26: Three sets of data contain 8, 7, and 5 observations and their geometric means are 8.52, 10.12, and 7.75, respectively. Find the combined geometric mean of 20 observations.

Solution: Applying the formula (3-12), the combined geometric mean can be obtained as follows:

$$\text{G.M.} = \text{Antilog}\left[\frac{n_1 \log G_1 + n_2 \log G_2 + n_3 \log G_3}{n_1 + n_2 + n_3}\right]$$

$$= \text{Antilog}\left[\frac{8 \log (8.52) + 7 \log (10.12) + 5 \log (7.75)}{8 + 7 + 5}\right]$$

$$= \text{Antilog}\left[\frac{(8 \times 0.9304) + (7 \times 1.0051) + (5 \times 0.8893)}{20}\right]$$

$$= \text{Antilog}\left(\frac{18.9254}{20}\right) = \text{Antilog}\ (0.94627) = 8.835$$

Hence, the combined G.M. of 20 observations is 8.835.

Example 3.27: The weighted geometric mean of four numbers 8, 25, 17, and 30 is 15.3. If the weights of the first three numbers are 5, 3, and 4 respectively, find the weight of fourth number.

Solution: Let weight of fourth number be w. Then the weighted geometric mean of four numbers can be calculated as shown in Table 3.20.

Table 3.20 Calculations of Weighted G.M.

Numbers (x)	*Weight of Each Number* (w)	$\log_{10} x$	$w \log_{10} x$
8	5	0.9031	4.5155
25	3	1.3979	4.1937
17	4	1.2304	4.9216
30	w	1.4771	$1.4771w$
	$12 + w$		$13.6308 + 1.4771w$

Thus the weighted G.M. is

$$\log \{\text{G.M. } (w)\} = \left[\left(\frac{1}{\Sigma w}\right)\Sigma\, w \log x\right]$$

or
$$\log (15.3) = \left[\left(\frac{1}{12 + w}\right)(13.6308 + 1.4771w)\right]$$

$$(1.1847)\ (12 + w) = 13.6308 + 1.4771w$$

$$14.2164 + 1.1847w = 13.6308 + 1.4771w$$

$$0.5856 = 0.2924\ w$$

or
$$w = \frac{0.5856}{0.2924} = 2 \text{ (approx.)}$$

Thus the weight of fourth number is 2.

3.6.3 Advantages, Disadvantages, and Applications of G.M.

Advantages

(i) The value of G.M. is not much affected by extreme observations and is computed by taking all the observations into account.

(ii) It is useful for averaging ratio and percentage as well as in determining rate of increase and decrease.

(iii) In the calculation of G.M. more weight is given to smaller values and less weight to higher values. For example, it is useful in the study of price fluctuations where the lower limit can touch zero whereas the upper limit may go upto any number.

(iv) It is suitable for algebraic manipulations. The calculation of weighted G.M. and combined G.M. are two examples of algebraic manipulations of the original formula of geometric mean.

Disadvantages

(i) The calculation of G.M. as compared to A.M., is more difficult and intricate.
(ii) The value of G.M. cannot be calculated when any of the observation in the data set is either negative or zero.
(iii) While calculating weighted geometric, mean equal importance (or weight) is not given to each observation in the data set.

Applications

(i) The concept of G.M. is used in the construction of index numbers.
(ii) Since G.M. $\leq$ A.M., therefore G.M. is useful in those cases where smaller observations are to be given importance. Such cases usually occur in social and economic areas of study.
(iii) The G.M. of a data set is useful in estimating the average rate of growth in the initial value of an observation per unit per period. For example, it is useful in finding the percentage increase in sales, profit, production, population, and so on. It is also useful in calculating the amount of money accumulated at the end of n periods, with an original principal amount of P_0. The formula is as follows:

$$P_n = P_0 (1 + r)^n$$

or

$$r = \left(\frac{P_n}{P_0}\right)^{\frac{1}{n}} - 1$$

where r = interest rate (rate of growth) per unit period
n = number of years or length of the period.

Conceptual Questions 3B

14. Define simple and weighted geometric mean of a given distribution. Under what circumstances would you recommend its use?

15. Discuss the advantages, disadvantages, and uses of geometric mean.

Self-Practice Problems 3B

3.13 Find the geometric mean of the following distribution of data:

Dividend declared (%) :	0–10	10–20	20–30	30–40	40–45
Number of companies :	5	7	15	25	8

3.14 The population of a country was 300 million in 1985. It became 520 million in 1995. Calculate the percentage compounded rate of growth per year.

3.15 Compared to the previous year, the overhead expenses went up by 32 per cent in 1994, increased by 40 per cent in the next year, and by 50 per cent in the following year. Calculate the average rate of increase in overhead expenses over the three years.

3.16 The rise in the price of a certain commodity was 5 per cent in 1995, 8 per cent in 1996, and 77 per cent in 1997. It is said that the average price rise between 1995 and 1997 was 26 per cent and not 30 per cent. Justify the statement and show how you would explain it before a layman.

3.17 The weighted geometric mean of the four numbers 20, 18, 12, and 4 is 11.75. If the weights of the first three numbers are 1, 3, and 4 respectively, find the weight of the fourth number.

3.18 A machinery is assumed to depreciate 44 per cent in value in the first year, 15 per cent in the second year, and 10 per cent per year for the next three years, each percentage being calculated on diminishing value. What is the average percentage of depreciation for the entire period?

3.19 The following data represent the percentage increase in the number of prisoners (a negative number indicates a percentage decrease) in a district jail:

Year	: 1995	1996	1997	1998	1999	2000
Per cent increase	: −2	3	7	4	5	−3

Calculate the average percentage increase using data from 1996–1999 as well as using data for all 6 years.

3.20 A manufacturer of electrical circuit borads, has manufactured the following number of units over the past 5 years:

2000	2001	2002	2003	2004
14,300	15,150	16,110	17,540	19,430

Calculate the average percentage increase in units produced over this time period, and use this to estimate production for 2006.

3.21 The owner of a warehouse is calculating the average growth factor for his warehouse over the last 6 years. Using a geometric mean, he comes up with an answer of 1.42. Individual growth factors for the first 5 years were 1.91, 1.53, 1.32, 1.91, and 1.40, but he lost the records for the sixth year, after he calculated the mean. What was it?

3.22 Industrial Gas Supplier keeps records on the cost of processing a purchase order. Over the last 5 years, this cost has been Rs 355, 358, 361, 365 and 366. What has supplier's average percentage increase been over this period? If this average rate stays the same for 3 more years, what will cost supplier to process a purchase order at that time?

3.23 A sociologist has been studying the yearly changes in the number of convicts assigned to the largest correctional facility in the state. His data are expressed in terms of the percentage increase in the number of prisoners (a negative number indicates a percentage decrease). The sociologist's most recent data are as follows:

1999	2000	2001	2002	2003	2004
5%	6%	9%	4%	7%	6%

(a) Calculate the average percentage increase using only the 1999-2002 data.

(b) A new penal code was passed in 1998. Previously, the prison population grew at a rate of about 2 percent per year. What seems to be the effect of the new penal code?

Hints and Answers

3.13 G.M. = 25.64 per cent.

3.14 Apply the formula: $P_n = P_0 (1 + r)^n$; r is the rate of growth in population. Since $P_{1995} = 520$ and $P_{1985} = 300$ and $n = 10$, therefore $520 = 300 (1 + r)^{10}$ or $r = 3.2$ per cent.

3.15 Apply the formula $P_n = P_0 (1 + r)^3 = P_0 (1 + 0.32) (1 + 0.40) (1 + 0.50)$; P_0 = Overhead expenses in 1994: $(1 + r)^3 = 1.32 \times 1.40 \times 1.50$. Taking log and simplified, we get $r = 40.5$ per cent.

3.16 Average price rise = 26 per cent (G.M.) If we use A.M. and take the price in the base year as 100, then (105 + 108 + 177)/3 = 130 or 30 per cent is the average change per year. Then the price in 1995 would be 130, price in 1996 would be 130 + 30 per cent increase on 130 = 169, and in the year 1997 it would be 169 + 30 per cent increase on 130 = 219.7

3.17 Apply log G.M. $= \dfrac{\Sigma\, w \log w}{\Sigma\, w}$ or log 11.75

$= \dfrac{9.4974 + 0.6021 w_4}{8 + w_4}$ or $w_4 = 0.850$ (approx.).

3.18

Depreciation rate	: 44	15	10	10	10
Diminishing value taking 100 as base (x)	: 56	85	90	90	90

Log x: 1.7582 1.9294 1.9542 1.9542 1.9542 = 9.5502

G.M. = Antilog ($\Sigma \log x/N$) = Antilog (9.5502/5)
= Antilog (1.91004)
= 81.28

The diminishing value is Rs 81.28 and average depreciation is 18.72 per cent.

3.20 G.M.: $\sqrt[4]{19430/14300} = \sqrt[4]{1.3587} = 1.07964$. So the average increase is 7.96 per cent per year.In 2006, the estimated production will be $19430 (1.0796)^2 = 22{,}646$ units (approx.)

3.21 Since G.M: $1.42 = x \times \sqrt[6]{1.91 \times 1.53 \times 1.32 \times 1.91 \times 1.40}$

$x = (1.42)^6/(1.91 \times 1.53 \times 1.32 \times 1.91 \times 1.40)$
$= 8.195/10.988 = 0.7458$

3.22 G.M. $= \sqrt[4]{366/355} = \sqrt[4]{1.0309} = 1.00765$. So the average increase is 0.765 per cent per year. In three more years the estimated cost will be $366 (1.00765)^3$ = Rs 757.600

3.23 (a) G.M: $\sqrt[4]{0.95 \times 1.06 \times 1.09 \times 1.04} = \sqrt[4]{2.5132}$
= 1.03364. So the average rate of increase from 1999-2002 was 3.364 per cent per year.

(b) G.M.: $\sqrt[6]{0.95 \times 1.06 \times 1.09 \times 1.04 \times 1.07 \times 0.94}$
$= \sqrt[6]{1.148156} = 1.01741$. So the new code appears to have slight effect on the rate of growth of convicts, which has decrease from 2 per cent to 1.741 per cent per year.

3.7 HARMONIC MEAN

The **harmonic mean** (H.M.) of a set of observations is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observations, that is,

$$\frac{1}{\text{H.M}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

or

$$\text{H.M.} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)} \qquad \text{(For ungrouped data)} \qquad (3\text{-}14)$$

If $f_1, f_2, \ldots, f_n$ are the frequencies of observations $x_1, x_2, \ldots, x_n$, then the harmonic mean is defined as:

$$\text{H.M.} = \frac{n}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)} \qquad \text{(For grouped data)} \qquad (3\text{-}15)$$

where $n = \sum_{i=1}^{n} f_i$.

Harmonic mean: A value that is the reciprocal of the mean of the reciprocals of a set of numbers.

Example 3.28: An investor buys Rs 20,000 worth of shares of a company each month. During the first 3 months he bought the shares at a price of Rs 120, Rs 160, and Rs 210. After 3 months what is the average price paid by him for the shares?

Solution: Since the value of shares is changing after every one month, therefore the required average price per share is the harmonic mean of the prices paid in first three months.

$$\text{H.M.} = \frac{3}{(1\div 120)+(1\div 160)+(1\div 210)} = \frac{3}{0.008+0.006+0.004}$$
$$= 3/0.018 = \text{Rs } 166.66$$

Example 3.29: Find the harmonic mean of the following distribution of data

Dividend yield (per cent)	:	2–6	6–10	10–14
Number of companies	:	10	12	18

Solution: The calculations of harmonic mean is shown in Table 3.21.

Table 3.21 Calculations of H.M.

Class (Dividend yield)	*Mid-value* (m_i)	*Number of Companies (frequency,* f_i*)*	*Reciprocal* $\left(\frac{1}{m_i}\right)$	$f_i\left(\frac{1}{m_i}\right)$
2–6	4	10	1/4	2.5
6–10	8	12	1/8	1.5
10–14	12	18	1/12	1.5
		N = 40		5.5

The harmonic mean is: $\text{H.M.} = \frac{n}{\sum_{i=1}^{3} f_i\left(\frac{1}{m_i}\right)} = \frac{40}{5.5} = 7.27$

Hence the average dividend yield of 40 companies is 7.27 per cent.

3.7.1 Advantages, Disadvantages, and Applications of H.M.

Advantages

(i) The H.M. of the given data set is also computed based on its every element.
(ii) While calculating H.M., more weightage is given to smaller values in a data set

because in this case the reciprocal of given values is taken for the calculation of H.M.

(iii) The original formula of H.M. can be extended to accommodate further analysis of data by certain algebraic manipulations.

Disadvantages

(i) The H.M. is not often used for analysing business problems.

(ii) The H.M. of any data set cannot be calculated if it has negative and/or zero elements.

(iii) The calculation of H.M. involves complicated calculations. For calculating the H.M. of a data set, the largest weight is given to smaller values of elements, therefore it does not represent the true characteristic of the data set.

Applications

The harmonic mean is particularly useful for computation of average rates and ratios. Such rates and ratios are generally used to express relations between two different types of measuring units that can be expressed reciprocally. For example, distance (in km), and time (in hours).

3.8 RELATIONSHIP BETWEEN A.M., G.M., AND H.M.

For any set of observations, its A.M., G.M., and H.M. are related to each other in the relationship

$$\text{A.M.} \geq \text{G.M.} \geq \text{H.M.}$$

The sign of '=' holds if and only if all the observations are identical.

If observations in a data set take the values $a, ar, ar^2, \ldots, ar^{n-1}$, each with single frequency, then

$$(\text{G.H.})^2 = \text{A.M.} \times \text{H.M.}$$

Self-Practice Problems 3C

3.24 In a certain factory, a unit of work is completed by A in 4 minutes, by B in 5 minutes, by C in 6 minutes, by D in 10 minutes, and by E in 12 minutes (a) What is the average rate of completing the work? (b) What is the average number of units of work completed per minute? (c) At this rate how many units will they complete in a six-hour day?

3.25 An investor buys Rs 12,000 worth of shares of a company each month. During the first 5 months he bought the shares at a price of Rs 100, Rs 120, Rs 150, Rs 200, and Rs 240 per share. After 5 months what is the average price paid by him for the shares?

3.26 Calculate the A.M., G.M., and H.M. of the following observations and show that A.M. > G.M. > H.M.

32 35 36 37 39 41 43

3.27 The profit earned by 18 companies is given below:

Profit (Rs in lakh) :	20	21	22	23	24	25
No. of companies :	4	2	7	1	3	1

Calculate the harmonic mean of profit earned.

3.28 Find the harmonic mean for the following distribution of data:

Class interval :	0–10	10–20	20–30	30–40
Frequency :	5	8	3	4

Hints and Answers

3.24 (a) Average rate of completing the work per minute = 6.25 (b) Average units/minute = $1 \div 6.25 = 0.16$; (c) Units completed in six-hours (360 minutes) day by all 5 workers = $360 \times 0.16 = 288$ units

3.25 Average price paid for shares = Rs 146.30

3.26 A.M. = 37.56; G.M. = 37.52; H.M. = 37.25

3.27 H.M. = Rs 21.9 lakh

3.28 H.M. = 9.09

3.9 AVERAGES OF POSITION

Different from mathematical averages—arithmetic mean, geometric mean, and harmonic mean, which are mathematical in nature and deal with those characteristics of a data set which can be directly measured quantitatively, such as: income, profit, level of production, rate of growth, etc. However, in cases where we want to guard against the influence of a few outlying observations (called outliers), and/or we need to measure qualitative characteristics of a data set, such as: honesty, intelligence, beauty, consumer acceptance, and so on. In all such cases another measures of central tendency namely *median*, *quartiles*, *deciles*, *percentiles*, and *mode* are used. These measures are also called *positional averages*. The term 'position' refers to the place of the value of an observation in the data set. These measures help in identifying the value of an observation of interest rather than computing it.

3.9.1 Median

Median may be defined as the *middle value* in the data set when its elements are arranged in a sequential order, that is, in either ascending or decending order of magnitude. It is called a middle value in an ordered sequence of data in the sense that half of the observations are smaller and half are larger than this value. The **median** is thus a measure of the *location* or *centrality* of the observations.

Median: A measure of central location such that one half of the observations in the data set is less than or equal to the given value.

The median can be calculated for both ungrouped and grouped data sets.

Ungrouped Data

In this case the data is arranged in either ascending or decending order of magnitude.

(i) If the number of observations (n) is an *odd number*, then the median (Med) is represented by the numerical value corresponding to the positioning point of $(n + 1)/2$ ordered observation. That is,

$$\text{Med} = \text{Size or value of } \left(\frac{n+1}{2}\right)\text{th observation in the data array}$$

(ii) If the number of observations (n) is an *even number*, then the median is defined as the arithmetic mean of the numerical values of $n/2$th and $(n/2 + 1)$th observations in the data array. That is,

$$\text{Med} = \frac{\frac{n}{2}\text{th} + \left(\frac{n}{2}+1\right)\text{th}}{2}$$

Example 3.30: Calculate the median of the following data that relates to the service time (in minutes) per customer for 7 customers at a railway reservation counter: 3.5, 4.5, 3, 3.8, 5.0, 5.5, 4

Solution: The data are arranged in ascending order as follows:

Observations in the data array :	1	2	3	4	5	6	7
Service time (in minutes) :	3	3.5	3.8	4	4.5	5	5.5

The median for this data would be

$$\text{Med} = \text{value of } (n + 1)/2 \text{ th observation in the data array}$$
$$= \{(7 + 1) \div 2\}\text{th} = 4\text{th observation in the data array} = 4$$

Thus the median service time is 4 minutes per customer.

