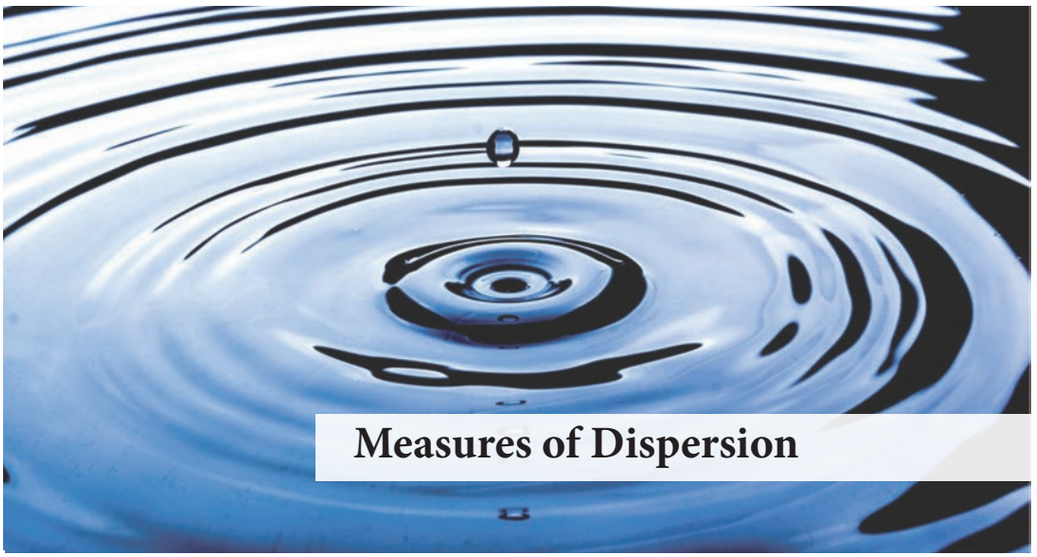
Chapter

6



Measures of Dispersion

Karl Pearson
(27 March 1857 – 27 April 1936)

Karl Pearson was an English mathematician and biostatistician. He has been credited with establishing the discipline of mathematical statistics. He founded the world's first university Statistics department at University College London in 1911, and contributed significantly to the field of biometrics, meteorology, theories of social Darwinism and eugenics. In fact, Pearson devoted much time during 1893 to 1904 in developing statistical techniques for biometry. These techniques, which are widely used today for statistical analysis.

'The 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

Learning Objectives

- Provides the importance of the concept of variability (dispersion)
- Measures the spread or dispersion and Identifiers the causes of dispersion
- Describes the spread - range and standard deviations
- Describes the role of Skewness and Kurtosis
- Explains about moments
- Illustrates the procedure to draw Box plot.

DI5PY

Introduction

The measures of central tendency describes the central part of values in the data set appears to concentrate around a central value called average. But these measures do

not reveal how these values are dispersed (spread or scattered) on each side of the central value. Therefore while describing data set it is equally important to know how for the item in the data are close around or scattered away from the measures of central tendency.

Example 6.1

Look at the runs scored by the two cricket players in a test match:

Players	I Innings	II Innings	Mean
Player 1	0	100	50
Player 2	40	60	50

Comparing the averages of the two players we may come to the conclusion that they were playing alike. But player 1 scored 0 runs in I innings and 100 in II innings. Player 2 scored nearly equal runs in both the innings. Therefore it is necessary for us to understand data by measuring dispersion.

6.1 Characteristics of a good Measure of Dispersion

An ideal measure of dispersion is to satisfy the following characteristics.

(i) It should be well defined without any ambiguity.

(ii) It should be based on all observations in the data set..

(iii) It should be easy to understand and compute.

(iv) It should be capable of further mathematical treatment.

(v) It should not be affected by fluctuations of sampling.

(vi) It should not be affected by extreme observations.

6.2 Types of measures of dispersion

Range, Quartile deviation, Mean deviations, Standard deviation and their Relative measures

The measures of dispersion are classified in two categories, namely

(i) Absolute measures

(ii) Relative measures.

6.3 Absolute Measures

It involves the units of measurements of the observations. For example, (i) the dispersion of salary of employees is expressed in rupees, and (ii) the variation of time

required for workers is expressed in hours. Such measures are not suitable for comparing the variability of the two data sets which are expressed in different units of measurements.

6.3.1 Range

Raw Data:

Range is defined as difference between the largest and smallest observations in the data set. Range(R) = Largest value in the data set (L) –Smallest value in the data set(S)

$$R = L - S$$

Grouped Data:

For grouped frequency distribution of values in the data set, the range is the difference between the upper class limit of the last class interval and the lower class limit of first class interval.

Coefficient of Range

The relative measure of range is called the coefficient of range

Co efficient of Range = (L- S) / (L+ S)

Example 6.2

The following data relates to the heights of 10 students (in cms) in a school. Calculate the range and coefficient of range.

158, 164, 168, 170, 142, 160, 154, 174, 159, 146

Solution:

L=174 S=142

Range = L – S = 174 – 142 = 32

Coefficient of range = (L - S)/(L + S)

= (174-142) / (174+142) = 32 / 316 = 0.101

Example 6.3

Calculate the range and the co-efficient of range for the marks obtained by 100 students in a school.

Marks	60-63	63-66	66-69	69-72	72-75
No. of students	5	18	42	27	8

Solution:

L = Upper limit of highest class = 75

S = lower limit of lowest class = 60

Range = L-S = 75-60 = 15

Co - efficient of Range = (L- S) / (L+ S)

= 15/ (75 + 60) = 15 / 135=0.111

Merits:

- Range is the simplest measure of dispersion.
- It is well defined, and easy to compute.
- It is widely used in quality control, weather forecasting, stock market variations etc.

Limitations:

- The calculations of range is based on only two values – largest value and smallest value.
- It is largely influenced by two extreme values.
- It cannot be computed in the case of open-ended frequency distributions.
- It is not suitable for further mathematical treatment.

6.3.2 Inter Quartile Range and Quartile Deviation

The quartiles Q_1, Q_2 and Q_3 have been introduced and studied in Chapter 5.

Inter quartile range is defined as: Inter quartile Range (IQR) = $Q_3 - Q_1$

Quartile Deviation is defined as, half of the distance between Q_1 and Q_3.

Quartile Deviation Q.D = $\frac{Q_3 - Q_1}{2}$

It is also called as semi-inter quartile range.

Coefficient of Quartile Deviation:

The relative measure corresponding to QD is coefficient of QD and is defined as:

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

Merits:

- It is not affected by the extreme (highest and lowest) values in the data set.
- It is an appropriate measure of variation for a data set summarized in open-ended class intervals.
- It is a positional measure of variation; therefore it is useful in the cases of erratic or highly skewed distributions.

Limitations:

- The QD is based on the middle 50 per cent observed values only and is not based on all the observations in the data set, therefore it cannot be considered as a good measure of variation.
- It is not suitable for mathematical treatment.
- It is affected by sampling fluctuations.
- The QD is a positional measure and has no relationship with any average in the data set.

6.3.3 Mean Deviation

The Mean Deviation (MD) is defined as the arithmetic mean of the absolute deviations of the individual values from a measure of central tendency of the data set. It is also known as the average deviation.

The measure of central tendency is either mean or median. If the measure of central tendency is mean (or median), then we get the mean deviation about the mean (or median).

$$\text{MD (about mean)} = \frac{\sum |D|}{n} \qquad D = (x - \bar{x})$$

$$\text{MD (about median)} = \frac{\sum |{}^{D}m|}{n} = D_m = x - \text{Median}$$

The coefficient of mean deviation (CMD) is the relative measure of dispersion corresponding to mean deviation and it is given by

$$\text{Coefficient of Mean Deviation (CMD)} = \frac{MD\,(mean\ or\ median)}{mean\ or\ median}$$

Example 6.4

The following are the weights of 10 children admitted in a hospital on a particular day.

Find the mean deviation about mean, median and their coefficients of mean deviation.

7, 4, 10, 9, 15, 12, 7, 9, 9, 18

Solution:

$n = 10$; Mean: $\bar{x} = \frac{\sum x}{n} = \frac{100}{10} = 10$

Median: The arranged data is: 4, 7, 7, 9, 9, 9, 10, 12, 15, 18

Median = $\frac{9+9}{2} = \frac{18}{2} = 9$:

Marks (x)	**$\lvert D \rvert = \lvert x - \bar{x} \rvert$**	**$\lvert D_m \rvert = \lvert x$- Median$\rvert$**
7	3	2
4	6	5
10	0	1
9	1	0
15	5	6
12	2	3
7	3	2
9	1	0
9	1	0
18	8	9
Total =100	**30**	**28**

Mean deviation from mean $= \frac{\sum |D|}{n} = \frac{30}{10} = 3$

Co-efficient mean deviation about mean $= \frac{\text{Mean deviation about mean}}{\bar{x}} = \frac{3}{10} = 0.3$

Mean deviation about median $= \frac{\sum |Dm|}{n}$

$= \frac{28}{10} = 2.8$

Co- efficient mean deviation about median $= \frac{\text{Mean deviation about median}}{\text{median}}$

$= \frac{2.8}{9} = 0.311$

6.3.4 Standard Deviation

Consider the following data sets.

10, 7, 6, 5, 4, 3, 2
10, 10, 10, 9, 9, 9, 2 ,2
10, 4, 4, 3, 2, 2, 2

It is obvious that the range for the three sets of data is 8. But a careful look at these sets clearly shows the numbers are different and there is a necessity for a new measure to address the real variations among the numbers in the three data sets. This variation is measured by standard deviation. The idea of standard deviation was given by Karl Pearson in 1893.

Definition:

'Standard deviation is the positive square root of average of the deviations of all the observation taken from the mean.' It is denoted by a greek letter σ.

(a) Ungrouped data

$x_1, x_2, x_3 \ldots x_n$ are the ungrouped data then standard deviation is calculated by

1. **Actual mean method:** Standard deviation $\sigma = \sqrt{\frac{\sum d^2}{n}}$, $d = x - \bar{x}$

2. **Assumed mean method:** Standard deviation $\sigma = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$, $d = x - A$

(b) Grouped Data (Discrete)

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \text{ , } d = x - A$$

Where, f = frequency of each class interval

N = total number of observation (or elements) in the population

x = mid – value of each class interval

where A is an assumed A.M.

(c) Grouped Data (continuous)

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times C, \quad d = \frac{x - A}{c}$$

Where, f = frequency of each class interval

N = total number of observation (or elements) in the population

c = width of class interval

x = mid-value of each class interval

where A is an assumed A.M.

Variance : Sum of the squares of the deviation from mean is known as Variance.

The square root of the variance is known as standard deviation.

NOTE

The simplified form of standard deviation formula may also be used

1. $\sigma = \frac{1}{n}\sqrt{n\sum d^2 - \left(\sum d\right)^2}$ (for raw data)
2. $\sigma = \frac{1}{N}\sqrt{N\sum fd^2 - \left(\sum fd\right)^2} \times C$

(for grouped data) where $d = (x\text{-A})/c$

Example 6.5

The following data gives the number of books taken in a school library in 7 days find the standard deviation of the books taken

7, 9, 12, 15, 5, 4, 11

Solution:

Actual mean method

$$\bar{x} = \frac{\sum x}{n}$$

$$\frac{7+9+.....+11}{7} = \frac{63}{7} = 9$$

x	$d = x - \bar{x}$	d^2
7	-2	4
9	0	0
12	3	9
15	6	36
5	-4	16
4	-5	25
11	2	4
		94

$$\sigma = \sqrt{\frac{\sum d^2}{n}}$$

$$= \sqrt{\frac{94}{7}}$$

$$= \sqrt{13.43}$$

$$= 3.66$$

NOTE

We can use two methods to find standard deviation

1. Direct method
2. Shortcut method

Merits:

- The value of standard deviation is based on every observation in a set of data.
- It is less affected by fluctuations of sampling.
- It is the only measure of variation capable of algebraic treatment.