Example 3.31: Calculate the median of the following data that relates to the number of patients examined per hour in the outpatient word (OPD) in a hospital: 10, 12, 15, 20, 13, 24, 17, 18

Solution: The data are arranged in ascending order as follows:

Observations in the data array :	1	2	3	4	5	6	7	8
Number of patients :	10	12	13	15	17	18	20	24

Since the number of observations in the data array are even, the average of $(n/2)$th = 4th observation, i.e. 15 and $(n/2) + 1$ = 5th observation, i.e. 17, will give the median, that is,

$$\text{Med} = (15 + 17) \div 2 = 16$$

Thus median number of patients examined per hour in OPD in a hospital are 16.

Grouped Data

To find the median value for grouped data, first identify the class interval which contains the median value or $(n/2)$th observation of the data set. To identify such class interval, find the cumulative frequency of each class until the class for which the cumulative frequency is equal to or greater than the value of $(n/2)$th observation. The value of the median within that class is found by using interpolation. That is, it is assumed that the observation values are evenly spaced over the entire class interval. The following formula is used to determine the median of grouped data:

$$\text{Med} = l + \frac{(n/2) - cf}{f} \times h$$

where l = lower class limit (or boundary) of the median class interval.

$c.f.$ = cumulative frequency of the class prior to the median class interval, that is, the sum of all the class frequencies upto, but not including, the median class interval

f = frequency of the median class

h = width of the median class interval

n = total number of observations in the distribution.

Example 3.32: A survey was conducted to determine the age (in years) of 120 automobiles. The result of such a survey is as follows:

Age of auto	:	0–4	4–8	8–12	12–16	16–20
Number of autos	:	13	29	48	22	8

What is the median age for the autos?

Solution: Finding the cumulative frequencies to locate the median class as shown in Table 3.22.

Table 3.22 Calculations for Median Value

Age of Auto (in years)	*Number of Autos (f)*	*Cumulative Frequency ($c.f$)*	
0– 4	13	13	
4– 8	29	42	
8–12	48	(90)	← Median class
12–16	22	112	
16–20	8	120	
	$n = 120$		

Here the total number of observations (frequencies) are $n = 120$. Median is the size of $(n/2)$th $= 120 \div 2 = 60$th observation in the data set. This observation lies in the class interval 8–12. Applying the formula (3-16), we have

$$\text{Med} = l + \frac{(n/2) - cf}{n} \times h$$

$$= 8 + \frac{(120 \div 2) - 42}{48} \times 4 = 8 + 1.5 = 9.5$$

Example 3.33: In a factory employing 3000 persons, 5 per cent earn less than Rs 150 per day, 580 earn from Rs 151 to Rs 200 per day, 30 per cent earn from Rs 201 to Rs 250 per day, 500 earn from Rs 251 to Rs 300 per day, 20 per cent earn from Rs 301 to Rs 350 per day, and the rest earn Rs 351 or more per day. What is the median wage?

Solution: Calculations for median wage per day are shown in Table 3.23.

Table 3.23 Calculations of Median Wage

Earnings (Rs)	*Percentage of Workers (Per cent)*	*Number of Persons (f)*	*Cumulative Frequency (c.f.)*	
Less than 150	5	150	150	
151–200	—	580	730	
201–250	30	900	(1630)	← Median class
251–300	—	500	2130	
301–350	20	600	2730	
351 and above	—	270	3000	
		$n = 3000$		

Median observation = $(n/2)$th = $(3000) \div 2$ = 1500th observation. This observation lies in the class interval 201–250.

Now applying the formula (3-16), we have

$$\text{Med} = l + \frac{(n/2) - cf}{f} \times h$$

$$= 201 + \frac{1500 - 730}{900} \times 50 = 201 + 42.77 = \text{Rs } 243.77$$

Hence, the median wage is Rs 243.77 per day.

3.9.2 Advantages, Disadvantages, and Applications of Median

Advantages

(i) Median is unique, i.e. like mean, there is only one median for a set of data.
(ii) The value of median is easy to understand and may be calculated from any type of data. The median in many situations can be located simply by inspection.
(iii) The sum of the absolute differences of all observations in the data set from median value is minimum. In other words, the absolute difference of observations from the median is less than from any other value in the distribution. That is, $\Sigma\,|x - \text{Med}|$ = a minimum value.
(iv) The extreme values in the data set does not affect the calculation of the median value and therefore it is the useful measure of central tendency when such values do occur.
(v) The median is considered the best statistical technique for studying the qualitative attribute of a an observation in the data set.
(vi) The median value may be calculated for an open-end distribution of data set.

Disadvantages

(i) The median is not capable of algebraic treatment. For example, the median of two or more sets of data cannot be determined.
(ii) The value of median is affected more by sampling variations, that is, it is affected by the number of observations rather than the values of the observations. Any observation selected at random is just as likely to exceed the median as it is to be exceeded by it
(iii) Since median is an average of position, therefore arranging the data in ascending or descending order of magnitude is time consuming in case of a large number of observations.
(iv) The calculation of median in case of grouped data is based on the assumption that values of observations are evenly spaced over the entire class-interval.

Applications

The median is helpful in understanding the characteristic of a data set when

(i) observations are qualitative in nature

(ii) extreme values are present in the data set
(iii) a quick estimate of an average is desired.

3.10 PARTITION VALUES—QUARTILES, DECILES, AND PERCENTILES

The basic purpose of all the measures of central tendency discussed so far was to know more and more about the characteristic of a data set. Another method to analyse a data set is by arranging all the observations in either ascending or descending order of their magnitude and then dividing this ordered series into two equal parts by applying the concept of median. However, to have more knowledge about the data set, we may decompose it into more parts of equal size. The measures of central tendency which are used for dividing the data into several equal parts are called *partition values*.

In this section, we shall discuss data analysis by dividing it into *four*, *ten*, and *hundred* parts of equal size. Corresponding partition values are called *quartiles*, *deciles*, and *percentiles*. All these values can be determined in the same way as median. The only difference is in their location.

Quartiles: The values which divide an ordered data set into 4 equal parts. The 2nd quartile is the median

Quartiles The values of observations in a data set, when arranged in an ordered sequence, can be divided into four equal parts, or quarters, using three quartiles namely Q_1, Q_2, and Q_3. The first quartile Q_1 divides a distribution in such a way that 25 per cent $(=n/4)$ of observations have a value less than Q_1 and 75 per cent $(= 3n/4)$ have a value more than Q_1, i.e. Q_1 is the median of the ordered values that are below the median.

The second quartile Q_2 has the same number of observations above and below it. It is therefore same as median value.

The quartile Q_3 divides the data set in such a way that 75 per cent of the observations have a value less than Q_3 and 25 per cent have a value more than Q_3, i.e. Q_3 is the median of the order values that are above the median.

The generalized formula for calculating quartiles in case of grouped data is:

$$Q_i = l + \left\{\frac{i\,(n/4) - cf}{f}\right\} \times h; \quad i = 1, 2, 3 \tag{3-17}$$

where cf = cumulative frequency prior to the quartile class interval
l = lower limit of the quartile class interval
f = frequency of the quartile class interval
h = width of the class interval

Deciles: The values which divides an ordered data set into 10 equal parts. The 5th decile is the median.

Deciles The values of observations in a data set when arranged in an ordered sequence can be divided into ten equal parts, using nine deciles, D_i $(i = 1, 2, \ldots, 9)$. The generalized formula for calculating deciles in case of grouped data is:

$$D_i = l + \left\{\frac{i\,(n/10) - cf}{f}\right\} \times h; \quad i = 1, 2, \ldots, 9 \tag{3-18}$$

where the symbols have their usual meaning and interpretation.

Percentiles: The values which divides an ordered data set into 100 equal parts. The 50th percentile is the median.

Percentiles The values of observations in a data when arranged in an ordered sequence can be divided into hundred equal parts using ninety nine percentiles, P_i $(i = 1, 2, \ldots, 99)$. In general, the ith percentile is a number that has $i\%$ of the data values at or below it and $(100 - i)\%$ of the data values at or above it. The lower quartible (Q_1), median and upper quartible (Q_3) are also the 25th percentile, 50 the percentile and 75th percentile, respectively. For example, if you are told that you scored at 90th percentile in a test (like the CAT), it indicates that 90% of the scores were at or below your score, while 10% were at or above your score. The generalized formula for calculating percentiles in case of grouped data is:

$$P_i = l + \left\{\frac{i\,(n/100) - cf}{f}\right\} \times h; \quad i = 1, 2, \ldots, 99 \tag{3-19}$$

where the symbols have their usual meaning and interpretation.

3.10.1 Graphical Method for Calculating Partition Values

The graphical method of determining various partition values can be summarized into following steps:

(i) Draw an ogive (cumulative frequency curve) by 'less than' method.

(ii) Take the values of observations or class intervals along the horizontal scale (i.e. x-axis) and cumulative frequency along vertical scale (i.e., y-axis).

(iii) Determine the median value, that is, value of $(n/2)$th observation, where n is the total number of observations in the data set.

(iv) Locate this value on the y-axis and from this point draw a line parallel to the x-axis meeting the ogive at a point, say P. Draw a perpendicular on x-axis from P and it meets the x-axis at a point, say M.

The other partition values such as quartiles, deciles, and percentiles can also be obtained by drawing lines parallel to the x-axis to the distance $i\,(n/4)$ $(i = 1, 2, 3)$; $i\,(n/10)$ $(i = 1, 2, \ldots, 9)$, and $i\,(n/100)$ $(i = 1, 2, \ldots, 99)$, respectively.

Example 3.34: The following is the distribution of weekly wages of 600 workers in a factory:

Weekly Wages (in Rs)	*Number of Workers*	*Weekly Wages (in Rs)*	*Number of Workers*
Below 875	69	1100 – 1175	58
875 – 950	167	1175 – 1250	24
950 –1025	207	1250 – 1325	10
1025 –1100	65		600

(a) Draw an ogive for the above data and hence obtain the median value. Check it against the calculated value.

(b) Obtain the limits of weekly wages of central 50 per cent of the workers.

(c) Estimate graphically the percentage of workers who earned weekly wages between 950 and 1250. *[Delhi Univ., MBA, 1996]*

Solution: (a) The calculations of median value are shown in Table 3.24.

Table 3.24 Calculations of Median Value

Weekly Wages (in Rs)	*Number of Workers (f)*	*Cumulative Frequency (Less than type)*	*Percent Cumulative Frequency*
Less than 875	69	69	11.50
Less than 950	167	236 ← Q_1 class	39.33
Less than 1025	207	443 ← Median class	73.83
Less than 1100	65	508 ← Q_3 class	84.66
Less than 1175	58	566	94.33
Less than 1250	24	590	98.33
Less than 1325	10	600	100.00

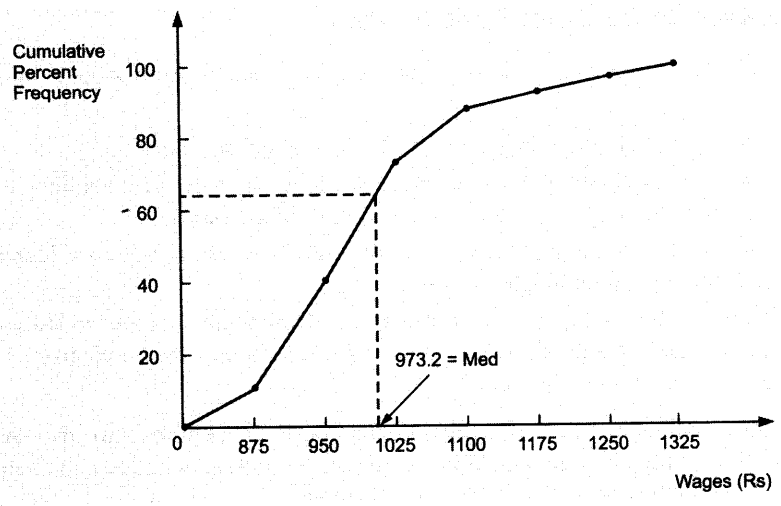
Since a median observation in the data set is the $(n/2)$th observation = $(600 \div 2)$th observation, that is, 300th observation. This observation lies in the class interval 950–1025. Applying the formula (3-16) to calculate median wage value, we have

$$\text{Med} = l + \frac{(n/2) - cf}{f} \times h$$

$$= 950 + \frac{300 - 236}{207} \times 75 = 950 + 23.2 = \text{Rs } 973.2 \text{ per week}$$

The median wage value can also be obtained by applying the graphical method as shown in Fig. 3.1.

Fig. 3.1
Cumulative Frequency Curve



Q_1 = value of $(n/4)$th observation
= value of (600/4)th = 150th observation

(b) The limits of weekly wages of central 50 per cent of the workers can be calculated by taking the difference of Q_1 and Q_3. This implies that Q_1 lies in the class interval 875–950. Thus

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

$$= 875 + \frac{150 - 69}{167} \times 75 = 875 + 36.38 = \text{Rs } 911.38 \text{ per week}$$

Similarly, Q_3 = Value of $(3n/4)$th observation
= Value of $(3 \times 600/4)$th = 450th observation

This value of Q_3 lies in the class interval 1025–1100. Thus

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

$$= 1025 + \frac{450 - 443}{65} \times 75 = 1025 + 8.08 = \text{Rs } 1033.08 \text{ per week}$$

Hence the limits of weekly wages of central 50 per cent workers are Rs 411.38 and Rs 533.08.

(c) The percentage of workers who earned weekly wages less than or equal to Rs 950 is 39.33 and who earned weekly wages less than or equal to Rs 1250 is 98.33. Thus the percentage of workers who earned weekly wages between Rs 950 and Rs 1250 is (98.33–39.33) = 59.

Example 3.35: You are working for the transport manager of a 'call centre' which hires cars for the staff. You are interested in the weekly distances covered by these cars. Kilometers recorded for a sample of hired cars during a given week yielded the following data:

Kilometers Covered	*Number of Cars*	*Kilometers Covered*	*Number of Cars*
100–110	4	150–160	8
110–120	0	160–170	5
120–130	3	170–180	0
130–140	7	180–190	2
140–150	11		40

(a) Form a cumulative frequency distribution and draw a cumulative frequency ogive.
(b) Estimate graphically the number of cars which covered less than 165 km in the week.
(c) Calculate Q_1, Q_2, Q_3 and P_{75}.

Solution: (a) The calculations to a cumulative frequency distribution and to draw ogive are shown in table 3.25.

Table 3.25

Kilometers Covered Less than	*Number of Cars*	*Cumulative Frequency*	*Percent Cumulative Frequency*
110	4	4	10.0
120	0	4	10.0
130	3	7	17.5
140	7	14 $\leftarrow Q_1$	35.0
150	11	25 $\leftarrow$ Me $= Q_2$	62.5
160	8	33 $\leftarrow Q_3$	82.5 $\leftarrow P_{75}$
170	5	38	95.0
180	0	38	95.0
190	2	40	100.0

Plotting cumulative frequency values on the graph paper, frequency polygon is as shown in Fig. 3.2.

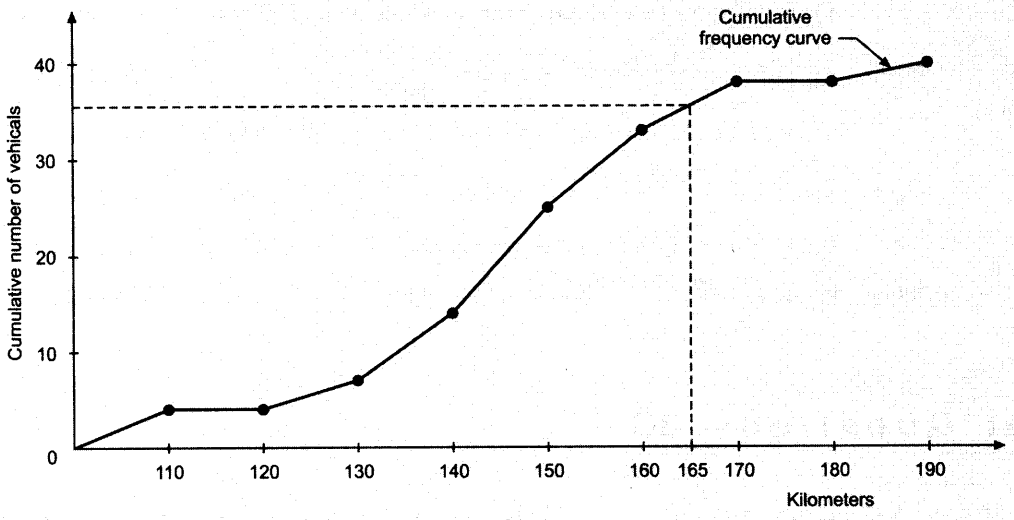


Fig. 3.2
Cumulative Frequency Curve

(b) The number of cars which covered less than 165 km in the week are 35 as shown in the Fig. 3.2.