Limitations:

- Compared to other measures of dispersion, calculations of standard deviation are difficult.
- While calculating standard deviation, more weight is given to extreme values and less to those near mean.
- It cannot be calculated in open intervals.
- If two or more data set were given in different units, variation among those data set cannot be compared.

Example 6.6

Raw Data:

Weights of children admitted in a hospital is given below calculate the standard deviation of weights of children.

13, 15, 12, 19, 10.5, 11.3, 13, 15, 12, 9

Solution:

$$\text{A.M.,} \quad \bar{x} = \frac{\sum x}{n}$$

$$= \frac{13 + 15 + \ldots\ldots + 49}{10}$$

$$= \frac{129.8}{10}$$

$$= 12.98$$

Deviation from actual mean

x	$d = x\text{-}12.98$	d^2
13	0.02	0.0004
15	2.02	4.0804
12	–0.98	0.9604
19	6.02	36.2404
10.5	2.48	6.1504

11.3	–1.68	2.8224
13	0.02	0.0004
15	2.02	4.0804
12	–0.98	0.9604
9	–3.98	15.8404
***n* =10**		**71.136**

Standard deviation $\sigma = \sqrt{\frac{\sum d^2}{n}}$

$= \sqrt{\frac{71.136}{10}}$

$= 2.67$

NOTE

If the mean value is not an integer, the calculation is difficult. In such a case we use the alternative formula for the calculation.

Example 6.7

Find the standard deviation of the first '*n*' natural numbers.

Solution:

The first *n* natural numbers are 1, 2, 3,…, *n*. The sum and the sum of squares of these *n* numbers are

$\sum x_i = 1+2+3+.....+n = \frac{n(n+1)}{2}$

$\sum x^2_i = 1^2 + 2^2 + 3^2 + ...n = \frac{n(n+1)(2n+1)}{6}$

Mean $\bar{x} = \frac{1}{n}\sum x_i = \frac{n(n+1)}{2n} = \frac{(n+1)}{2}$

$\frac{\sum x_i^2}{n} = \frac{(n+1)(2n+1)}{6}$

Standard deviation, $\sigma = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$

$= \sqrt{\frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}}$

$= \sqrt{\frac{2(n+1)(2n+1) - 3(n+1)^2}{12}}$

$= \sqrt{\frac{(n+1)[2(2n+1) - 3(n+1)]}{12}}$

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

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

Example 6.8

The wholesale price 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 is cumbersome when x values are large. So, another method is used, which will reduce the calculation time. Here we take the deviations from an assumed mean or arbitrary value A such that $d = x - A$

In this question, if we take deviation from an assumed A.M. =255. The calculations then for standard deviation will be as shown in below Table;

Observations (x)	**$d = x - A$**	**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**

$$\text{Variance} = \sigma^2 = \frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2$$

$$= \frac{1617}{7} - \left(\frac{35}{7}\right)^2$$

$$= 231 - 5^2$$

$$= 231 - 25$$

$$= 206$$

$$\text{Standard deviation } \sigma = \sqrt{variance}$$

$$\sigma = \sqrt{206} = 14.35$$

Example 6.9

The mean and standard deviation from 18 observations are 14 and 12 respectively. If an additional observation 8 is to be included, find the corrected mean and standard deviation.

Solution:

The sum of the 18 observations is = $n \times \bar{x}$ = 18 × 14 = 252.

The sum of the squares of these 18 observations

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

$$12^2 = \frac{\sum x^2}{18} - 14^2$$

$$144+196 = \frac{\sum x^2}{18}$$

$$\frac{\sum x^2}{18} = 340$$

$$\sum x^2 = 340 \times 18 = 6120$$

When the additional observation 8 is included, then n=19 ,

Σx = 252+8 = 260

Therefore, Corrected Mean = 260/19 = 13.68

$$\text{Corrected } \sum x^2 = \sum x^2 + 8^2$$

$$= 6120+64$$

$$= 6184$$

$$\text{Corrected Variance } \sigma^2 = \frac{6184}{19} - 13.68^2$$

$$= 325.47 - 187.14$$

$$= 138.33;$$

$$\text{Corrected Standard deviation } \sigma = \sqrt{138.33}$$

$$\sigma = 11.76$$

Example 6.10

A study of 100 engineering companies gives the following information

Profit (₹ 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:

A = 35 C = 10

Profit (Rs. in Crore)	Mid-value (x)	$d = \frac{x-A}{C}$	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	0	30	0	0
40 – 50	45	1	20	20	20
50 – 60	55	2	10	20	40
Total			**100**	**-28**	**200**

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

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

$$= \sqrt{2 - (0.078)} \times 10$$

$$= 13.863$$

Activity

Find the standard deviation for this problem using the other two formulae.

6.4 Combined Mean and Combined Standard Deviation

Combined arithmetic mean can be computed if we know the mean and number of items in each group of the data.

$\overline{x_1}, \overline{x_2}, \sigma_1, \sigma_2$ are mean and standard deviation of two data sets having n_1 and n_2 as number of elements respectively.

$$\text{combined mean } \overline{x_{12}} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2} \quad \text{(if two data sets)}$$

$$\overline{x_{123}} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + n_3\overline{x_3}}{n_1 + n_2 + n_3} \quad \text{(if three data sets)}$$

Combined standard deviation

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

$$d_1 = \overline{x_{12}} - \overline{x_1}$$

$$d_2 = \overline{x_{12}} - \overline{x_2}$$

Example 6.11

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

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

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

(ii) Find the combined standard deviation?

Solution:

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

Total wage bill paid monthly by *X* and *Y* is

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

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

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

(ii) For calculating the combined variance, we will first calculate the combined mean

$$\overline{x_{12}} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

$$= \frac{2750000 + 29250000}{550 + 650}$$

$$= \text{Rs. } 4729.166$$

Combined standard deviation

$$d_1 = \overline{x_{12}} - \overline{x_1} = 4729.166 — 5000 = -270.834$$

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

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

$$= \sqrt{\frac{4,08,38,080.55 + 3,51,76,082.50}{1200}} = \sqrt{633445} = 251.68$$

6.5 Relative Measures

It is a pure number independent of the units of measurements. This measure is useful especially when the data sets are measured in different units of measurement.

For example, suppose a nutritionist would like to compare the obesity of school children in India and England. He collects data from some of the schools in these two countries. The weight is normally measured in kilograms in India and in pounds in England. It will be meaningless, if we compare the obesity of students using absolute measures. So it is sensible to compare them in relative measures.

6.5.1 Coefficient of Variation

The standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, ie., heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

The coefficient of variation is obtained by dividing the standard deviation by the mean and multiplying it by 100. Symbolically,

$$\text{Coefficient of Variation } (C.V) = \frac{\sigma}{\bar{x}} \times 100$$

If we want to compare the variability of two or more series, we can use *C.V.* The series or groups of data for which the *C.V* is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the *C.V* is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

Merits:

- The *C.V* is independent of the unit in which the measurement has been taken, but standard deviation depends on units of measurement. Hence one should use the coefficient of variation instead of the standard deviation.

Limitations:

- If the value of mean approaches 0, the coefficient of variation approaches infinity. So the minute changes in the mean will make major changes.

Example 6.12

If the coefficient of variation is 50 per cent and a standard deviation is 4, find the mean.

Solution:

$$\text{Coefficient of Variation} = \frac{\sigma}{\bar{x}} \times 100$$

$$50 = \frac{4}{\bar{x}} \times 100$$

$$\bar{x} = \frac{4}{50} \times 100 = 8$$

Example 6.13

The scores of two batsmen, *A* and *B*, in ten innings during a certain season, are as under:

A: Mean score = 50; Standard deviation = 5

B: Mean score = 75; Standard deviation = 25

Find which of the batsmen is more consistent in scoring.

Solution:

$$\text{Coefficient of Variation } (C.V) = \frac{\sigma}{\bar{x}} \times 100$$

$$C.V \text{ for batsman } A = \frac{5}{50} \times 100 = 10\%$$

$$C.V \text{ for batsman } B = \frac{25}{75} \times 100 = 33.33\%$$

The batsman with the smaller *C.V* is more consistent.

Since for Cricketer *A*, the *C.V* is smaller, he is more consistent than *B*.

Example 6.14

The weekly sales of two products *A* and *B* were recorded as given below

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

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

Solution:

For comparing the fluctuations in sales of two products, we will prefer to calculate coefficient of variation for both the products.

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

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

$$\bar{x} = A + \frac{\sum fd}{N}$$

$$= 56 - \frac{57}{7} = 47.86$$

$$\text{Variance } \sigma^2 = \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2$$

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

$$= 412.14 - 66.30 = 345.84$$

$$\text{Standard deviation } \sigma = \sqrt{345.84} = 18.59$$

$$\text{Coefficient of variation } (C.V) = \frac{\sigma}{\bar{x}} \times 100$$

$$= \frac{18.59}{47.86} \times 100$$

$$= 38.84\ \%$$

Product *B*

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

$$\bar{x} = A + \frac{\sum fd}{N}$$

$$= 225 + \frac{15}{7}$$

$$= 225 + 2.14 = 227.14$$

$$\text{Variance } \sigma^2 = \frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2$$

$$= \frac{35125}{7} - \left(\frac{15}{7}\right)^2$$

$$= 5017.85 - 4.59$$

$$= 5013.26$$

$$\text{Standard deviation} = \sqrt{5013.26} = 70.80$$

$$\text{Coefficient of variation } (C.V)\ B = \frac{70.80}{227.14} \times 100$$

$$= 31.17\%$$

Since the coefficient of variation for product *A* is more than that of product *B*,

Therefore the fluctuation in sales of product *A* is higher than product *B*.

6.6 Moments

6.6.1 Raw moments:

Raw moments can be defined as the arithmetic mean of various powers of deviations taken from origin. The r^{th} Raw moment is denoted by μ_r', r = 1,2,3... Then the first raw moments are given by

Raw moments	Raw data ($d=x-A$)	Discrete data ($d=x-A$)	Continuous data ($d=(x-A)/c$)
μ_1'	$\frac{\sum d}{n}$	$\frac{\sum fd}{N}$	$\frac{\sum fd}{N} \times c$
μ_2'	$\frac{\sum d^2}{n}$	$\frac{\sum fd^2}{N}$	$\frac{\sum fd^2}{N} \times c^2$
μ_3'	$\frac{\sum d^3}{n}$	$\frac{\sum fd^3}{N}$	$\frac{\sum fd^3}{N} \times c^3$
μ_4'	$\frac{\sum d^4}{n}$	$\frac{\sum fd^4}{N}$	$\frac{\sum fd^4}{N} \times c^4$

6.6.2 Central Moments:

Central moments can be defined as the arithmetic mean of various powers of deviation taken from the mean of the distribution. The r^{th} central moment is denoted by μ_r, $r = 1, 2, 3...$

Central moments	Raw data	Discrete data	Continuous data $d' = \frac{(x-\bar{x})}{c}$
μ_1	$\frac{\sum (x-\bar{x})}{n} = 0$	$\frac{\sum f(x-\bar{x})}{N} = 0$	$\frac{\sum fd'}{N} \times c$
μ_2	$\frac{\sum f(x-\bar{x})^2}{N} = 0$	$\frac{\sum f(x-\bar{x})^2}{N} = \sigma^2$	$\frac{\sum fd'^2}{N} \times c^2$
μ_3	$\frac{\sum (x-\bar{x})^3}{n}$	$\frac{\sum f(x-\bar{x})^3}{N}$	$\frac{\sum fd'^3}{N} \times c^3$
μ_4	$\frac{\sum (x-\bar{x})^4}{n}$	$\frac{\sum f(x-\bar{x})^4}{N}$	$\frac{\sum fd'^4}{N} \times c^4$

In general, given n observation $x_1, x_2, \ldots, x_n$ the r^{th} order raw moments ($r = 0,1,2,...$) are defined as follows:

$$\mu'_r = \frac{1}{N}\sum f(x-A)^r \text{ (about } A\text{)}$$

$$\mu'_r = \frac{\sum fx^r}{N} \text{ (about origin)}$$

$$\mu_r = \frac{1}{N}\sum f(x-\bar{x})^r \text{ (about mean)}$$

NOTE

Raw moments are denoted by μ'_r and central moments are denoted by μ_r

6.6.3 Relation between raw moments and central moments

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2(\mu'_1)^2 - 3(\mu_1')^4$$

Example 6.15

The first two moments of the distribution about the value 5 of the variable, are 2 and 20.find the mean and the variance.