(c) Since there are 40 observations in the data set, we can take 10th, 20th and 30th cumulative values to corresponds to Q_1, Q_2 and Q_3 respectively. These values from the graph gives $Q_1 = 134$, $Q_2 = 146$ and $Q_3 = 156$.

$$P_{75} = \frac{(75n/100) - cf}{f} \times h = 150 + \frac{30 - 25}{8} \times 10 = 156.25$$

This implies that 75 per cent of cars covered less than or equal to 156.25 kilometers.

Example 3.36: The following distribution gives the pattern of overtime work per week done by 100 employees of a company. Calculate median, first quartile, and seventh decile.

Overtime hours	:	10–15	15–20	20–25	25–30	30–35	35–40
No. of employees	:	11	20	35	20	8	6

[*Kurukshetra Univ., MBA, 1997*]

Calculate Q_1, D_7 and P_{60}.

Solution: The calculations of median, first quartile (Q_1), and seventh decile (D_7) are shown in Table 3.26.

Table 3.26

Overtime Hours	*Number of Employees*	*Cumulative Frequency (Less than type)*
10–15	11	11
15–20	20	31 ← Q_1 class
20–25	35	66 ← Median and P_{60} class
25–30	20	86 ← D_7 class
30–35	8	94
35–40	6	100
	100	

Since the number of observations in this data set are 100, the median value is $(n/2)$th $= (100 \div 2)$th $= 50$th observation. This observation lies in the class interval 20–25. Applying the formula (3-16) to get median overtime hours value, we have

$$\text{Med} = l + \frac{(n/2) - cf}{f} \times h$$

$$= 20 + \frac{50 - 31}{35} \times 5 = 20 + 2.714 = 22.714 \text{ hours}$$

Q_1 = value of $(n/4)$ th observation = value of (100/4)th = 25th observation

Thus $Q_1 = l + \frac{(n/4) - cf}{f} \times h = 15 + \frac{25 - 11}{20} \times 5 = 15 + 3.5 = 18.5$ hours

D_7 = value of $(7n/10)$th observation = value of $(7 \times 100)/10 = 70$th observation

Thus $D_7 = l + \frac{(7n/10) - cf}{f} \times h = 25 + \frac{70 - 66}{20} \times 5 = 25 + 1 = 26$ hours

P_{60} = Value of $(60n/100)$ th observation $= 60 \times (100/100) = 60$th observation

Thus $P_{60} = l + \frac{(60 \times n/100) - cf}{f} \times h = 20 + \frac{60 - 31}{35} \times 5 = 24.14$ hours

Conceptual Questions 3C

16. Define median and discuss its advantages and disadvantages.

17. Why is it necessary to interpolate in order to find the median of a grouped data?

18. When is the use of median is considered more appropriate than mean?

19. Write a short criticism of the following statement: 'Median is more representative than mean because it is relatively less affected by extreme values'.

20. What are quartiles of a distribution? Explain their uses.

21. It has been said that the same percentage of frequencies falls between the first and ninth decile for symmetric and skewed distributions. Criticize or explain this statement. Generalize your answer to other percentiles.

22. Describe the similarities and differences among median, quartiles, and percentiles as descriptive measures of position.

23. You obtained the following answers to a statement while conducting a survey on reservation for women in politics strongly disagree, disagree, mildly disagree, agree some what, agree, strongly agree. Of these answers, which is the median?

Self-Practice Problems 3D

3.29 On a university campus 200 teachers are asked to express their views on how they feel about the performance of their Union's president. The views are classified into the following categories:

Disapprove strongly	= 94
Disapprove	= 52
Approve	= 43
Approve strongly	= 11

What is the median view?

3.30 The following are the profit figures earned by 50 companies in the country

Profit (Rs in lakh)	*Number of Companies*
10 or less	4
20 or less	10
30 or less	30
40 or less	40
50 or less	47
60 or less	50

Calculate

(a) the median, and

(b) the range of profit earned by the middle 80 per cent of the companies. Also verify your results by graphical method.

3.31 A number of particular items has been classified according to their weights. After drying for two weeks the same items have again been weighted and similarly classified. It is known that the median weight in the first weighting was 20.83 g, while in the second weighing it was 17.35 g. Some frequencies a and b in the first weighing and x and y in the second weighing are missing. It is known that $a = x/3$ and $b = y/2$. Find out the values of the missing frequencies.

Class	*Frequencies*		*Class*	*Frequencies*	
	I	*II*		*I*	*II*
0–5	a	x	15–20	52	50
5–10	b	y	20–25	75	30
10–15	11	40	25–30	22	28

3.32 The length of time taken by each of 18 workers to complete a specific job was observed to be the following:

Time (in min) :	5–9	10–14	15–19	20–24	25–29
Number of workers :	3	8	4	2	1

(a) Calculate the median time

(b) Calculate Q_1 and Q_3

3.33 The distribution of the insurance money paid by an automobile insurance company to owners of automobiles in a particular year is given below:

Amount Paid (in Rs)	*Frequency*	*Amount Paid (in Rs)*	*Frequency*
below 1500	52	3500–3499	816
1500–1999	108	4000–4499	993
2000–2499	230	4500–4999	825
2500–2999	528	5000 and above	650
3000–3499	663		

Calculate the median amount of money paid.

3.34 The following distribution is with regard to weight (in g) of mangoes of a given variety. If mangoes less than 443 g in weight be considered unsuitable for the foreign market, what is the percentage of total yield suitable for it? Assume the given frequency distribution to be typical of the variety.

Weight (in gms)	*Number of Mangoes*	*Weight (in gms)*	*Number of Mangoes*
410–419	10	450–459	45
420–429	20	460–469	18
430–439	42	470–479	7
440–449	54		

Draw an ogive of 'more than' type of the above data and deduce how many mangoes will be more than 443 g.

3.35 Gupta Machine Company has a contract with his customers to supply machined pump gears. One requirement is that the diameter of gears be within specific limits. Here are the diameters (in inches) of a sample of 20 gears:

4.01	4.00	4.02	4.02	4.03	4.00
3.98	3.99	3.99	4.01	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]

3.36 Given the following frequency distribution with some missing frequencies:

Class	*Frequency*	*Class*	*Frequency*
10–20	185	50–60	136
20–30	—	60–70	—
30–40	34	70–80	50
40–50	180		

If the total frequency is 685 and median is 42.6, find out the missing frequencies.

Hints and Answers

3.29 Disapprove

3.30 Med = 27.5; $P_{90} - P_{10} = 47.14 - 11.67 = 35.47$

3.31 $a = 3, b = 6; x = 9, y = 12$

3.32 (a) 13.25 (b) $Q_3 = 17.6$, $Q_1 = 10.4$

3.34 52.25%; 103

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

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

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

$$= \sqrt{\frac{1}{19}\{(320.32 - 20\,(4.002)^2\}}$$

$= 0.016$ inches 95% of the gears will have diameter in the interval : $\bar{x} \pm 2s = (4.002 \pm 0.016)$

3.36 20–30(77)

3.11 MODE

Mode value: A measure of location recognised by the location of the most frequently occurring value of a set of data.

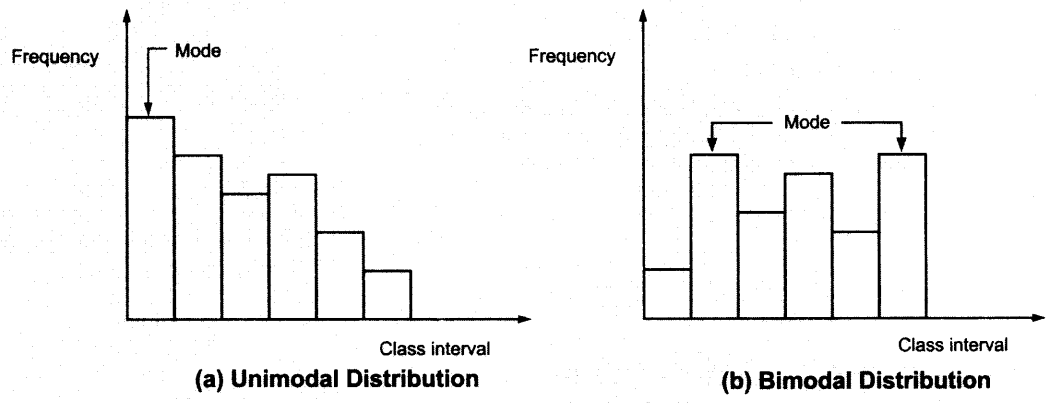
The **mode** is that value of an observation which occurs most frequently in the data set, that is, the point (or class mark) with the highest frequency.

The concept of mode is of great use to large scale manufacturers of consumable items such as ready-made garments, shoe-makers, and so on. In all such cases it is important to know the size that fits most persons rather than 'mean' size.

There are many practical situations in which arithmetic mean does not always provide an accurate characteristic (reflection) of the data due to the presence of extreme values. For example, in all such statements like 'average man prefers ... brand of cigarettes', 'average production of an item in a month', or 'average service time at the service counter'. The term 'average' means majority (i.e., mode value) and not the arithmetic mean. Similarly, the median may not represent the characteristics of the data set completely owing to an uneven distribution of the values of observations. For example, suppose in a distribution the values in the lower half vary from 10 to 100 (say), while the same number of observations in the upper half very from 100 to 7000 (say) with most of them close to the higher limit. In such a distribution, the median value of 100 will not provide an indication of the true nature of the data. Such shortcomings stated above for mean and median are removed by the use of *mode*, the third measure of central tendency.

The mode is a poor measure of central tendency when most frequently occurring values of an observation do not appear close to the centre of the data. The mode need not even be unique value. Consider the frequency distributions shown in Fig. 3.3(a) and (b). The distribution in Fig. 3.3(a) has its mode at the lowest class and certainly cannot be considered representative of central location. The distribution shown Fig. 3.3(b) has two modes. Obviously neither of these values appear to be representative of the central location of the data. For these reasons the mode has limited use as a measure of central tendency for decision-making. However, for descriptive analysis, mode is a useful measure of central tendency.

Fig. 3.3
Frequency Distribution



Calculation of Mode It is always preferable to calculate mode from grouped data. Table 3.27, for example, shows the sales of an item per day for 20 days period. The mode of this data is 71 since this value occurs more frequently (four times than any other value). However, it fails to reveal the fact that most of the values are under 70.

Table 3.27 Sales During 20 Days Period
(Data arranged in ascending order)

53,	56,	57,	58,	58,	60,	61,	63,	63,	64
64,	65,	65,	67,	68,	71,	71,	71,	71,	74

Converting this data into a frequency distribution as shown in Table 3.28:

Table 3.28 Frequency Distribution of Sales Per Day

Sales volume (Class interval) :	53–56	57–60	61–64	65–68	69–72	72 and above
Number of days (Frequency) :	2	4	5	4	4	1

Table 3.28 shows that a sale of 61–64 units of the item was achieved on 5 days. Thus this class is more representative of the sales per day.

In the case of grouped data, the following formula is used for calculating mode:

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

where l = lower limit of the model class interval
f_{m-1} = frequency of the class preceding the mode class interval
f_{m+1} = frequency of the class following the mode class interval
h = width of the mode class interval

Example 3.37: Using the data of Table 3.28, calculate the mode of sales distribution of the units of item during the 20 days period.

Solution: Since the largest frequency corresponds to the class interval 61–64, therefore it is the mode class. Then we have, $l = 61, f_m = 5, f_{m-1} = 4, f_{m+1} = 4$ and $h = 3$. Thus

$$M_0 = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \times h$$

$$= 61 + \frac{5-4}{10-4-4} \times 3 = 61 + 1.5 = 62.5$$

Hence, the modal sale is of 62.5 units.

Example 3.38: In 500 small-scale industrial units, the return on investment ranged from 0 to 30 per cent; no unit sustaining loss. Five per cent of the units had returns ranging from zero per cent to (and including) 5 per cent, and 15 per cent of the units earned returns exceeding 5 per cent but not exceeding 10 per cent. The median rate of return was 15 per cent and the upper quartile 2 per cent. The uppermost layer of returns exceeding 25 per cent was earned by 50 units.

(a) Present the information in the form of a frequency table as follows:
Exceeding 0 per cent but not exceeding 5 per cent
Exceeding 5 per cent but not exceeding 10 per cent
Exceeding 10 per cent but not exceeding 15 per cent
and so on.

(b) Find the rate of return around which there is maximum concentration of units.

Solution: (a) The given information is summarized in the form of a frequency distribution as shown in Table 3.29.

Table 3.29

Rate of Return	*Industrial Units*
Exceeding 0 per cent but not exceeding 5 per cent	$500 \times \frac{5}{100} = 25$
Exceeding 5 per cent but not exceeding 10 per cent	$500 \times \frac{15}{100} = 75$
Exceeding 10 per cent but not exceeding 15 per cent	$250 - 100 = 150$
Exceeding 15 per cent but not exceeding 20 per cent	$375 - 250 = 125$
Exceeding 20 per cent but not exceeding 25 per cent	$500 - 375 - 50 = 75$
Exceeding 25 per cent but not exceeding 30 per cent	50

(b) Calculating mode to find out the rate of return around which there is maximum concentration of the units. The mode lies in the class interval 10–15. Thus

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

$$= 10 + \frac{150 - 75}{2 \times 150 - 75 - 125} \times 5 = 10 + 3.75 = 13.75 \text{ per cent}$$

3.11.1 Graphical Method for Calculating Mode Value

The procedure of calculating mode using the graphical method is summarized below:

(i) Draw a histogram of the data, the tallest rectangle will represent the modal class.
(ii) Draw two diagonal lines from the top right corner and left corner of the tallest rectangle to the top right corner and left corner of the adjacent rectangles.
(iii) Draw a perpendicular line from the point of intersection of the two diagonal lines on the x-axis. The value on the x-axis marked by the line will represent the modal value.

Example 3.39: Calculate the mode using the graphical method for the following distribution of data:

Sales (in units) :	53–56	57–60	61–64	65–68	69–72	73–76
Number of days :	2	4	5	4	4	1

Solution: Construct a histogram of the data shown in Fig. 3.4 and draw other lines for the calculation of mode value.

The mode value from Fig. 3.4 is 62.5 which is same as calculated in Example 3.37.

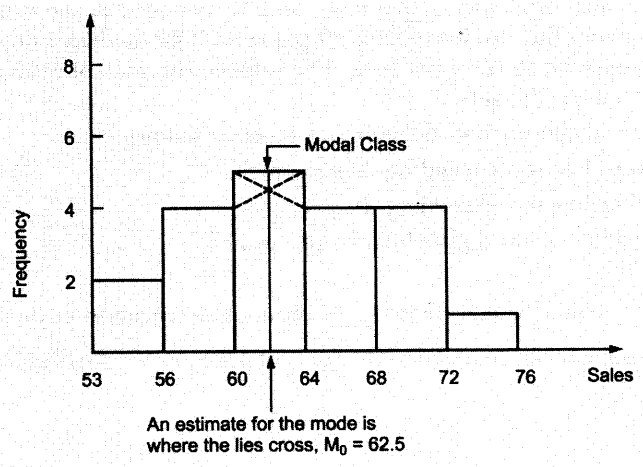


Figure 3.4
Graph for Modal Value

3.11.2 Advantages and Disadvantages of Mode Value

Advantages

(i) Mode value is easy to understand and to calculate. Mode class can also be located by inspection.
(ii) The mode is not affected by the extreme values in the distribution. The mode value can also be calculated for open-end frequency distributions.
(iii) The mode can be used to describe quantitative as well as qualitative data. For example, its value is used for comparing consumer preferences for various types of products, say cigarettes, soaps, toothpastes, or other products.

Disadvantages

(i) Mode is not a rigidly defined measure as there are several methods for calculating its value.
(ii) It is difficult to locate modal class in the case of multi-modal frequency distributions.
(iii) Mode is not suitable for algebraic manipulations.
(iv) When data sets contain more than one modes, such values are difficult to interpret and compare.

3.12 RELATIONSHIP BETWEEN MEAN, MEDIAN, AND MODE

In a *unimodal* and symmetrical distribution the values of mean, median, and mode are equal as indicated in Fig. 3.5. In other words, when these three values are all not equal to each other, the distribution is not symmetrical.