Solution:

$\mu'_1 = 2, \mu'_2 = 20$ and $A = 5$

$$\bar{x} = \mu'_1 + A$$

$$\bar{x} = 2+5=7$$

$$\sigma^2 = \mu'_2 - (\mu'_1)^2$$

$$\sigma^2 = 20 - 2^2 = 16$$

Mean = 7 and Variance = 16

6.7 SKEWNESS AND KURTOSIS

There are two other comparable characteristics called skewness and kurtosis that help us to understand a distribution.

6.7.1 Skewness

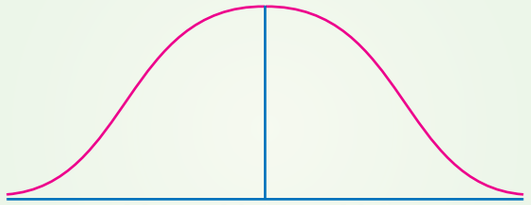
Skewness means 'lack of symmetry'. We study skewness to have an idea about the shape of the curve drawn from the given data. When the data set is not a symmetrical distribution, it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed.

The concept of skewness will be clear from the following three diagrams showing a symmetrical distribution, a positively skewed distribution and negatively skewed distribution.

We can see the symmetricity from the following diagram.

(a) Symmetrical Distribution

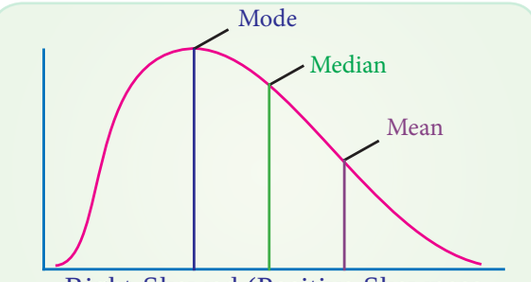
It is clear from the diagram below that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the centre point of the curve.



Mean = Median = Mode
Fig.6.1

(b) Positively Skewed Distribution

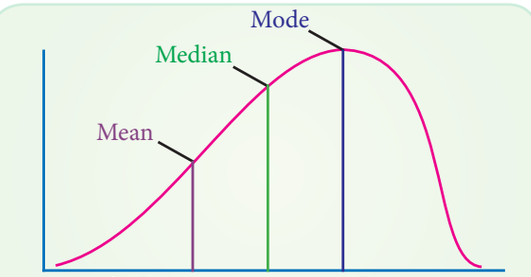
In the positively skewed distribution the value of the mean is maximum and that of mode is least – the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the high-value end of the curve (the right-hand side) than they are on the low – value end. For a positively skewed distribution, Mean>Median> Mode



Right-Skewed (Positive Skewness
Fig.6.2

(c) Negatively skewed distribution

In a negatively skewed distribution the value of mode is maximum and that of mean least-the median lies in between the two. In the negatively skewed distribution the position is reversed, i.e., the excess tail is on the left-hand side.



Left-Skewed (Positive Skewness
Fig.6.3

It should be noted that in moderately symmetrical distribution the interval between the mean and the median is approximately one-third of the interval between the mean and the mode. It is this relationship which provides a means of measuring the degree of skewness.

d. Some important Measures of Skewness

(i) Karl-Pearson coefficient of skewness

(ii) Bowley's coefficient of skewness

(iii) Coefficient of skewness based on moments

(i) Karl – Person coefficient of skewness

According to Karl-Pearson the absolute measure of skewness = Mean – Mode.

Karl- Pearson coefficient of skewness = $\frac{Mean - Mode}{S.D}$

Example 6.16

From the known data, mean = 7.35, mode = 8 and Variance = 1.69 then find the Karl- Pearson coefficient of skewness.

Solution:

Karl- Pearson coefficient of skewness $= \frac{Mean - Mode}{S.D}$

Variance $= 1.69$

Standard deviation $= \sqrt{1.69} = 1.3$

Karl- Pearson coefficient of skewness $= \frac{7.35 - 8}{1.3}$

$= \frac{-0.65}{1.3} = -0.5$

(ii) Bowley's coefficient of skewness

In Karl Pearson method of measuring skewness the whole of the series is needed. Prof. Bowley has suggested a formula based on position of quartiles. In symmetric distribution quartiles will be equidistance from the median. $Q_2 - Q_1 = Q_3 - Q_2$, but in skewed distributions it may not happen. Hence

Bowley's coefficient of skewness (SK) = $\frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$

Example 6.17

If $Q_1 = 40$, $Q_2 = 50$, $Q_3 = 60$, find Bowley's coefficient of skewness

Solution:

Bowley's coefficient of skewness

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

Bowley's coefficient of skewness

$= \frac{40 + 60 - 2 \times 50}{60 - 40} = \frac{0}{20} = 0$

$\therefore$ Given distribution is symmetric.

NOTE

If the difference between the mean and median or mean and mode is greater, the data is said to be more dispersed.