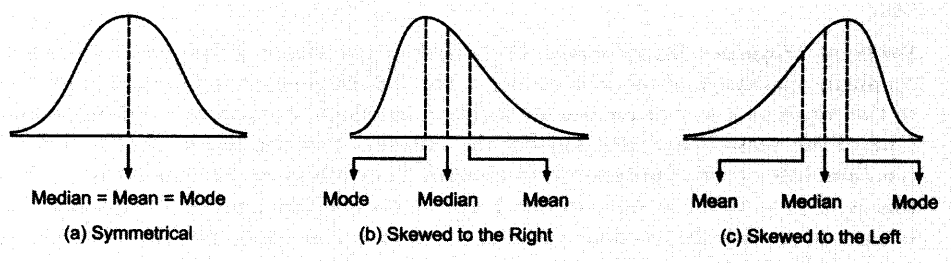


Figure 3.5
A comparison of Mean, Median, and Mode for three Distributional Shapes

A distribution that is not symmetrical, but rather has most of its values either to the right or to the left of the mode, is said to be *skewed*. For such asymmetrical distribution, Karl Pearson has suggested a relationship between these three measures of central tendency as:

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median}) \tag{3-22}$$

or

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean}$$

This implies that the value of any of these three measures can be calculated provided we know any two values out of three. The relationship (3-22) is shown in Fig. 3.5(b) and (c).

If most of the values of observations in a distribution fall to the right of the mode as shown in Fig. 3.5(b), then it is said to be skewed to the right *or positively skewed*. Distributions that are skewed right contain a few unusually large values of observations. In this case mode remains under the peak (i.e., representing highest frequency) but the median (value that depends on the number of observations) and mean move to the right (value that is affected by extreme values). The order of magnitude of these measures will be

$$\text{Mean} > \text{Median} > \text{Mode}$$

But if the distribution is skewed to the left or *negatively skewed* (i.e., values of lower magnitude are concentrated more to the left of the mode) then mode is again under the peak whereas median and mean move to the left of mode. The order of magnitude of these measures will be

$$\text{Mean} < \text{Median} < \text{Mode}$$

In both these cases, the difference between mean and mode is 3 times the difference between mean and median.

In general, for a single-peaked skewed distribution (non-symmetrical), the median is preferred to the mean for measuring location because it is neither influenced by the frequency of occurrence of a single observation value as mode nor it is affected by extreme values.

3.13 COMPARISON BETWEEN MEASURES OF CENTRAL TENDENCY

In this chapter, we have already presented three methods to understand the characteristics of a data set. However, the choice of which method to use for describing a distribution of values of observations in a data set is not always easy. The choice to use any one of these three is mainly guided by their characteristics. The characteristics of these three differ from each other with regard to three factors:

(i) Presence of outlier data values
(ii) Shape of the frequency distribution of data values
(iii) Status of theoretical development

Outlier: A very small or very large value in the data set.

1. ***The Presence of Outlier Data Values:*** The data values that differ in a big way from the other values in a data set are known as *outliers* (either very small or very ligh values). As mentioned earlier that the median is not sensitive to outlier values because its value depend only on the number of observations and the value always lies in the middle of the ordered set of values, whereas mean, which is calculated using all data values is sensitive to the outlier values in a data set. Obviously, smaller the number of observations in a data set, greater the influence of any outliers on the mean. The median is said to be *resistant* to the presence of outlier data values, but the mean is not.

2. ***Shape of Frequency Distribution:*** The effect of the shape of frequency distribution on mean, median, and mode is shown in Fig. 3.5. In general, the median is preferred to the mean as a way of measuring location for single peaked, skewed distributions. One of the reasons is that it satisfies the criterion that the *sum of absolute* difference (i.e., absolute error of judgment) of median from values in the data set is minimum, that is, $\Sigma \mid x - \text{Med} \mid = \min$. In other words, the smallest sum of the absolute errors is associated with the median value is the data set as compared to either mean or mode. When data is multi-modal, there is no single measure of central location and the mode can vary dramatically from one sample to another, particularly when dealing with small samples.

3. ***The Status of Theoretical Development:*** Although the three measures of central tendency—Mean, Median, and Mode, satisfy different mathematical criteria but the objective of any statistical analysis in *inferential statistics* is always to minimize the *sum of squared deviations* (*errors*) taken from these measures to every value in the data set. The criterion of the sum of squared deviations is also called *least squares criterion.* Since A.M. satisfies the least squares criterion, it is mathematically consistent with several techniques of statistical inference.

   As with the median, it can not be used to develop theoretical concepts and models and so is only used for basic descriptive purposes.

Conceptual Questions 3 D

24. Give a brief description of the different measures of central tendency. Why is arithmetic mean so popular?

25. How would you explain the choice of arithmetic mean as the best measure of central tendency. Under what circumstances would you deem fit the use of median or mode?

26. What are the advantages and disadvantages of the three common averages: Mean, Median, and Mode?

27. Describe the relationship between the mean and median of a set of data to indicate the skewness of the distribution of values.

28. Identify the mathematical criteria associated with mean, median, and mode and briefly explain the meaning of each criterion.

29. It is said that the use of a particular average depends upon the particular problem in hand. Comment and indicate at least one instance of the use of mean, median, mode, geometric, and harmonic mean.

30. How would you account for the predominant choice of arithmetic mean as a measure of central tendency? Under what circumstances would it be appropriate to use mode or median? [*Delhi Univ., MBA, 2000*]

31. Under what circumstances would it be appropriate to use mean, median, or mode? Discuss [*Delhi Univ., MBA, 1996*]

32. Explain the properties of a good average. In the light of these properties which average do you think is best and why? [*Jodhpur Univ., MBA, 1996*]

33. Give a brief note of the measures of central tendency together with their merits and demerits. Which is the best measure of central tendency and why? [*Osmania Univ., MBA, 1998*]

34. What is a statistical average? What are the desirable properties for an average to possess? Mention the different types of averages and state why arithmetic mean is most commonly used amongst them.

35. What is the relationship between mean, median, and mode? Under what circumstances are they equal?

36. It has been said that the less variability that exists, the more an average is representative of a set of data. Comment on the meaning of this statement.

37. Which measure of central tendency is usually be preferred if the distribution is known to be single peaked and skewed? Why?

38. Suppose the average amount of cash (in pocket, wallet, purse, etc.) possessed by 60 students attending a class is Rs 125. The median amount carried is Rs 90.

(a) What characteristics of the distribution of cash carried by the students can be explained. Why is mean larger than the median?

(b) Identify the process or population to which inferences based on these results might apply.

Self-Practice Problems 3E

3.37 Given below is the distribution of profits (in '000 rupees) earned by 94 per cent of the retail grocery shops in a city.

Profits	*Number of Shops*	*Profit*	*Number of Shops*
0–10	0	50–60	68
10–20	5	60–70	83
20–30	14	70–80	91
30–40	27	80–90	94
40–50	48		

Calculate the modal value.

3.38 Compute mode value from the following data relating to dividend paid by companies in a particular financial year.

Dividend (in per cent) Value	*Number of (in per cent)*	*Dividend of the Share of the Share Value*	*Number of Companies Companies*
5.0– 7.5	182	15.0–17.5	280
7.5–10.0	75	17.5–20.0	236
10.0–12.5	59	20.0–22.5	378
12.5–15.0	127	22.5–25.0	331

3.39 Following is the cumulative frequency distribution of the preferred length of kitchen slabs obtained from the preference study on 50 housewives:

Length (metres) more than	*Number of Housewives*
1.0	50
1.5	46
2.0	40
2.5	42
3.0	10
3.5	3

A manufacturer has to take a decision on what length of slabs to manufacture. What length would you recommend and why?

3.40 The management of Doordarshan holds a preview of a new programme and asks viewers for their reaction. The following results by age groups, were obtained.

Age group	*Under 20*	*20–39*	*40–59*	*60 and above*
Liked the program :	140	75	50	40
Dislike the program :	60	50	50	20

Using a suitable measure of central tendency, suggest towards which age group the management should aim its advertising campaign.

3.41 A sample of 100 households in a given city revealed the following number of persons per household:

Number of Persons	*No. of Households*
1	16
2	28
3–4	37
5–6	12
7–11	7

(a) What is the modal category for the 100 households observed?

(b) What proportion of the households have more than four persons.

3.42 The number of solar heating systems available to the public is quite large, and their heat storage capacities are quite varied. Here is a distribution of heat storage capacity (in days) of 28 systems that were tested recently by a testing agency

Days	*Frequency*	*Days*	*Frequency*
0–0.99	2	4–4.99	5
1–1.99	4	5–5.99	3
2–2.99	6	6–6.99	1
3–3.99	7		

The agency knows that its report on the tests will be widely circulated and used as the basis for solar heat allowances.

(a) Compute the mean, median, and mode of these data.

(b) Select the answer from part (a) which best reflects the central tendency of the test data and justify your choice.

3.43 Mr Pandey does statistical analysis for an automobile racing team. The data on fuel consumption (in km per litre) for the team's cars in recent races are as follows:

14.77 16.11 16.11 15.05 15.99 14.91
15.27 16.01 15.75 14.89 16.05 15.22
16.02 15.24 16.11 15.02

(a) Calculate the mean and median fuel consumption.

(b) Group the data into five equally-sized classes. What is the fuel consumption value of the modal class?

(c) Which of the three measures of central tendency is best to use? Explain.

3.44 An agriculture farm sells grab bags of flower bulbs. The bags are sold by weight; thus the number of bulbs in each bag can vary depending on the varieties included. Below are the number of bulbs in each of the 20 bags sampled:

21 33 37 56 47 25 33 32 47 34
36 23 26 33 37 26 37 37 43 45

(a) What are the mean and median number of bulbs per bag?

(b) Based on your answer, what can you conclude about the shape of the distribution of number of bulbs per bag?

3.45 The table below is the frequency distribution of ages to the nearest birthday for a random sample of 50 employees in a large company

Age to nearest birthday	:	20–29	30–39	40–49	50–59	60–69
Number of employees	:	5	12	13	8	12

Compute the mean, median, and mode for these data.

3.46 A track coach is in the process of selecting one of the two sprinters for the 200 meter race at the upcoming games. He has the following data of the results of five races (time in seconds) of the two sprinters run with 15 minutes rest intervals in between.

Athlete	*Races*				
	1	*2*	*3*	*4*	*5*
Vibhor	24.2	24.1	24.1	28.9	24.2
Prasant	24.4	24.5	24.5	24.6	24.5

Based on these data, which of the two sprinters should the coach select? Why?

3.47 The following data are the yields (in per cent) in the money market of 10 companies listed at the Bombay Stock Exchange (BSE) as on 18 October 2001, the day before the BSE index average passed the 3000 mark.

Company	*Money Market Yield (Per cent)*
Tata Power	10.0
HCL Infosys	7.5
ITC	5.7
NIIT	5.4
Cipla	4.6
Reliance Petro	4.1
Reliance	4.0
Dr. Reddy's Lab	3.9
Digital Glob	3.0
ICICI	2.9

(a) Compute the mean, median, mode, quartile, and decile deviation for these yields.

(b) What other information would you want to know if you were deciding to buy shares of one of these companies? Prepare a list of questions that you would like to ask a broker.

3.48 In order to estimate how much water will need to be supplied to a locality in East Delhi area during the summer of 2002, the minister asked the General Manager of the water supply department to find out how much water a sample of families currently uses. The sample of 20 families used the following number of gallons (in thousands) in the past years.

9.3 19.6 14.5 17.8 14.7 15.0 13.9 12.7
10.0 13.0 25.0 16.3 11.2 20.2 15.4 11.6
16.5 11.0 12.2 10.9

(a) What is the mean and median amount of water used per family?

(b) Suppose that 10 years from now, the government expects that there will be 1800 families living in that colony. How many gallons of water will be needed annually, if rate of consumption per family remains the same?

(c) In what ways would the information provided in (a) and (b) be useful to the government? Discuss.

(d) Why might the government have used the data from a survey rather than just measuring the total consumption in Delhi?

Hints and Answers

3.37 $M_0 = 3\ \text{Med} - 2\bar{x}$ = Rs 50.04 thousand

3.38 M_0 = 21.87 (per cent of the share value)

3.41 (a) Persons between 3–4

(b) 19 per cent house holds

Formulae Used

1. Summation of n numbers

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Simplified expression for the summation of n numbers

$$\Sigma x_i = x_1 + x_2 + \ldots + x_n$$

2. Sample mean, $\bar{x} = \dfrac{\Sigma x_i}{n}$

Population mean, $\mu = \dfrac{\Sigma x_i}{N}$

Sample mean for grouped data, $\bar{x} = \dfrac{\Sigma f_i m_i}{n}$

where $n = \Sigma f_i$ and m_i = mid-value of class intervals

3. Weighted mean for a population or a sample,

$$\bar{x}_w \text{ or } \mu_w = \frac{\Sigma w_i x_i}{\Sigma w_i}$$

where w_i = weight for observation i,

4. Position of the median in an ordered set of observation belong to a population or a sample is, $\text{Med} = x_{(n/2)+(1/2)}$

Median for grouped data, $\text{Med} = l + \left[\dfrac{(n/2) - cf}{f}\right]h$

5. Quartile for a grouped data

$$Q_i = l + \left[\frac{i(n/4) - cf}{f}\right]h;\quad i = 1, 2, 3$$

Decile for a grouped data

$$D_i = l + \left[\frac{i(n/10) - cf}{f}\right]h;\quad i = 1, 2, \ldots, 9$$

Percentile for a grouped data

$$P_i = l + \left[\frac{i(n/100) - cf}{f}\right]h;\quad i = 1, 2, \ldots, 99$$

6. Mode for a grouped data

$$M_0 = l + \left[\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}}\right]h$$

Mode for a multimode frequency distribution

$$M_0 = 3\ \text{Median} - 2\ \text{Mean}$$

Chapter Concepts Quiz

True or False

1. Inferential statistics are used to describe specific characteristics of the data. (T/F)
2. With nominal data, the mean should be used as a measure of central tendency. (T/F)
3. With ordinal data, we can use both the mode and the mean as a measure of central tendency. (T/F)
4. When the data are interval or ratio, we can use the mean as a measure of central tendency. (T/F)
5. Harmonic mean is the reciprocal of arithmetic mean. (T/F)
6. Sum of absolute deviations from median is minimum. (T/F)
7. The mean of a data set remains unaffected if an observation equal to mean is included in it. (T/F)
8. With continuous data, the median is the most appropriate measure of central tendency. (T/F)
9. Weighted mean is useful in problems relating to the construction of index numbers and standardized birth and death rates. (T/F)
10. It is possible to have data with three different values for measures of central tendency. (T/F)
11. The median is less affected than the mean by extreme values of observations in a distribution. (T/F)
12. The sum of deviations from mean is zero. (T/F)
13. If the number of observations is even, the median is in the middle of the distribution. (T/F)
14. The mode is always found at the highest point of a graph of a frequency distribution. (T/F)
15. For grouped data, it is possible to calculate an approximate mean by assuming that each value in a given class is equal to its mid point. (T/F)

Multiple Choice

16. Which of the following statements is not correct?
(a) Few data sets do not have arithmetic means
(b) Calculation of arithmetic mean is affected by extreme data values
(c) The weighted mean should be used when it is necessary to take the importance of each value into account
(d) All these statements are correct

17. The algebraic sum of the deviations from mean is:
(a) maximum (b) minimum
(c) zero (d) none of the above

18. Given a normally distributed continuous variable the best measure of central tendency is:
(a) mode (b) median
(c) mean (d) none of the above

19. Calculation of a mode from grouped data is usually better than from ungrouped data because:
(a) ungrouped data tend to be bimodal
(b) regardless of the skewness of the distribution, mode for the grouped data remains same
(c) grouped data is least affected by extreme values
(d) mode is likely to be most representative of the data set

20. The arithmetic mean of the first n natural numbers, 1, 2, ..., n is:
(a) $n/2$ (b) $(n + 1)/2$
(c) $n(n + 1)/2$ (d) none of the above

21. The sum of squares of deviations from mean is:
(a) maximum (b) minimum
(c) zero (d) none of the above

22. The measure of central tendency which is most strongly influenced by extreme values in the 'tail' of the distribution is:
(a) mean (b) median
(c) mode (d) none of the above

23. The major assumption for computing a mean value from a group data is:
(a) each class contains equal number of observations
(b) each observation in a class is equal to the mid-value
(c) no observation occurs more than once
(d) none of the above

24. If an observation in the data set in zero, then its geometric mean is:
(a) positive (b) negative
(e) zero (d) indeterminant

25. If an observation in the data set is negative, then its geometric mean is:
(a) positive (b) negative
(c) zero (d) indeterminant

26. For the given set of observations 1, 4, 4, 4, and 7, it can be said that the:
(a) mean is larger than either median or mode
(b) mean = median = mode
(c) mean ≠ median ≠ mode
(d) none of the above

27. If in a set of discrete values of observations, 50 per cent values are greater than 25, then Q_2 is:
(a) 20 (b) 25
(c) 50 (d) 75

28. The relationship between AM, GM and HM is:
(a) G.M. = (A.M.) × (H.M.)
(b) $(G.M.)^2 = (A.M.) \times (H.M.)$
(c) $G.M. = (A.M. \times H.M.)^2$
(d) $(G.M.)^2 = (A.M.)^2 \times (H.M.)^2$

29. Which of the following relationship is true in a symmetrical distribution?
(a) Median $- Q_1 = Q_3 -$ Median
(b) Median $- Q_1 > Q_3 -$ Median
(c) Median $- Q_1 < Q_3 -$ Median
(d) None of the above

30. Which of the following relationship is true in a multimodal distribution?
(a) Mean – Mode = 3 (Mean – Median)
(b) Mode = 3 Median – 2 Mean
(c) 3 Median = (2 Mean + Mode)
(d) All of the above

Concepts Quiz Answers

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

Review Self-Practice Problems

3.49 The following is the data on profit margin (in per cent) of three products and their corresponding sales (in Rs) during a particular period.