(iii) Measure of skewness based on Moments

The Measure of skewness based on moments is denoted by β_1 and is given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

Example 6.18

Find β_1 for the following data $\mu_1 = 0$, $\mu_2 = 8.76$, $\mu_3 = -2.91$

Solution:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_1 = \frac{(-2.91)^2}{(8.76)^3} = \frac{8.47}{672.24}$$

$$= 0.0126$$

6.7.2 Kurtosis

Kurtosis in Greek means 'bulginess'. In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of distribution is measured relative to the peakedness of normal curve. In other words, measures of kurtosis tell us the extent of which a distribution is more peaked or flat-topped than the normal curve.

The following diagram illustrates the shape of three different curves mentioned below:

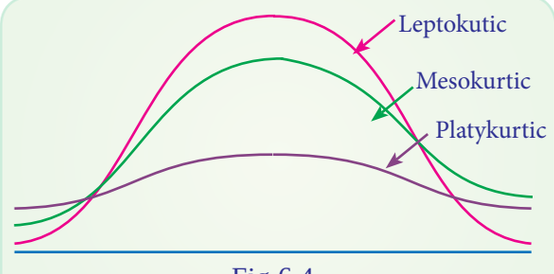


Fig.6.4

If a curve is more peaked than the normal curve, it is called 'leptokurtic'. In such a case items are more closely bunched around the mode. On the other hand if a curve is more flat-topped than the normal curve, it is called 'platykurtic'. The bell shaped normal curve itself is known as 'mesokurtic'. We can find how much the frequency curve is flatter than the normal curve using measure of kurtosis.

Measures of Kurtosis

The most important measure of kurtosis is the value of the coefficient. It is defined as: coefficient of kurtosis $\beta_2 = \frac{\mu_4}{\mu_2^2}$

NOTE

The greater the value of β_2 , the more peaked the distribution.

(i) The normal curve and other curves with $\beta_2 = 3$ are called mesokurtic.

(ii) When the value of β_2 is greater than 3, the curve is more peaked than the normal curve, i.e., leptokurtic.

(iii) When the value of β_2 is less than 3 the curve is less peaked than the normal curve, i.e., platykurtic.

Example 6.19

Find the value of β_2 for the following data $\mu_1 = 0\ \mu_2 = 4\ \mu_3 = 0\ \mu_4 = 37.6$

Solution:

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

$$\beta_2 = \frac{37.6}{4^2} = \frac{37.6}{16} = 2.35 < 3$$

$\beta_2 < 3$, The curve is platykurtic.

6.8 Box plot

A box plot can be used to graphically represent the data set. These plots involve five specific values:

(i) The lowest value of the data set (i.e., minimum), (ii) Q_1 (iii) The median (iv) Q_3, (v) The highest value of the data set (i.e maximum)

These values are called a five- number summary of the data set.

A box plot is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1 and a horizontal line from Q_3 to the maximum data value, and drawing a box by vertical lines passing through Q_1 and Q_3, with a vertical line inside the box passing through the median or Q_2.

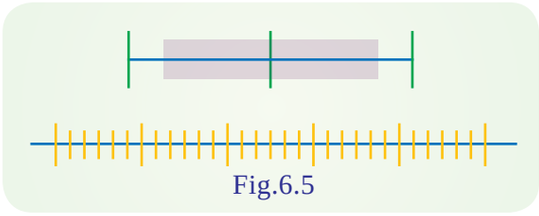


Fig.6.5

6.8.1 Description of boxplot

(i) If the median is near the center of the box, the distribution is approximately symmetric

(ii) If the median falls to the left of the center of the box, the distribution is positively skewed.

(iii) If the median falls to the right of the center of the box, the distribution is negatively skewed.

(iv) If the lines are about the same length, the distribution is approximately symmetric

(v) If the right line is larger than the left line. the distribution is positively skewed.

(vi) If the left line is larger than the right line. the distribution is negatively skewed.

Remark:

(i) The line drawn from minimum value of the dataset to Q_1 and Q_3 to the maximum value of the data set is called whisker.

(ii) Box plot is also called Box – Whisker plot.

(iii) A box and whisker plot illustrate the spread of the distribution and also gives an idea of the shape of the distribution

Example 6.20

The following data gives the Number of students studying in XI standard in 10 different schools 89,47,164,296,30,215,138,78,48, 39 construct a boxplot for the data.

Solution:

Step 1: Arrange the data in order

30,39,47,48,78,89,138,164,215,296

Step 2: Find the median

Median = $\frac{78+89}{2}$ = 83.5

Step 3: Find Q_1

30,39,47,48,78

Q_1 = 47

Step 4: Find Q_3

89,138,164,215,296

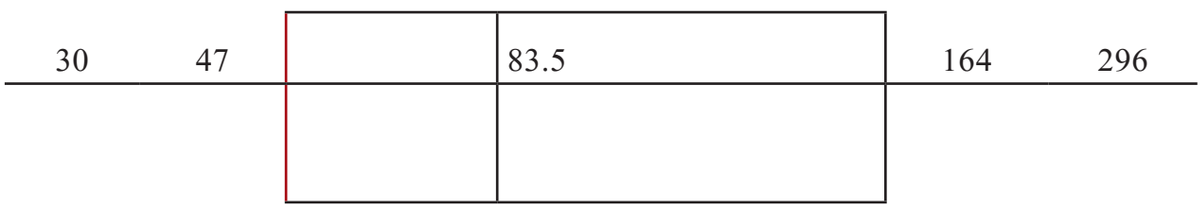
$Q_3 = 164$

Step 5: Find the minimum and maximum values.

Step 6: Locate the lowest value, Q_1, median, Q_3 and the highest value on the scale.

Step 7: Draw a box through Q_1 and Q_3

Box plot



Example 6.21

Construct a box –whisker plot for the following data

96, 151, 167, 185, 200, 220, 246, 269, 238, 252, 297, 105, 123, 178, 202

Solution:

Step 1: Arrange the data in code

96,105,123,151,167,178,185,200,202,220,238,246,252,269,297.

Step 2: Find the Median

8^{th} term Median = 200

Step 3: Find Q_1 (middle of previous terms of 200)

96,105,123,151,167,178,185

$Q_1 = 151$

Step 4 : Find Q_3 (middle of successive terms of 200)

202,220,238,246,252,269,297

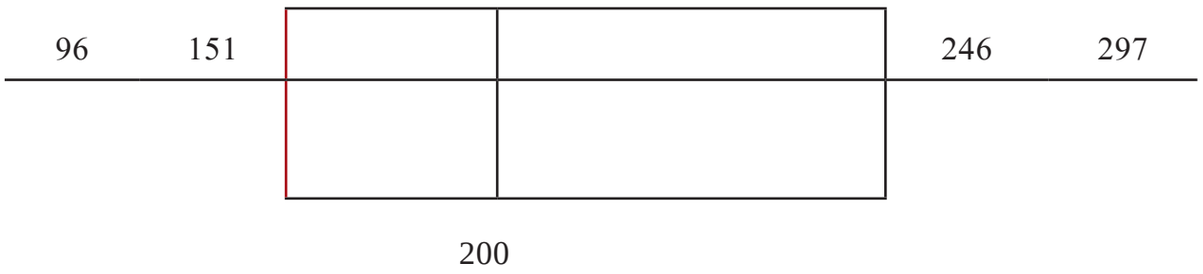
$Q_3 = 246$

Step 5: Minimum value = 96, Maximum Value = 297

Step 6: Draw a scale for the data on the x axis

Step 7: Locate the five numbers in the scale and draw a box around

Box plot



Points to Remember

- The range is the difference between the largest and smallest observations.
- The inter quartile range (IQR) is the difference between the upper and lower quartiles.
- The variance is the average of the squares of the values of $x-\overline{x}$
- The standard deviation (s.d) is the square root of the variance and has the same units as x
- If a population is approximately **symmetric** in a sample the mean and the median will have similar values. Typically their values will also be close to that of the mode of the population (if there is one!)
- A population that is not symmetric is said to be **skewed**. A distribution with a long 'tail' of high values is said to be **positively skewed**, in which case the mean is usually greater than the mode or the median. If it has a long tail of low values it is said to be **negatively skewed**, then the mean is likely to be the lowest of the three location measures of the distribution
- Box plots (Box-whisker diagrams): indicate the least and greatest values together with the quartiles and the median.

EXERCISE : 6

I. Choose the best answer:

1. When a distribution is symmetrical and has one mode, the highest point on the curve is called the

   (a) Mode (b) Median (c) Mean (d) All of these.

2. When referring to a curve tails to the left end, you would call it.

   (a) Symmetrical (b) Negatively skewed
   (c) Positively skewed (d) All of these

3. Disadvantages of using the range as a measure of dispersion include all of the following except

   (a) It is heavily influenced by extreme values

   (b) It can change drastically from one sample to the next

   (c) It is difficult to calculate

   (d) It is determined by only two points in the data set.

4. Which of the following is true?

   (a) The variance can be calculated for grouped or ungrouped data.

   (b) The standard deviation can be calculated for grouped or ungrouped data.

   (c) The standard deviation can be calculated for grouped or ungrouped data but the variance can be calculated only for ungrouped data.

   (d) (a) and (b), but not (c).

5. The squareroot of the variance of a distribution is the

   (a) Standard deviation (b) Mean
   (c) Range (d) Absolute deviation

6. The standard deviation of a set of 50 observations in 8. If each observation is multiplied by 2, then the new value of standard deviation will be:

   (a) 4 (b) 2 (c)16 (d) 8

7. In a more dispersed (spread out) set of data:

   (a) Difference between the mean and the median is greater

   (b) Value of the mode is greater

(c) Standard deviation is greater

(d) Inter-quartile range is smaller

8. Which of the following is a relative measure of dispersion?

(a) standard deviation (b) variance
(c) coefficient of variation (d) all of the above

9. If quartile deviation is 8, then value of the standard deviation will be:

(a) 12 (b) 16 (c) 24 (d) none of the above

10. If the difference between the mean and the mode is 35 and the standard deviation is 10 then the coefficient of skewness is

(a) 2.5 (b) 1.5 (c) 3.5 (d) 6.5

II. Fill in the Blanks:

11. The difference between the values of the first and third quartiles is the _____

12. The measure of the average squared difference between the mean and each item in the population is the _____.The positive square root of this value is the _____

13. The expression of the standard deviation as a percentage of the mean is the _____

14. The number of observations lies above or below the median is called the _____

15. If $\beta_2 = 3$ the curve is_____

III Answer shortly:

16. What is dispersion? What are various measures of dispersion?

17. What is meant by relative measure of dispersion? Describe its uses.

18. Define mean deviation. How does it differ from standard deviation?

19. What is standard deviation? Explain its important properties? What is variance?

20. What are the measures of skewness?

21. Write the measures used in box plot.

IV Answer in brief:

22. Explain dispersion and write their uses?

23. What are the requisites of a good measure of variation?

24. Explain how measures of central tendency and measures of variations are complementary to each other in the context of analysis of data.

25. Distinguish between absolute and relative measures of variation. Give a broad classification of the measures of variation.

26. Explain and illustrate how the measures of variation afford a supplement to averages in frequency distribution.

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

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

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

30. In a small business firm, two typists are employed-typist A and typist B. Typist A types out, on an average, 30 pages per day with a standard deviation of 6. Typist B, on an average, types out 45 pages with a standard deviation of 10. Which typist shows greater consistency in his output?

V Calculate the following:

31. Calculate mean deviation about mean for the following frequency distribution:

Age in Years	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45
No.of persons	7	10	16	32	24	18	10	5	1

32. The mean of two sample sizes 50 and 100 respectively are 54.1 and 50.3 and the standard deviations are 8 and 7. Find the mean and standard deviation of the sample size of 150 obtained by combining the two samples.

33. Following data represents the life of two models of refrigerators A and B.

Life (No. of years)		0-2	2-4	4-6	6-8	8-10	10-12
Refrigerator	Model A	5	16	13	7	5	4
	Model B	2	7	12	19	9	1

Find the mean life of each model. Which model has greater uniformity? Also obtain mode for both models.