Product	*Profit Margin (Per cent)*	*Sales (Rs in thousand)*
A	12.5	2,000
B	10.3	6,000
C	6.4	10,000

(a) Determine the mean profit margin.
(b) Determine the weighted mean considering the rupee sales as weight for each product.
(c) Which of the means calculated in part (a) and (b) is the correct one?

3.50 The number of cars sold by each of the 10 car dealers during a particular month, arranged in ascending order, is 12, 14, 17, 20, 20, 20, 22, 22, 24, 25. Considering this scale to be the statistical population of interest, determine the mean, median, and mode for the number of cars sold.
(a) Which value calculated above best describes the 'typical' sales volume per dealer?
(b) For the given data, determine the values at the (i) quartile Q_1 and (ii) percentile P_{30} for these sales amounts.

3.51 A quality control inspector tested nine samples of each of three designs A, B and C of certain bearing for a new electrical winch. The following data are the number of hours it took for each bearing to fail when the winch motor was run continuously at maximum output, with a load on the winch equivalent to 1.9 times the intended capacity.

A :	16	16	53	15	31	17	14	30	20
B :	18	27	23	21	22	26	39	17	28
C :	31	16	42	20	18	17	16	15	19

Calculate the mean and median for each group and suggest which design is best and why?

[IIPM, PGDM, 2002]

3.52 Calculate the mean, median, and mode for the following data pertaining to marks in statistics. There are 80 students in a class and the test is of 140 marks.

Marks more than :	0	20	40	60	80	100	120
Number of students :	80	76	50	28	18	9	3

[M.D. Univ., MBA, 1994]

3.53 A company invests one lakh rupees at 10 per cent annual rate of interest. What will be the total amount after 6 years if the principal is not withdrawn?

3.54 Draw an ogive for the following distribution. Read the median from the graph and verify your result by the mathematical formula. Also obtain the limits of income of the central 50% of employees.

Weekly Income (Rs)	*Number of Employees*	*Weekly Income (Rs)*	*Number of Employees*
Below 550	6	700–750	16
550–600	10	750–800	12
600–650	22	800 and above	15
650–700	30		

[Delhi Univ., MBA, 1999]

3.55 In the production of light bulbs, many bulbs are broken. A production manager is testing a new type of conveyor system in the hope of reducing the percentage of bulbs broken each day. For ten days he observes bulb breakage with the current conveyor. He then records bulb breakage for ten days with the new system, after allowing a few days for the operator to learn to use it. His data are as follows:

Conveyor System	*Percentage of Bulbs Broken Daily*									
Old	8.7	11.1	4.4	3.7	9.2	6.6	7.8	4.9	6.9	8.3
New	10.8	6.2	3.2	4.6	5.3	6.5	4.6	7.1	4.9	7.2

(a) Compute the mean and median for each conveyor system.
(b) Based on these results, do you think this test establishes that the new system lowers the breakage rate? Explain.

3.56 The following are the weekly wages in rupees of 30 workers of a firm:

140	139	126	114	100	88	62	77	99
103	108	129	144	148	134	63	69	148
132	118	142	116	123	104	95	80	85
106	123	133						

The firm gave bonus of Rs 10, 15, 20, 25, 30, and 35 for individuals in the respective salary slabs: exceeding 60 but not exceeding 75; exceeding but not exceeding 90; and so on up to exceeding 135 and not exceeding 150. Find the average bonus paid.

3.57 The mean monthly salaries paid to 100 employees of a company was Rs 5,000. The mean monthly salaries paid to male and female employees were Rs 5,200 and Rs 4,200 respectively. Determine the percentage of males and females employed by the company.

3.58 The following is the age distribution of 2,000 persons working in a large textile mill:

Age Group	*No. of Persons*	*Age Group*	*No. of Persons*
15 but less than 20	80	45 but less than 50	268
20 but less than 25	250	50 but less than 55	150
25 but less than 30	300	55 but less than 60	75
30 but less than 35	325	60 but less than 65	25
35 but less than 40	287	65 but less than 70	20
40 but less than 45	220		

Because of heavy losses the management decides to bring down the strength to 40 per cent of the present number according to the following scheme:
(i) To retrench the first 10 per cent from lowest age group 15–20.
(ii) To absorb the next 40 per cent in other branches.
(iii) To make 10 per cent from the highest age group, 40–45 retire prematurely.

What will be the age limits of persons retained in the mill and of those transferred to other branches? Also calculate the average age of those retained.

3.59 A factory pays workers on piece rate basis and also a bonus to each worker on the basis of individual output in each quarter. The rate of bonus payable is as follows:

Output (in units)	*Bonus (Rs)*	*Output (in units)*	*Bonus (Rs)*
70–74	40	90–94	70
75–79	45	95–99	80
80–84	50	100–104	100
85–89	60		

The individual output of a batch of 50 workers is given below:

94	83	78	76	88	86	93	80	91	82
89	97	92	84	92	80	85	83	98	103
87	88	88	81	95	86	99	81	87	90
84	97	80	75	93	101	82	82	89	72
85	83	75	72	83	98	77	87	71	80

By suitable classification you are required to find:

(a) Average bonus per worker for the quarter

(b) Average output per worker.

[*Pune Univ., MBA, 1998*]

3.60 An economy grows at the rate of 2 per cent in the first year, 2.5 per cent in the second year, 3 per cent in the third year, 4 per cent in the fourth year . . . and 10 per cent in the tenth year. What is the average rate of growth of the company?

3.61 A man travelled by car for 3 days. He covered 480 km each day. On the first day he drove for 10 hours at 48 km an hour, on the second day he drove for 12 hours at 40 km an hour, and on the last day he drove for 15 hours at 32 km per hour. What was his average speed? [*Bangalore Univ., BCom, 1996*]

3.62 The monthly income of employees in an industrial concern are given below. The total income of 10 employees in the class over Rs 25,000 is Rs 3,00,000. Compute the mean income. Every employee belonging to the top 25 per cent of the earners is required to pay 5 per cent of his income to the workers' relief fund. Estimate the contribution to this fund.

Income (Rs)	*Frequency*	*Income (Rs)*	*Frequency*
Below 5000	90	15,000–20,000	80
5000–10000	150	20,000–25,000	70
10000–15000	100	25,000 and above	10

[*Kakatiya Univ., MCom, 1997*]

3.63 In a factory there are 100 skilled, 250 semi-skilled, and 150 unskilled workers. It has been observed that on an average a unit length of a particular fabric is woven by a skilled worker in 3 hours, by a semi-skilled worker in 4 hours, and by an unskilled worker in 5 hours. Unskilled workers are expected to become semi-skilled workers and semi-skilled workers are expected to become skilled. How much less time will be required after 2 years of training for weaving the unit length of fabric by an average worker?

3.64 The price of a certain commodity in the first week of January is 400 g per rupee; it is 600 g per rupee in the second week and 500 g per rupee in the third week. Is it correct to say that the average price is 500 g per rupee? Verify.

3.65 Find the missing information in the following table:

	A	*B*	*C*	*Combined*
Number	10	8	—	24
Mean	20	—	6	15
Geometric Mean	10	7	—	8.397

[*Delhi Univ., BCom (Hons), 1998*]

3.66 During a period of decline in stock market prices, a stock is sold at Rs 50 per share on one day, Rs 40 on the next day, and Rs 25 on the third day.

(a) If an investor bought 100, 120, and 180 shares on the respective three days, find the average price paid per share.

(b) If the investor bought Rs 1000 worth of shares on each of the three days, find the average price paid per share. [*Delhi Univ., BA (Hons Econ.), 1998*]

Hints and Answers

3.49 (a) $\mu = \dfrac{\Sigma x_i}{N} = \dfrac{29.2}{3} = 9.73$ per cent

(b) $\mu_w = \dfrac{\Sigma w_i x_i}{\Sigma w_i} = \dfrac{1{,}50{,}800}{18{,}000} = 8.37$ per cent

(c) The weighted mean of 8.37 per cent considering sales (in Rs) as weights is the correct mean profit margin. Percentages should never the averaged without being weighted.

3.50 $\mu = \dfrac{\Sigma x_i}{N} = \dfrac{196}{10} = 19.6$

$$\text{Med} = \frac{\left(\frac{n}{2}\right) + \left(\frac{n}{2} + 1\right)}{2} = \frac{5\text{th} + 6\text{th}}{2} = 20.0$$

M_0 = most frequent value = 20

(a) Median is best used as the 'typical' value because of the skewness in the distribution of values

(b) $Q_1 = x_{(n/4) + (1/2)} = x_{(10/4) + (1/2)} = x_{3.0} = 17$

$P_{30} = x_{(3n/10) + (1/2)} = x_{3.5} = 17 + 0.5\ (20 - 17)$
$= 18.5$

3.51 Listing the data in ascending order:

A : 14 15 16 16 17 20 30 31 53

B : 17 18 21 22 23 26 27 28 39

C : 15 16 16 17 18 19 20 31 42

$\bar{x}_A = 212/9 = 23.56$; Med (A) = 17

$\bar{x}_B = 221/9 = 24.56$; Med (B) = 23

$\bar{x}_C = 194/9 = 21.56$; Med (C) = 18

Since medians are the fifth observation in each data set, therefore design B is best because both the mean and median are highest.

3.52 Arrange the marks in statistics into following class intervals:

Marks	0–20	20–40	40–60	60–80	80–100	100–120	120–140
No. of students	4	26	22	10	9	6	3

Mean = 56, Median = 49.09, and Mode = 36.92

3.53 Principal amount, A = Rs 1,00,000; $r = 10$ and $n = 6$;

$$P_n = A\left(1+\frac{r}{100}\right)^n = 1{,}00{,}000\left(1+\frac{10}{100}\right)^6$$

Taking log P_n and then antilog, we get P_n = Rs 1,77,2000

3.54 Median = Rs 679.20; Limits of income of central 50% of the employees = Rs 626.7 to Rs 747.7

3.56 Prepare a frequency distribution as follows:

Weekly Wages (Rs)	*Frequency (f)*	*Bonus Paid (x)*	*Weekly Wages (Rs)*	*Frequency (f)*	*Bonus Paid (x)*
61– 75	3	10	106–120	5	25
76– 90	4	15	121–135	7	30
91–105	5	20	136–150	6	35

$$\text{Average bonus paid} = \frac{\Sigma fx}{n} = \frac{375}{30} = \text{Rs } 24.5$$

3.57 Given $n_1 + n_2 = 100$, $\bar{x}_{12} = 5000$, $\bar{x}_1 = 5200$ and $\bar{x}_2 = 4200$

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\text{or} \quad 5000 = \frac{n_1(5200) + n_2(4200)}{n_1 + n_2}$$

$$= \frac{n_1(5200) + (100 - n_1)(4200)}{100}$$

$1000\,n_1 = 80{,}000$ or $n_1 = 80$ and $n_2 = 100 - n_1 = 20$

3.58 The number of persons to be retrenched from the lower group are: (20,000 × 10) ÷ 100 = 200. Eighty of these will be in the 15–20 age group and the rest 120 (= 200 – 80) in the 20–25 age group.

The persons to be absorbed in other branches = (2,000 × 10) ÷ 100 = 800. These persons belong to the following age groups:

Age Group		*No. of Persons*
20–25	(250–120)	130
25–30	—	300
30–35	—	325
35–40	(287–242)	45
		800

Those who will be retiring (2,000 × 10) ÷ 100 = 200 and these persons belong to highest age group as shown below:

Age Group		*No. of Persons*
65–70	—	20
60–65	—	25
55–60	—	75
50–55	(150–70)	80
		200

Age limits of persons who are retained:

Age Group	*No. of Persons*
35–40	242
40–45	220
45–50	268
50–55	70
	800

Average age of those retained: 43.54 years.

3.59

Output (in units)	*Frequency (f)*	*Bonus (Rs)*	*Output (in units)*	*Frequency (f)*	*Bonus (Rs)*
70–74	3	40	90–94	7	70
75–79	5	45	95–99	6	80
80–84	15	50	100–104	2	100
85–89	12	60			

(a) Average bonus/worker for quarter, $\bar{x} = \Sigma fx/n$ = 2,985/50 = Rs 59.7

(b) Total quarterly bonus paid = Rs 59.7 × 50 = Rs 2,985

(c) Average output/worker, $\bar{x}$ = 86.1 units

3.60

Year	1	2	3	4	5	6	7	8	9	10
Growth rate	2	2.5	3	4	5	6	7	8	9	10
Value at the end of year x	102	102.5	103	104	105	106	107	108	109	110

G.M. = Antilog ($\Sigma \log x \div n$) = Antilog (20.237 ÷ 10) = 105.6

Average growth rate = 105.6 – 100 = 5.6 per cent

3.61 H.M. $= n\Big/\left(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}\right) = 3\Big/\left(\frac{1}{48}+\frac{1}{40}+\frac{1}{32}\right)$

= 38.98 km per hour

3.62 Number of employees belonging to the top 25% of the earners is = (25 ÷ 100) × 500 = 125 and the distribution of these top earners is as follows:

Income (Rs)	*Frequency*
25,000 and above	10
20,000–25,000	70
15,000–20,000	45

80 persons have income in the range 15,000–20,000 = Rs 500 and therefore 45 persons will have income in the range (500 ÷ 80) × 45 = 281.25 or 281.

The top 45 earners in the income group 15,000–20,000 will have salaries ranging from (20,000–281 to 20,000, i.e. 19,719 to 20,000. Thus the distribution of top 125 persons is as follows:

Income (Rs)	*Mid-value (m)*	*Frequency (f)*	*Total Income (f × m)*
25,000 and above	—	10	3,00,000 (given)
20,000–25,000	22,500	70	15,75,000
19,719–20,000	19,859.5	45	8,93,677.5
		125	27,68,677.5

Hence the total income of the top 25% of earners is Rs 2,71,177.5. Contribution to the fund = 5% of 2,71,177.5 = Rs 93,598.

3.63 Average time per worker before training:

$$\frac{(100\times 3)+(250\times 4)+(150\times 5)}{100+250+150} = \frac{2050}{500} = 4.1 \text{ hours}$$

After training the composition of workers is as follows:

Skilled workers = 100 + 250 = 350
Semi-skilled workers = 150
Unskilled workers = Nil

Average time per worker after training is:

$$\frac{(350\times 3)+(150\times 4)}{350+150} = \frac{1050+600}{500} = 3.3 \text{ hours}$$

Thus after 2 years 0.8 hours less would be required.

3.64 Harmonic Mean $= n\Big/\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}\right)$

$$= 3\Big/\left(\frac{1}{400}+\frac{1}{600}+\frac{1}{500}\right) = 0.486 \text{ g per rupee}$$

3.65 Mean : Let x be the mean of B. Then

$(20 \times 10) + (8 \times x) + (6 \times 6) = (15 \times 24)$

$8x = 124$ or $x = 15.5$

Hence mean of B = 15.5

Geometric mean: Let x be the geometric mean of C. Then

$(10)^{10} \times (7)^8 \times x^6 = (8.397)^{24}$

$10 \log 10 + 8 \log 7 + 6 \log x = 24 \log 8.397$

$10 + (8 \times 0.8451) + 6 \log x = 24\ (0.9241)$

$6 \log x = 5.4176$ or x = Antilog 0.9029 = 7.997

Hence, geometric mean of C is 7.997.

3.66 Average price paid per share $= \frac{\Sigma wx}{\Sigma w} = \frac{14{,}300}{400}$

$= 35.75.$

Case Studies

Case 3.1: Kanta Bread

'Kanta Bread' company is manufacturing bread. The selling price of bread is fixed by the government. In view of increasing raw material prices the only way to sustain the profit level is by efficient control over the production process to reduce the cost of production. A large quantity of production shows small savings per unit of production and results in a significant amount cumulatively. Among various factors affecting the cost of production, the yield of the finished product obtained from a given quantity of raw materials is of considerable significance.

The process of manufacturing bread involves mixing of wheat flour (maida) with the required quantity of water and other ingredients as per the standard formula. The mixed bulk dough is subjected to yeast fermentation for a given period. Then this bulk quantity of dough is transferred to a machine called Divider, which divides the dough mechanically into small pieces of required weight, which are individually further processed and baked to get the finished product called bread.

For manufacturing a 800 g loaf of bread the required weight of dough piece coming from the Divider should be between 880 g and 885 g. This weight is termed as *dividing weight*. There is provision on the Divider machine to set the required dough piece weight. The smallest division (or increment) in weight that can be set on this machine is 5 g. As a measure of safety, 885 g is the weight that is usually set on the machine.

The performance of the machine, with respect to the accuracy of the weight of dough pieces delivered by it, is dependent on various factors such as the level of lubrication and the consistency of the dough. But these variations are less and the weight variation would be within ± 30 g of the set weight for a good Divider. Larger variations than this in weight occur only if the dough dividing mechanism (called Divider Head) has undergone considerable wear and tear.

The weight variation in the dough pieces from the Divider results in weight variation in the final product (bread). But the Bread weight, which is declared as 800 g is covered by the Weights and Measures (Packaged Commodities) Act, 1976. As per the rules of this Act, the average net weight of a random sample (Sample size is 20 bread) should not be less than the declared weight. The maximum permissible error in relation to the quantity contained in the individual package is 6 per cent of the declared weight. Also, it is stipulated that the number of individual packages showing an error in deficiency of weight greater than the maximum permissible error (i.e., 6%) should not be more than 5 per cent of the packages drawn as samples. The sample size as per the rules as applicable to bread is 20 individual packages. If the product does not confirm to the rules, then punitive action can be taken against the manufacturer.

In order to ensure that the rules under this Act are not violated, the rules are applied to the weights of the divided dough piece itself. Thus in terms of these rules it is required:

(i) that the average weight of a random sample of 20 divided dough pieces should be 880 g to 885 g.
(ii) that the weight of an individual dough piece should not be less than 832 g (i.e., 885 g – 6% = 832 g).
(iii) That not more than one piece out of 20 (i.e., 5%) should have a weight less than 832 g.

Thus the performance of the Divider should confirm to these standards. However it was complained by the production department that the weight variation at the Divider is very large and hence in order that the law is not violated, they are forced to set the Divider at a higher dividing weight than the normally required weight setting of 885 g, resulting in less on yield of the product.

Questions for Discussion

1. What is the average weight of dough pieces delivered by the Divider with the weight setting at 885 g?
2. What is the percentage of dough pieces with weight less than 832 g, if any?
3. What should be the lowest dividing weight setting on the Divider, so that the number of dough pieces with weight less than 832 g are less than 5 per cent, as stipulated by the law, (if already not so)?
4. The monetary loss to the company on account of setting the higher dividing weight as decided in Q.3 above could be calculated, which would help in deciding whether the Divider head should be replaced or not.

Data: The output of the Divider is 1000 pieces per hour. The weight of 20 dough pieces per hour. The weight of 20 dough pieces were taken using an accurate weighing scale. (Recording this weight is a routine quality control check). Ten sets of such dough weight on the machine were taken at different times. All the individual observations on these ten sets were pooled together to get the population of data with total 200 observations, given below:

SNo.	*Dividing Weight*	*SNo.*	*Dividing Weight*	*SNo.*	*Dividing Weight*	*SNo.*	*Dividing Weight*	*SNo.*	*Dividing Weight*
1.	875	41.	895	81.	836	121.	912	161.	870
2.	870	42.	910	82.	823	122.	912	162.	895
3.	852	43.	907	83.	910	123.	885	163.	870
4.	880	44.	912	84.	889	124.	895	164.	910
5.	909	45.	890	85.	897	125.	840	165.	910
6.	909	46.	895	86.	885	126.	860	166.	885
7.	893	47.	862	87.	892	127.	866	167.	920
8.	875	48.	875	88.	886	128.	875	168.	877
9.	830	49.	895	89.	886	129.	868	169.	909
10.	859	50.	867	90.	897	130.	861	170.	915
11.	827	51.	910	91.	897	131.	873	171.	920
12.	907	52.	900	92.	880	132.	893	172.	884
13.	909	53.	880	93.	870	133.	883	173.	910
14.	910	54.	900	94.	920	134.	878	174.	925
15.	915	55.	910	95.	927	135.	873	175.	900
16.	920	56.	897	96.	930	136.	893	176.	875
17.	905	57.	875	97.	925	137.	825	177.	895
18.	890	58.	905	98.	915	138.	830	178	930
19.	925	59.	975	99.	875	139.	845	179	863
20.	900	60.	909	100.	890	140.	830	180	913
21.	895	61.	890	101.	890	141.	840	181	903
22.	875	62.	880	102.	923	142.	835	182	843
23.	903	63.	890	103.	828	143.	830	183	878
24.	863	64.	825	104.	825	144.	860	184	883
25.	913	65.	845	105.	910	145.	855	185	878
26.	903	66.	875	106.	866	146.	860	186	897
27.	893	67.	890	107.	825	147.	835	187	916
28.	878	68.	892	108.	830	148.	886	188	907
29.	878	69.	821	109.	845	149.	888	189	912
30.	883	70.	826	110.	830	150.	890	190	895
31.	897	71.	904	111.	855	151.	878	191	885

Contd...

SNo.	*Dividing Weight*	*SNo.*	*Dividing Weight*	*SNo.*	*Dividing Weight*	*SNo.*	*Dividing Weight*	*SNo.*	*Dividing Weight*
32.	916	72.	915	112.	835	152.	858	192.	862
33.	813	73.	900	113.	855	153.	875	193.	879
34.	863	74.	900	114.	836	154.	858	194.	900
35.	900	75.	905	115.	860	155.	868	195.	872
36.	905	76.	885	116.	835	156.	888	196.	900
37.	898	77.	890	117.	865	157.	868	197.	905
38.	878	78.	889	118.	915	158.	865	198.	885
39.	878	79.	865	119.	930	159.	880	199.	895
40.	898	80.	845	120.	865	160.	905	200.	902

Chapter 4

There never was in the world two opinions alike, no more than two hairs or two grains; the most universal quality is diversity.

—Michel de Montaigne

I feel like a fugitive from the law of averages.

—Bill Mauldin

Measures of Dispersion

LEARNING OBJECTIVES

After studying this chapter, you should be able to

- provide the importance of the concept of variability (dispersion).
- measure the spread or dispersion, understand it, and identify its causes to provide a basis for action.

4.1 INTRODUCTION

Just as central tendency can be measured by a number in the form of an average, the amount of variation (dispersion, spread, or scatter) among the values in the data set can also be measured. The measures of central tendency describe that the major part of values in the data set appears to concentrate (cluster) around a central value called *average* with the remaining values scattered (spread or distributed) on either sides of that value. But these measures do not reveal how these values are dispersed (spread or scatter) on each side of the central value. The dispersion of values is indicated by the extent to which these values tend to spread over an interval rather than cluster closely around an average.

The statistical techniques to measure such dispersion are of two types:

(a) Techniques that are used to measure the extent of variation or the deviation (also called degree of variation) of each value in the data set from a measure of central tendency usually the mean or median. Such statistical techniques are called *measures of dispersion* (or *variation*).

(b) Techniques that are used to measure the direction (away from uniformity or symmetry) of variation in the distribution of values in the data set. Such statistical techniques are called *measures of skewness*, discussed in Chapter 5.

To measure the dispersion, understand it, and identify its causes is very important in statistical inference (estimation of parameter, hypothesis testing, forecasting, and so on). A small dispersion among values in the data set indicates that data are clusted closely around the mean. The mean is therefore considered representative of the data,

i.e. mean is a reliable average. Conversely, a large dispersion among values in the data set indicates that the mean is not reliable, i.e. it is not representative of the data.

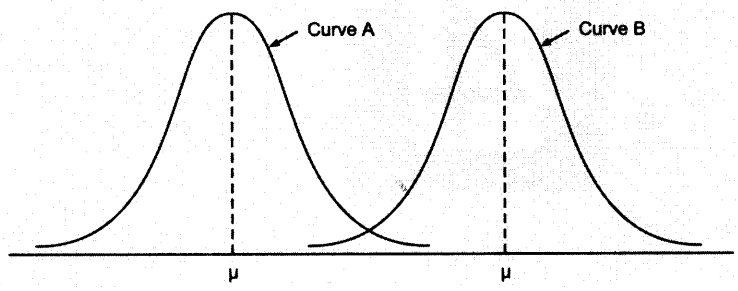


Figure 4.1
Symmetrical Distributions with Unequal Mean and Equal Standard Deviation

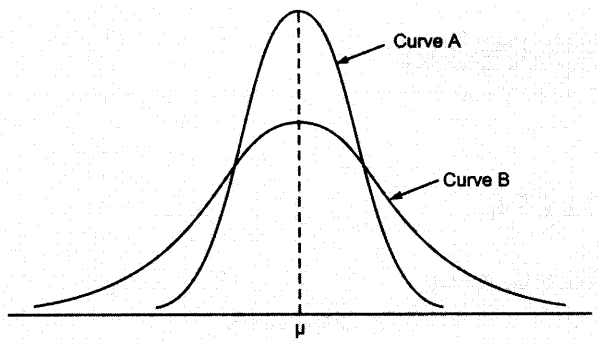


Figure 4.2
Symmetrical Distributions with Equal Mean and Unequal Standard Deviation

The symmetrical distribution of values in two or more sets of data may have same variation but differ greatly in terms of A.M. On the other hand, two or more sets of data may have the same A.M. values but differ in variation as shown in Fig. 4.2.

Illustration Suppose over the six-year period the net profits (in percentage) of two firms is as follows:

Firm 1	:	5.2,	4.5,	3.9,	4.7,	5.1,	5.4
Firm 2	:	7.8,	7.1,	5.3,	14.3,	11.0,	16.1

Since average amount of profit is 4.8 per cent for both firms, therefore operating results of both the firms are equally good and that a choice between them for investment purposes must depend on other considerations. However, the difference among the values is greater in Firm, 2, that is, profit is varying from 5.3 to 16.1 per cent, while the net profit values of Firm 1 were varying from 3.9 to 5.4 per cent. This shows that the values in data set 2 are spread more than those in data set 1. This implies that Firm 1 has a consistent performance while Firm 2 has a highly inconsistent performance. Thus for investment purposes a comparison of the average (mean) profit values alone should not be sufficient.

4.2 SIGNIFICANCE OF MEASURING DISPERSION (VARIATION)

Following are some of the purposes for which measures of variation are needed.

1. ***Test the reliability of an average:*** Measures of variation are used to test to what extent an average represents the characteristic of a data set. If the variation is small, that is, extent of dispersion or scatter is less on each side of an average, then it indicates high unformity of values in the distribution and the average represents an individual value in the data set. On the other hand, if the variation is large, then it indicates a lower degree of uniformity in values in the data set, and the average may be unreliable. No variation indicates perfect uniformity and, therefore, values in the data set are identical.
2. ***Control the variability:*** Measuring of variation helps to identify the nature and causes of variation. Such information is useful in controlling the variations. According to Spurr and Bonini, *'In matters of health, variations, in body temperature, pulse beat and*

blood pressure are the basic guides to diagnosis. Prescribed treatment is designed to control their variation. In industrial production efficient operation requires control of quality variation, the causes of which are sought through inspection and quality control programmes.' In social science, the measurement of 'inequality' of distribution of income and wealth requires the measurement of variability.

3. ***Compare two or more sets of data with respect to their variability:*** Measures of variation help in the comparison of the spread in two or more sets of data with respect to their uniformity or consistency. For example, (i) the measurement of variation in share prices and their comparison with respect to different companies over a period of time requires the measurement of variation, (ii) the measurement of variation in the length of stay of patients in a hospital every month may be used to set staffing levels, number of beds, number of doctors, and other trained staff, patient admission rates, and so on.
4. ***Facilitate the use of other statistical techniques:*** Measures of variation facilitate the use of other statistical techniques such as correlation and regression analysis, hypothesis testing, forecasting, quality control, and so on.

4.2.1 Essential Requisites for a Measure of Variation

The essential requisites for a good measure of variation are listed below. These requisites help in identifying the merits and demerits of individual measure of variation.

(i) It should be rigidly defined.
(ii) It should be based on all the values (elements) in the data set.
(iii) It should be calculated easily, quickly, and accurately.
(iv) It should not be unduly affected by the fluctuations of sampling and also by extreme observations.
(v) It should be amenable to further mathematical or algebraic manipulations.

4.3 CLASSIFICATION OF MEASURES OF DISPERSION

The various measures of dispersion (variation) can be classified into two categories:

(i) Absolute measures, and
(ii) Relative measures

Absolute measures are described by a number or value to represent the amount of variation or differences among values in a data set. Such a number or value is expressed in the same unit of measurement as the set of values in the data such as rupees, inches, feet, kilograms, or tonnes. Such measures help in comparing two or more sets of data in terms of absolute magnitude of variation, provided the variable values are expressed in the same unit of measurement and have almost the same average value.

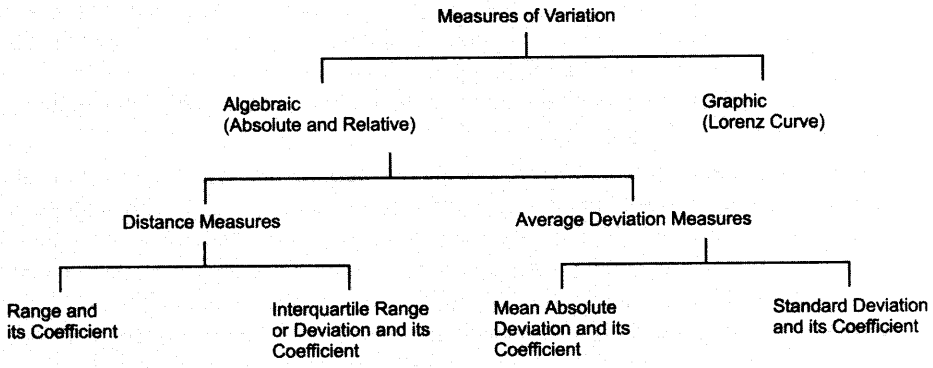
The *relative measures* are described as the ratio of a measure of absolute variation to an average and is termed as *coefficient of variation*. The word 'coefficient' means a number that is independent of any unit of measurement. While computing the relative variation, the average value used as base should be the same from which the absolute deviations were calculated.

Another classification of the measures of variation is based on the method employed for their calculations:

(i) Distance measures, and
(ii) Average deviation measures

The *distance measures* describe the spread or dispersion of values of a variable in terms of difference among values in the data set. The *average deviation measures* describe the average deviation for a given measure of central tendency.

The above-mentioned classification of various measures of dispersion (variation) may be summarized as shown below:



4.4 DISTANCE MEASURES

As mentioned above, two distance measures discussed in this section are namely:

(i) Range, and
(ii) Interquartile deviation

4.4.1 Range

Range: A measure of variability, defined to be the difference between the largest and lowest values in the data set.

The range is the most simple measure of dispersion and is based on the location of the largest and the smallest values in the data. Thus the **range** *is defined to be the difference between the largest and lowest observed values in a data set.* In other words, it is the length of an interval which covers the highest and lowest observed values in a data set and thus measures the dispersion or spread within the interval in the most direct possible way.

$$\text{Range (R)} = \text{Highest value of an observation} - \text{Lowest value of an observation}$$
$$= H - L \qquad (4\text{-}1)$$

For example, if the smallest value of an observation in the data set is 160 and largest value is 250, then the range is 250 – 160 = 90.

For grouped frequency distributions of values in the data set, the range is the difference between the upper class limit of the last class and the lower class limit of first class. In this case the range obtained may be higher than as compared to ungrouped data because of the fact that the class limits are extended slightly beyond the extreme values in the data set.

Coefficient of Range

The relative measure of range, called the coefficient of range is obtained by applying the following formula:

$$\text{Coefficient of range} = \frac{H - L}{H + L} \qquad (4\text{-}2)$$

Example 4.1: The following are the sales figures of a firm for the last 12 months

Months	:	1	2	3	4	5	6	7	8	9	10	11	12
Sales (Rs '000)	:	80	82	82	84	84	86	86	88	88	90	90	92

Calculate the range and coefficient of range for sales.

Solution: Given that H = 92 and L = 80. Therefore

$$\text{Range} = H - L = 92 - 80 = 12$$

and $$\text{Coefficient of range} = \frac{H - l}{H + L} = \frac{92 - 80}{92 + 80} = \frac{12}{172} = 0.069$$

Example 4.2: The following data show the waiting time (to the nearest 100th of a minute) of telephone calls to be matured:

Waiting Time	*Frequency*	*Waiting Time*	*Frequency (Minutes)*
0.10–0.35	6	0.88–1.13	8
0.36–0.61	10	1.14–1.39	4
0.62–0.87	8		

Calculate the range and coefficient of range.

Solution: Given that, H = 1.39 and L = 0.10. Therefore

$$\text{Range} = H - L = 1.39 - 0.10 = 1.29 \text{ minutes}$$

and $$\text{Coefficient of Range} = \frac{H-l}{H+L} = \frac{1.39-0.10}{1.39+0.10} = \frac{1.29}{1.49} = 0.865$$

Advantages, Disadvantages and Applications of Range The major advantages and disadvantages of range may be summarized as follows:

Advantages

(i) It is independent of the measure of central tendency and easy to calculate and understand.

(ii) It is quite useful in cases where the purpose is only to find out the extent of extreme variation, such as industrial quality control, temperature, rainfall, and so on.

Disadvantages

(i) The calculation of range is based on only two values—largest and smallest in the data set and fail to take account of any other observations.