34. Calculate the quartile deviation and coefficient of quartile deviation from the following data:

Size	06	09	12	15	18
Frequency	7	12	19	10	2

35. Calculate the appropriate measure of dispersion from the following data:

Wages in ₹	Below 35	35-37	38-40	41-43	Above 43
Earners	14	60	95	24	7

36. Two brands of tyres are tested with the following results:

Life (in 1000 miles)	20-25	25-30	30-35	35-40	40-45	45-50
Brand A	8	15	12	18	13	9
Brand B	6	20	32	30	12	0

Which is more consistent ?

37. Find the S.D. for the number of days patients admitted in a hospital.

Days of confinement	5	6	7	8	9
No. of patients	18	14	9	3	1

38. Calculate the quartile deviation and its coefficient from the following data:

Class Interval	10 -15	15 - 20	20 - 25	25 - 30	30 - 35
Frequency	8	12	14	10	6

39. Calculate the Mean Deviation from mean and its Coefficient from the following data, relating to Height (to the nearest cm) of 100 children:

Height(cms)	60	61	62	63	64	65	66	67	68
No. of Children	2	0	15	29	25	12	10	4	3

40. Find the standard deviation for the distribution given below:

x	1	2	3	4	5	6	7
Frequency	10	20	30	35	14	10	2

41. Calculate the coefficient of range separately for the two sets of data:

Set 1	10	20	9	15	10	13	28
Set 2	35	42	50	32	49	39	33

42. Blood serum cholesterol levels of 10 persons are 240, 260, 290, 245, 255, 288, 272, 263, 277 and 250. Calculate the standard deviation with the help of assumed mean.

43. Two groups of people played a game which reveals the quick operation of a particular key in a computer and the fraction of reaction times nearest to the tenth of a second is given below .

	Minimum	**I Quartile**	**Median**	**III Quartile**	**Maximum**
Group I	0.6	0.8	1.0	1.5	1.9
Group II	0.4	0.7	1.0	1.3	1.6

Draw two Box-Whisker plots and compare reaction times of the two groups.

44. Draw Box – Whisker plot for the following

(i) 3, 5, 10, 11, 12, 16, 17, 17, 19, 20, 22

(ii) –7, –5, –4, –4, –3, –3, –2, –1, 0, 1, 4, 6, 8, 9

Answers

I.1. d 2. b 3. c 4. d 5. a 6. c 7. a 8. c 9. d 10. c

II. 11. IQR 12. variance, standard deviation 13.coefficient of variation

14. 50% 15. normal (or) symmetric V.29. 2nd Formula 30. typist B

31. 7.13 32. 51.57, 7.67 33. (i) 5.12, 6.16 (ii) model B

34. 3, 0.25 35. QD=1.8, CQD=0.047 36. brand B 37. 1.03 38. 4.94,0.2265

39. 1.24, 0.0194 40. 1.4 41. 0.51,0.22 42. 16.5

Self-Practice Problems

1. The following samples shows the weekly number of road accidents in a city during a two-year period:

Number of Accidents	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	Total
Frequency	5	12	32	27	11	9	4	3	1	104

2. 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.

Activity

Age Group (years)	< 1	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	> 40
Frequency	15	113	122	91	110	119	132	65	46	15

Based on the figure answer the following question:

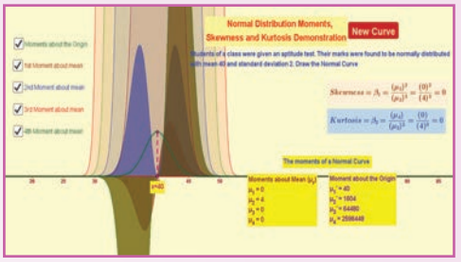
1. Whether the mean height can be taken as a representative of this group? If not what would be appropriate measure?

2. What would be the shape of the distribution of the height? In other words whether it is Symmetry?, if not what is the nature of skewness?

ICT CORNER

MOMENTS OF NORMAL CURVE

This activity helps to understand what is moments in probability using normal probability distribution



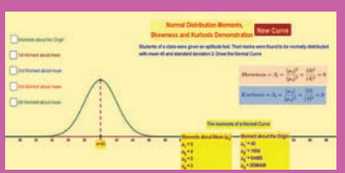
Steps:

- Open the browser and type the URL given (or) scan the QR code. GeoGebra work book called "11th Standard Statistics" will appear. In this several work sheets for statistics are given, open the worksheet named "Moments of Normal Curve"
- Normal curve will appear. You can change the normal curve by pressing on "New Curve". Observe the changes.
- You have to observe that for normal curve skewness and kurtosis is always zero. First you observe the calculation on the screen by changing the curve as in step-2. The reason you can find in next steps.
- First click on "Moments about the origin" check box. You can see the graph for moments about the origin in different colours. You get various values for moment about the origin. Observe the data at right bottom.
- If you click on "2nd moment about Mean" check box you see the curve is entirely on the top and it has some value.
- If you click on 1st, 3rd and 4th moment about mean you get top and bottom equal curves which are positive and negative. That is why these values are zero. That is why skewness and kurtosis formula lead to zero. (Observe)

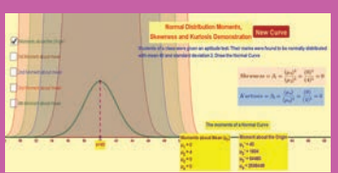
Step-1

Step-2

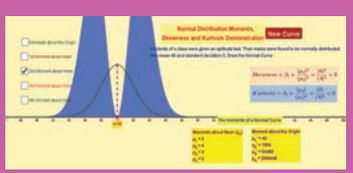
Step-3



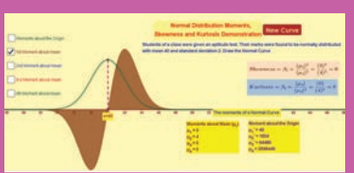
Step-4



Step-5



Step-6



Pictures are indicatives only*

URL:

https://ggbm.at/uqVhSJWZ

18H3Z