(ii) It is largely influenced by two extreme values and completely independent of the other values. For example, range of two data sets {1, 2, 3, 7, 12} and {1, 1, 1, 12, 12} is 11, but the two data sets differ in terms of overall dispersion of values

(iii) Its value is sensitive to changes in sampling, that is, different samples of the same size from the same population may have widely different ranges.

(iv) It cannot be computed in case of open-end frequency distributions because no highest or lowest value exists in open-ended class.

(v) It does not describe the variation among values in the data between two extremes. For example, each of the following set of data

Set 1	:	9	21	21	21	21	21	21	21
Set 2	:	9	9	9	9	21	21	21	21
Set 3	:	9	10	12	14	15	19	20	21

has a range of 21 – 9 = 12, but the variation of values is quite different in each case between the hightest and lowest values.

Applications of Range

(i) *Fluctuation in share prices:* The range is useful in the study of small variations among values in a data set, such as variation in share prices and other commodities that are very sensitive to price changes from one period to another.

(ii) *Quality control:* It is widely used in industrial quality control. Quality control is exercised by preparing suitable *control charts*. These charts are based on setting an upper control limit (range) and a lower control limit (range) within which produced items shall be accepted. The variation in the quality beyond these ranges requires necessary correction in the production process or system.

(iii) *Weather forecasts:* The concept of range is used to determine the difference between maximum and minimum temperature or rainfall by meteorological departments to announce for the knowledge of the general public.

4.4.2 Interquartile Range or Deviation

Interquartile range: A measure of variability, defined to be the difference between the quartiles Q_3 and Q_1.

The limitations or disadvantages of the range can partially be overcome by using another measure of variation which measures the spread over the middle half of the values in the data set so as to minimise the influence of outliers (extreme values) in the calculation of range. Since a large number of values in the data set lie in the central part of the frequency distribution, therefore it is necessary to study the **Interquartile Range** (also called midspread). To compute this value, the entire data set is divided into four parts each of which contains 25 per cent of the observed values. The quartiles are the highest values in each of these four parts. The *interquartile range* is *a measure of dispersion or spread of values in the data set between the third quartile, Q_3 and the first quartile, Q_1*. In other words, the *interquartile range or deviation* (IQR) is the range for the middle 50 per cent of the data. The concept of IQR is shown in Fig. 4.3:

$$\text{Interquartile range (IQR)} = Q_3 - Q_1 \tag{4-3}$$

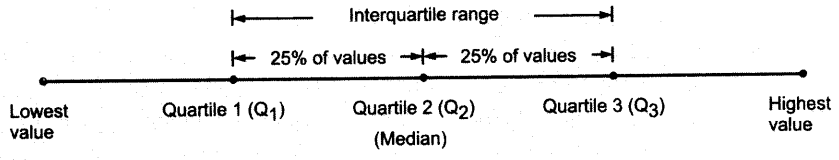
Half the distance between Q_1 and Q_3 is called the *semi-interquartile range* or the *quartile deviation* (QD).

$$\text{Quartile deviation (QD)} = \frac{Q_3 - Q_1}{2} \tag{4-4}$$

The median is not necessarily midway between Q_1 and Q_3, although this will be so for a symmetrical distribution. The median and quartiles divide the data into equal numbers of values but do not necessarily divide the data into equally wide intervals.

As shown above the quartile deviation measures the average range of 25 per cent of the values in the data set. It represents the spread of all observed values because its value is computed by taking an average of the middle 50 per cent of the observed values rather than of the 25 per cent part of the values in the data set.

Figure 4.3
Interquartile Range



In a non-symmetrical distribution, the two quartiles Q_1 and Q_3 are at equal distance from the median, that is, Median $- Q_1 = Q_3 -$ Median. Thus, Median $\pm$ Quartile Deviation covers exactly 50 per cent of the observed values in the data set.

A smaller value of quartile deviation indicates high uniformity or less variation among the middle 50 per cent observed values around the median value. On the other hand, a high value of quartile deviation indicates large variation among the middle 50 per cent observed values.

Coefficient of Quartile Deviation

Since quartile deviation is an absolute measure of variation, therefore its value gets affected by the size and number of observed values in the data set. Thus, the Q.D. of two or more than two sets of data may differ. Due to this reason, to compare the degree of variation in different sets of data, we compute the relative measure corresponding to Q.D., called the *coefficient of Q.D.*, and it is calculated as follows:

$$\text{Coefficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \tag{4-5}$$

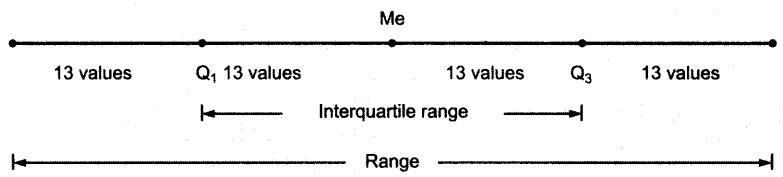
Example 4.3: Following are the responses from 55 students to the question about how much money they spent every day.

55	60	80	80	80	85	85	85	90	90	90
90	92	94	95	95	95	95	100	100	100	100
100	100	105	105	105	105	109	110	110	110	110
112	115	115	115	115	115	120	120	120	120	120
124	125	125	125	130	130	140	140	140	145	150

Calculate the range and interquartile range and interpret your result.

Solution: The median of the given values in the data set is the (55 + 1)/2 = 28th value which is 105. From this middle value of 105, there are 27 values at or below of 105 and another 27 at or above of 105.

The lower quartile of Q_1 = (27 + 1)/2 = 14th value from bottom of the data i.e. Q_1 = 94 and upper quartile is the 14th value from the top, i.e. Q_3 = 120. The 55 values have been partitioned as follows:



The interquartile range is, IQR = 120 – 94 = 26 while the range is R = 150 – 55 = 95. The middle 50% of the data fall in relatively narrow range of only Rs 26. This means responses are more densely clustered near the centre of the data and more spread out towards the extremes. For instance, lowest 25% of the students had responses, ranging over 55 to 94, i.e. Rs 39, while the next 25% had responses ranging over 94 to 105, i.e. only Rs 11. Similarly, the third quarter had responses from 105 to 110, i.e. only Rs 5, while the top 25% had responses in the interval (120 to 150), i.e. Rs 30.

The median and quartiles divide the data into equal numbers of values but not necessarily divide the data into equally wide intervals.

Example 4.4: Use an appropriate measure to evaluate the variation in the following data:

Farm Size (acre)	*No. of Farms*	*Farm Size (acre)*	*No. of Farms*
below 40	394	161–200	169
41–80	461	201–240	113
81–120	391	241 and above	148
121–160	334		

Solution: Since the frequency distribution has open-end class intervals on the two extreme sides, therefore Q.D. would be an appropriate measure of variation. The computation of Q.D. is shown in Table 4.1.

Table 4.1 Calculations of Quartile Deviation

Farm Size (acre)	*No. of Farms*	*Cumulative Frequency (cf) (less than)*	
below 40	394	394	
41–80	461	855	← Q_1 class
81–120	391	1246	
121–160	334	1580	← Q_3 class
161–200	169	1749	
201–240	113	1862	
241 and above	148	2010	
	2010		

Q_1 = Value of $(n/4)$th observation = 2010 ÷ 4 or 502.5th observation

This observation lies in the class 41–80. Therefore

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

$$= 41 + \frac{502.5 - 394}{461} \times 40 = 41 + 9.41 = 50.41 \text{ acres}$$

Q_3 = Value of $(3n/4)$th observation = (3 × 2010) ÷ 4 or 1507.5th observation

This observation lies in the class 121–160. Therefore

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

$$= 121 + \frac{1507.5 - 1246}{334} \times 40 = 121 + 31.31 = 152.31 \text{ acres}$$

Thus the quartile deviation is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{152.31 - 50.41}{2} = 50.95 \text{ acres}$$

and $$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{50.95}{202.72} = 0.251$$

Advantages and Disadvantages of Quartile Deviation The major advantages and disadvantages of quartile deviation are summarized as follows:

Advantages

(i) It is not difficult to calculate but can only be used to evaluate variation among observed values within the middle of the data set. Its value is not affected by the extreme (highest and lowest) values in the data set.
(ii) It is an appropriate measure of variation for a data set summarized in open-end class intervals.
(iii) Since it is a positional measure of variation, therefore it is useful in case of erratic or highly skewed distributions, where other measures of variation get affected by extreme values in the data set.

Disadvantages

(i) The value of Q.D. is based on the middle 50 per cent observed values in the data set, therefore it cannot be considered as a good measure of variation as it is not based on all the observations.
(ii) The value of Q.D. is very much affected by sampling fluctuations.
(iii) The Q.D. has no relationship to any particular value or an average in the data set for measuring the variation. Its value is not affected by the distribution of the individual values within the interval of the middle 50 per cent observed values.

Conceptual Questions 4A

1. Explain the term variation. What does a measure of variation serve? In the light of these, comment on some of the well-known measures of variation.
[Delhi Univ., MBA, 2001]
2. What are the requisites of a good measure of variation?
3. Explain how measures of central tendency and measures of variation are complementary to each other in the context of analysis of data.
4. Distinguish between absolute and relative measures of variation. Give a broad classification of the measures of variation.
5. (a) Critically examine the different methods of measuring variation.
(b) Explain with suitable examples the term 'variation'. Mention some common measures of variation and describe the one which you think is the most important. *[Delhi Univ., MBA, 1998]*
6. Explain and illustrate how the measures of variation afford a supplement to the information about frequency distribution furnished by averages.
[Delhi Univ., MBA, 1999]
7. What do you understand by 'coefficient of variation'? Discuss its importance in business problems.

Self-Practice Problems 4A

4.1 The following are the prices of shares of a company from Monday to Saturday:

Days	*Price (Rs)*	*Days*	*Price (Rs)*
Monday	200	Thursday	160
Tuesday	210	Friday	220
Wednesday	208	Saturday	250

Calculate the range and its coefficient.

4.2 The days sales figures (in Rs) for the last 15 days at Nirula's ice-cream counter, arranged in ascending order of magnitude, is recorded as follows: 2000, 2000, 2500, 2500, 2500, 3500, 4000, 5300, 9000, 12,500, 13,500, 24,500, 27,100, 30,900, and 41,000. Determine the range and middle 50 per cent range for this sample data.

4.3 The following distribution shows the sales of the fifty largest companies for a recent year:

Sales (Million of rupees)	*Number of Companies*
0– 9	18
10–19	19
20–29	6
30–39	2
40–49	5

Calculate the coefficient of range

4.4 You are given the frequency distribution of 292 workers of a factory according to their average weekly income.

Weekly Income (Rs)	*No. of Workers*	*Weekly Income (Rs)*	*No. of Workers*
Below 1350	8	1450–1470	22
1350–1370	16	1470–1490	15
1370–1390	39	1490–1510	15
1390–1410	58	1510–1530	9
1410–1430	60	1530 and above	10
1430–1450	40		

Calculate the quartile deviation and its coefficient from the above mentioned data.

[Kurukshetra Univ., MBA, 1998]

4.5 You are given the data pertaining to kilowatt hours of electricity consumed by 100 persons in a city.

Consumption (kilowatt hour)	*No. of Users*
0–10	6
10–20	25
20–30	36
30–40	20
40–50	13

Calculate the range within which the middle 50 per cent of the consumers fall.

4.6 The following sample shows the weekly number of road accidents in a city during a two-year period:

Number of Accidents	*Frequency*	*Number of Accidents*	*Frequency*
0– 4	5	25–29	9
5– 9	12	30–34	4
10–14	32	35–39	3
15–19	27	40–44	1
20–24	11		

Find the interquartile range and coefficient of quartile deviation.

4.7 A City Development Authority subdivided the available land for housing into the following building lot sizes:

Lot Size (Square meters)	*Frequency*
Below 69.44	19
69.44–104.15	25
104.16–208.32	42
208.33–312.49	12
312.50–416.65	5
416.66 and above	17

Find the interquartile range and quartile deviation.

4.8 The cholera cases reported in different hospitals of a city in a rainy season are given below:
Calculate the quartile deviation for the given distribution and comment upon the meaning of your result.

Age Group (Years)	*Frequency*	*Age Group*	*Frequency*
Less than 1	15	25–35	132
1–5	113	35–45	65
5–10	122	45–65	46
10–15	91	65 and above	15
15–25	229		

Hints and Answers

4.1 Range = Rs 90, Coefficient of range = 0.219

4.2 Range = Rs 39,000;

Middle 50%, R $= P_{75} - P_{25}$

$= x_{(75n/100) + (1/2)} - x_{(25n/100) + (1/2)}$

$= x_{(11.25 + 0.50)} - x_{(3.75 + 0.50)}$

$= x_{11.75} - x_{4.25}$

$= (13{,}500 + 8250) - (2500 + 00)$

$= 19{,}250$

Here $x_{11.75}$ is the interpolated value for the 75% of the distance between 11th and 12th ordered sales amount. Similarly, $x_{4.25}$ is the interpolated value for the 25% of the distance between 4th and 5th order sales amount.

4.3 Coefficient of range = 1

4.4 Quartile deviation = 27.76; Coeff. of Q.D. = 0.020; Q_1 = 1393.48; Q_3 = 1449

4.5 $Q_3 - Q_1 = 34 - 17.6 = 16.4$

4.6 $Q_3 - Q_1$ = 30.06; Coefficient of Q.D. = 0.561

4.7 $Q_3 - Q_1$ = 540.26; Q.D. = 270.13

4.8 Q.D. = 10 years

4.5 AVERAGE DEVIATION MEASURES

The range and quartile deviation indicate overall variation in a data set, but do not indicate spread or scatteredness around the centrilier (i.e. mean, median or mode). However, to understand the nature of distribution of values in the data set, we need to measure the 'spread' of values around the mean to indicate how representative the mean is.

In this section, we shall discuss two more measures of dispersion to measure the mean (or average) amount by which all values in a data set (population or sample) vary from their mean. These measures deal with the average deviation from some measure of central tendency—usually mean or median. These measures are:

(a) Mean Absolute Deviation or Average Deviation
(b) Variance and Standard Deviation

4.5.1 Mean Absolute Deviation

Since two measures of variation, range and quartile deviation, discussed earlier do not show how values in a data set are scattered about a central value or disperse themselves throughout the range, therefore it is quite reasonable to measure the variation as a degree (amount) to which values within a data set deviate from either mean or median.

The mean of deviations of individual values in the data set from their actual mean is always zero so such a measure (zero) would be useless as an indicator of variation. This problem can be solved in two ways:

(i) Ignore the signs of the deviations by taking their absolute value, or
(ii) Square the deviations because the square of a negative number is positive.

Since the absolute difference between a value x_i of an observation from A.M. is always a positive number, whether it is less than or more than the A.M., therefore we take the absolute value of each such deviation from the A.M. (or median). Taking the average of these deviations from the A.M., we get a measure of variation called the *mean absolute deviation* (MAD). In general, the mean absolute deviation is given by

$$\text{MAD} = \frac{1}{\text{N}} \sum_{i=1}^{\text{N}} |x - \mu|, \quad \text{for a population} \tag{4-6}$$

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |x - \bar{x}|, \quad \text{for a sample}$$

where | | indicates the absolute value. That is, the signs of deviations from the mean are disregarded.

For a grouped frequency distribution, MAD is given by

$$\text{MAD} = \frac{\sum_{i=1}^{n} f_i |x_i - \bar{x}|}{\sum f_i} \tag{4-7}$$

Formulae (4-6) and (4-7), in different contexts, indicates that the MAD provides a useful method of comparing the relative tendency of values in the distribution to scatter around a central value or to disperse themselves throughout the range.

While calculating the mean absolute deviation, the median is also considered for computing because the sum of the absolute values of the deviations from the median is smaller than that from any other value. However, in general, arithmetic mean is used for this purpose.

If a frequency distribution is symmetrical, then A.M. and median values concide and the same MAD value is obtained. In such a case $\bar{x} \pm$ MAD provides a range in which 57.5 per cent of the observations are included. Even if the frequency distribution

is moderately skewed, the interval $\bar{x} \pm$ MAD includes the same percentage of observations. This shows that more than half of the observations are scattered within one unit of the MAD around the arithmetic mean.

The MAD is useful in situations where occasional large and erratic deviations are likely to occur. The standard deviation, which uses the squares of these large deviations, tends to over-emphasize them.

Coefficient of MAD

The relative measure of mean absolute deviation (MAD) called the *coefficient of MAD* is obtained by dividing the MAD by a measure of central tendency (arithmetic mean or median) used for calculating the MAD. Thus

$$\text{Coefficient of MAD} = \frac{\text{Mean absolute deviation}}{\bar{x} \text{ or Me}} \tag{4-8}$$

If the value of relative measure is desired in percentage, then

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

Example 4.5: The number of patients seen in the emergency ward of a hospital for a sample of 5 days in the last month were: 153, 147, 151, 156 and 153. Determine the mean deviation and interpret.

Solution: The mean number of patients is, $\bar{x} = (153 + 147 + 151 + 156 + 153)/5 = 152$. Below are the details of the calculations of MAD using formula (4–6).

Numer of Patients (x)	$x - \bar{x}$	*Absolute Deviation* $\lvert x - \bar{x} \rvert$
153	153 – 152 = 1	1
147	147 – 152 = –5	5
151	151 – 152 = –1	1
156	156 – 152 = 4	4
153	153 – 152 = 1	1
		12

$$\text{MAD} = \frac{1}{n}\Sigma\lvert x - \bar{x}\rvert = \frac{12}{5} = 2.4 \cong 3 \text{ patients (approx)}$$

The mean absolute deviation is 3 patients per day. The number of patients deviate on the average by 3 patients from the mean of 152 patients per day.

Example 4.6: Calculate the mean absolute deviation and its coefficient from median for the following data

Year	*Sales (Rs thousand)*	
	Product A	*Product B*
1996	23	36
1997	41	39
1998	29	36
1999	53	31
2000	38	47

Solution: The median sales (Me) of the two products A and B is Me = 38 and Me = 36, respectively. The calculations of MAD in both the cases are shown in Table 4.2.

Table 4.2 Calculations of MAD

Product A		*Product B*	
Sales (x)	$\lvert x - Me\rvert = \lvert x - 38\rvert$	*Sales (x)*	$\lvert x - Me\rvert = \lvert x - 36\rvert$
23	15	31	5
29	9	36	0
38	0	36	0
41	3	39	3
53	15	47	11
$n = 5$	$\sum \lvert x - Me\rvert = 42$	$n = 5$	$\sum \lvert x - Me\rvert = 19$

Product A:

$$MAD = \frac{1}{n}\sum |x - Me| = \frac{42}{5} = 8.4$$

$$\text{Coefficient of MAD} = \frac{MAD}{Me} = \frac{8.4}{38} = 0.221$$

Product B:

$$MAD = \frac{1}{n}\sum |x - Me| = \frac{19}{5} = 3.8$$

$$\text{Coefficient of MAD} = \frac{MAD}{Me} = \frac{3.8}{36} = 0.106$$

Example 4.7: Find the mean absolute deviation from mean for the following frequency distribution of sales (Rs in thousand) in a co-operative store.

Sales	:	50–100	100–150	150–200	200–250	250–300	300–350
Number of days	:	11	23	44	19	8	7

Solution: The mean absolute deviation can be calculated by using the formula (4-6) for mean. The calculations for MAD are shown in Table 4.3. Let the assumed mean be, A = 175.

Table 4.3 Calculations for MAD

Sales (Rs)	*Mid-Value (m)*	*Frequency (f)*	$(m - 175)/50$ $(= d)$	fd	$\lvert x - \bar{x}\rvert$ $= \lvert x - \bar{x}\rvert$	$f\lvert x - \bar{x}\rvert$
50–100	75	11	−2	−22	104.91	1154.01
100–150	125	23	−1	−23	54.91	1262.93
150–200	(175) ← A	44	0	0	4.91	216.04
200–250	225	19	1	19	45.09	856.71
250–300	275	8	2	16	95.09	760.72
300–350	325	7	3	21	145.09	1015.63
		112		11		5266.04

$$\bar{x} = A + \frac{\sum fd}{\sum f} \times h = 175 + \frac{11}{112} \times 50 = \text{Rs } 179.91 \text{ per day}$$

$$MAD = \frac{\sum f\lvert x - \bar{x}\rvert}{\sum f} = \frac{5266.04}{112} = \text{Rs } 47.01$$

Thus the average sales is Rs 179.91 thousand per day and the mean absolute deviation of sales is Rs 47.01 thousand per day.

Example 4.8: A welfare organization introduced an education scholarship scheme for school going children of a backward village. The rates of scholarship were fixed as given below:

Age Group (Years)	*Amount of Scholarship per Month (Rs)*
5–7	300
8–10	400
11–13	500
14–16	600
17–19	700

The ages of 30 school children are noted as; 11, 8, 10, 5, 7, 12, 7, 17, 5, 13, 9, 8, 10, 15, 7, 12, 6, 7, 8, 11, 14, 18, 6, 13, 9, 10, 6, 15, 3, 5 years respectively. Calculate mean and standard deviation of monthly scholarship. Find out the total monthly scholarship amount being paid to the students. *[IGNOU, MBA, 2002]*

Solution: The number of students in the age group from 5–7 to 17–19 are calculated as shown in table 4.4:

Table 4.4

Age Group (Years)	*Tally Bars*	*Number of Students*
5–7	𝍸 𝍸	10
8–10	𝍸 |||	8
11–13	𝍸 ||	7
14–16	|||	3
17–19	||	2
		30

The calculations for mean and standard deviation are shown in Table 4.5.

Table 4.5 Calculations for Mean and Standard Deviation

Age Group (Years)	*Number of Students (f)*	*Mid-value (m)*	$d = \frac{m-A}{h} = \frac{m-12}{3}$	*fd*	fd^2
5–7	10	6	–2	–20	40
8–10	8	9	–1	–8	8
11–13	7	A → (12)	0	0	0
14–16	3	15	1	3	3
17–19	2	18	2	4	8
	30			–21	59

$$\text{Mean, } \bar{x} = A + \frac{\Sigma fd}{\Sigma f} \times h = 12 - \frac{21}{30} \times 3 = 12 - 2.1 = 9.9$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times h = \sqrt{\frac{59}{30} - \left(\frac{-21}{30}\right)^2} \times 3$$

$$= \sqrt{1.967 - 0.49} \times 3 = 1.2153 \times 3 = 3.6459$$

Calculations for monthly scholarship paid to 30 students are shown in Table 4.6.

Table 4.6 Calculations for Monthly Scholarship

Number of Students	*Amount of Scholarship per Month (Rs)*	*Total Monthly Scholarship (Rs)*
10	300	3000
8	400	3200
7	500	3500
3	600	1800
2	700	1400
		12,900

Advantages and Disadvantages of MAD The advantages and disadvantages of MAD are summarized below:

Advantages

(i) The calculation of MAD is based on all observations in the distribution and shows the dispersion of values around the measure of central tendency.

(ii) The value of MAD is easy to compute and therefore makes it popular among those users who are not even familiar with statistical methods.

(iii) While calculating MAD, equal weightage is given to each observed value and thus it indicates how far each observation lies from either the mean or median.

(iv) Average deviation from mean is always zero in any data set. The MAD avoids this problem by using absolute values to eliminate the negative signs.

Disadvantages

(i) The algebraic signs are ignored while calculating MAD. If the signs are not ignored, then the sum of the deviations taken from arithmetic mean will be zero and close to zero when deviations are taken from median.

(ii) The value of MAD is considered to be best when deviations are taken from median. However, median does not provide a satisfactory result in case of a high degree of variability in a data set.

Moreover, the sum of the deviations from mean (ignoring signs) is greater than the sum of the deviations from median (ignoring signs). In such a situation, computations of MAD by taking deviations from mean is also not desirable.

(iii) The MAD is generally unwieldy in mathematical discussions.

Inspite of all these demerits, the knowledge of MAD would help the reader to understand another important measure of dispersion called the *standard deviation*.

4.5.2 Variance and Standard Deviation

Another way to disregard the signs of negative deviations from mean is to square them. Instead of computing the absolute value of each deviation from mean, we square the deviations from mean. Then the sum of all such squared deviations is divided by the number of observations in the data set. This value is a measure called **population variance** and is denoted by σ^2 (a lower-case Greek letter sigma). It is usually referred to as 'sigma squared'. Symbolically, it is written as:

Variance: A measure of variability based on the squared deviations of the observed values in the data set about the mean value.

$$\text{Population variance, } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{4-9}$$

$$= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - (\mu)^2$$

(Deviation is taken from actual population A.M.)

$$= \frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2$$ (Deviation is taken from assumed A.M.)

where $d = x - A$ and A is any constant (also called assumed A.M.)

Since σ^2 is the average or mean of squared deviations from arithmetic mean, it is also called the *mean square average*.

The population variance is basically used to measure variation among the values of observations in a population. Thus for a population of N observations (elements) and with μ denoting the population mean, the formula for population variance is shown in Eqn. (4-9). However, in almost all applications of statistics, the data being analyzed is a sample data. As a result, population variance is rarely determined. Instead, we compute a sample variance to estimate population variance, σ^2.

It was shown that if the *sum of the squared* deviations about a sample mean $\bar{x}$ in Eqn. (4-9) is divided by n (sample size), then it invariably tends to cause the resulting estimate of σ^2 to be lower than its actual value. This undesirable condition is called *bias*. However, this *bias* in the estimation of population variance from a sample can be removed by dividing the sum of the squared deviations between the sample mean and each element in the population by $n - 1$ rather than by n. Thus the *unbiased* sample variance denoted by s^2 is defined as follows:

$$\text{Sample variance, } s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{\Sigma x^2}{n-1} - \frac{n\bar{x}^2}{n-1} = \frac{\Sigma x^2}{n-1} - \frac{(\Sigma x)^2}{n(n-1)} \tag{4-10}$$

The numerator $\Sigma(x-\bar{x})^2$ in Eqn. (4-10) is called the *total sum of squares*. This quantity measures the total variation among values in a data set (whereas the variance measures only the *average variation*). The larger the value of $\Sigma(x-\bar{x})^2$, the greater the variation among the values in a data set.

Standard Deviation

The numerical value of population or a sample variance is difficult to interpret because it is expressed in square units. To reach a interpretable measure of variance expressed in the units of original data, we take a positive square root of the variance, which is known as the **standard deviation** or *root-mean square deviation*. The standard deviation of population and sample is denoted by σ and s, respectively. We can think of the standard deviation as roughly the *average distance values fall from the mean.*

Standard deviation: A measure of variability computed by taking the positive square root of the variance.

(a) *Ungrouped Data*

$$\text{Population standard deviation, } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\Sigma(x-\mu)^2} = \sqrt{\frac{1}{N}\Sigma x^2 - (\mu)^2}$$

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

$$\text{Sample standard deviation, } s = \sqrt{\frac{\Sigma x^2}{n-1} - \frac{n\bar{x}^2}{n-1}}\text{; where } n = \text{sample size}$$

(b) *Grouped Data*

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

where f = frequency of each class interval

$N = \Sigma f$ = total number of observations (or elements) in the population

h = width of class interval

m = mid-value of each class interval

$d = \dfrac{m - A}{h}$, where A is any constant (also called assumed A.M.)

$$\text{Sample standard deviation, } s = \sqrt{s^2} = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma fx^2}{n-1} - \frac{(\Sigma fx)^2}{n(n-1)}} \tag{4-11}$$

Remarks: 1. For any data set, MAD is always less than the σ because MAD is less sensitive to the extreme observations. Thus when a data contains few very large observations, the MAD provides a more realistic measure of variation than σ. However σ is often used in statistical applications because it is amenable to mathematical development.

2. When sample size (n) becomes very large, ($n - 1$) becomes indistinguishable and becomes irrelevant.

Advantages and Disadvantages of Standard Deviation The advantages and disadvantages of the standard deviation are summarized below:

Advantages

(i) The value of standard deviation is based on every observation in a set of data. It is the only measure of variation capable of algebraic treatment and less affected by fluctuations of sampling as compared to other measures of variation.

(ii) It is possible to calculate the combined standard deviation of two or more sets of data.

(iii) Standard deviation has a definite relationship with the area under the symmetric curve of a frequency distribution. Due to this reason, standard deviation is called a *standard* measure of variation.

(iv) Standard deviation is useful in further statistical investigations. For example, standard deviation plays a vital role in comparing skewness, correlation, and so on, and also widely used in sampling theory.

Disadvantages

(i) As compared to other measures of variation, calculations of standard deviation are difficult.

(ii) While calculating standard deviation, more weight is given to extreme values and less to those near mean. Since for calculating S.D., the deviations from the mean are squared, therefore large deviations when squared are proportionately more than small deviations. For example, the deviations 2 and 10 are in the ratio of 1 : 5 but their squares 4 and 100 are in the ratio of 1 : 25.

Example 4.9: The wholesale prices of a commodity for seven consecutive days in a month is as follows:

Days	:	1	2	3	4	5	6	7
Commodity price/quintal	:	240	260	270	245	255	286	264

Calculate the variance and standard deviation.

Solution: The computations for variance and standard deviation are shown in Table 4.7.

Table 4.7 Computations of Variance and Standard Deviation by Actual Mean Method

Observation (x)	$x - \bar{x} = x - 260$	$(x-\bar{x})^2$
240	− 20	400
260	0	0
270	10	100
245	− 15	225
255	− 5	25
286	26	676
264	4	16
1820		1442

$$\bar{x} = \frac{\Sigma x}{N} = \frac{1820}{7} = 260$$

$$\text{Variance } \sigma^2 = \frac{\Sigma(x-\bar{x})}{N} = \frac{1442}{7} = 206$$

$$\text{Standard deviation } \sigma = \sqrt{\sigma^2} = \sqrt{206} = 14.352$$

In this question, if we take deviation from an assumed A.M. = 255 instead of actual A.M. = 260. The calculations then for standard deviation will be as shown in Table 4.8.

Table 4.8 Computation of Standard Deviation by Assumed Mean Method

Observation (x)	$d = x - A = x - 255$	d^2
240	- 15	225
260	5	25
270	15	225
245	- 10	100
(255) ← A	0	0
286	31	961
264	9	81
	35	1617

$$\text{Standard deviation } \sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} = \sqrt{\frac{1617}{7} - \left(\frac{35}{7}\right)^2}$$

$$= \sqrt{231 - 25} = \sqrt{206} = 14.352$$

This result is same as obtained earlier in Table 4.7.

Remark: When actual A.M. is not a whole number, assumed A.M. method should be used to reduce the computation time.

Example 4.10: A study of 100 engineering companies gives the following information

Profit (Rs in crore) :	0–10	10–20	20–30	30–40	40–50	50–60
Number of companies :	8	12	20	30	20	10

Calculate the standard deviation of the profit earned.

Solution: Let assumed mean, A be 35 and the value of h be 10. Calculations for standard deviation are shown in Table 4.9.

Table 4.9 Calculations of Standard Deviation

Profit (Rs in crore)	*Mid-value (m)*	$d = \frac{m-A}{h} = \frac{m-35}{10}$	f	fd	fd^2
0–10	5	- 3	8	- 24	72
10–20	15	- 2	12	- 24	48
20–30	25	- 1	20	- 20	20
30–40	(35) ← A	0	30	0	0
40–50	45	1	20	20	20
50–60	55	2	10	20	40
				- 28	200

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

$$= \sqrt{\frac{200}{100} - \left(\frac{-28}{100}\right)^2} \times 10 = \sqrt{2 - 0.078} \times 10 = 13.863$$

Example 4.11: Mr. Gupta, a retired government servant is considering investing his money in two proposals. He wants to choose the one that has higher average net present value and lower standard deviation. The relevant data are given below. Can you help him in choosing the proposal?

Proposal A:	*Net Present Value (NPV)*	*Chance of the Possible Outcome of NPV*
	1559	0.30
	5662	0.40
	9175	0.30

Proposal B:	*Net Present Value (NPV)*	*Chance of the Possible Outcome of NPV*
	– 10,050	0.30
	5,812	0.40
	20,584	0.30

Solution: To suggest to Mr. Gupta a proposal for high average net present value, first calculate the expected (average) net present value for both the proposals.

Proposal A: Expected NPV = $1559 \times 0.30 + 5662 \times 0.40 + 9175 \times 0.30$
$= 467.7 + 2264.8 + 2752.5 =$ Rs 5485

Proposal B: Expected NPV = $-10{,}050 \times 0.30 + 5812 \times 0.40 + 20{,}584 \times 0.30$
$= -3015 + 2324.8 + 6175.2 =$ Rs 5485

Since the expected NPV in both the cases is same, he would like to choose the less risky proposal. For this we have to calculate the standard deviation in both the cases.

Standard deviation for proposal A:

$NPV(x_i)$	*Expected NPV* $(\bar{x})$	$x - \bar{x}$	f	$f(x-\bar{x})^2$
1559	5485	– 3926	0.30	46,24,042.8
5662	5485	177	0.40	12,531.6
9175	5485	3690	0.30	40,84,830.0
			1.00	87,21,404.4

$$s_A = \sqrt{\frac{\Sigma f(x-x)^2}{\Sigma f}} = \sqrt{87,21,404.4} = \text{Rs } 2953.20$$

Standard deviation for proposal B:

$NPV(x_i)$	*Expected NPV* $(\bar{x})$	$x - \bar{x}$	f	$f(x-\bar{x})^2$
– 10,050	5485	– 15,535	0.30	7,24,00,867.5
5812	5485	327	0.40	42,771.6
20,584	5485	15,099	0.30	6,83,93,940
			1.00	14,08,37,579

$$s_B = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} = \sqrt{14,08,37,579} = \text{Rs } 11,867.50$$

The $s_A < s_B$ indicates uniform net profit for proposal A. Thus proposal A may be chosen.

4.5.3 Mathematical Properties of Standard Deviation

1. **Combined standard deviation:** The combined standard deviation of two sets of data containing n_1 and n_2 observations with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2 respectively is given by