

Chapter 2

Descriptive Analysis

Learning Objectives

After completing this chapter, you should be able to do the following:

- Learn the importance of descriptive studies.
- Know the various statistics used in descriptive studies.
- Understand the situations in management research for undertaking a descriptive study.
- Describe and interpret various descriptive statistics.
- Learn the procedure of computing descriptive statistics using SPSS.
- Know the procedure of developing the profile chart of a product or organization.
- Discuss the findings in the outputs generated by the SPSS in a descriptive study.

Introduction

Descriptive studies are carried out to understand the profile of any organization that follows certain common practice. For example, one may like to know or be able to describe the characteristics of an organization that implement flexible working timing or that have a certain working culture. Descriptive studies may be undertaken to describe the characteristics of a group of employees in an organization. The purpose of descriptive studies is to prepare a profile or to describe interesting phenomena from an individual or an organizational point of view.

Although descriptive studies can identify sales pattern over a period of time or in different geographical locations but cannot ascertain the causal factors. These studies are often very useful for developing further research hypotheses for testing. Descriptive research may include case studies, cross-sectional studies, or longitudinal investigations.

Different statistics are computed in descriptive studies to describe the nature of data. These statistics computed from the sample provide summary of various measures. Descriptive statistics are usually computed in all most every experimental research

study. The primary goal in a descriptive study is to describe the sample at any specific point of time without trying to make inferences or causal statements. Normally, there are three primary reasons to conduct such studies:

1. To understand an organization by knowing its system
2. To help in need assessment and planning resource allocation
3. To identify areas of further research

Descriptive studies help in identifying patterns and relationships that might otherwise go unnoticed.

A descriptive study may be undertaken to ascertain and be able to describe the characteristics of variables of interest, in a given situation. For instance, a study of an organization in terms of percentage of employee in different age categories, their job satisfaction level, motivation level, gender composition, and salary structure can be considered as descriptive study. Quite frequently descriptive studies are undertaken in organizations to understand the characteristics of a group or employees such as age, educational level, job status, and length of service in different departments.

Descriptive studies may also be undertaken to know the characteristics of all those organizations that operate in the same sector. For example, one may try to describe the production policy, sales criteria, or advertisement campaign in pharmacy companies. Thus, the goal of descriptive study is to offer the researcher a profile or to describe relevant aspects of the phenomena of interest in an organization, industry, or a domain of population. In many cases, such information may be vital before considering certain corrective steps.

In a typical profile study, we compute various descriptive statistics like mean, standard deviation, coefficient of variation, range, skewness, and kurtosis. These descriptive statistics explain different features of the data. For instance, mean explains an average value of the measurement, whereas standard deviation describes variation of the scores around their mean value; the coefficient of variation provides relative variability of scores; range gives the maximum variation; skewness explains the symmetry; and kurtosis describes the variation in the data set.

In descriptive studies, one tries to obtain information regarding current status of different phenomena. Purpose of such study is to describe “What exists?” with respect to situational variables.

In descriptive research, the statement of problem needs to be defined first and then identification of information is planned. Once the objectives of the study are identified, method of data collection is planned to obtain an unbiased sample, and therefore, it is important to define the population domain clearly. Further, an optimum sample size should be selected for the study as it enhances the efficiency in estimating population characteristics.

Once the data is collected, it should be compiled in a meaningful manner for further processing and reporting. The nature of each variable can be studied by looking to the values of different descriptive statistics. If purpose of the study is

analytical as well, then these data may further be analyzed for testing different formulated hypotheses.

On the basis of descriptive statistics and graphical pictures of the parameters, different kinds of generalizations and predictions can be made. While conducting descriptive studies, one gets an insight to identify the future scope of the related research studies.

Measures of Central Tendency

Researchers are often interested in defining a value that best describes some characteristics of the population. Often, this characteristic is a measure of central tendency. A measure of central tendency is a single score that attempts to describe a set of data by identifying the central position within that set of data. The three most common measures of central tendency are the mean, the median, and the mode. Measures of central tendency are also known as central location. Perhaps, you are more familiar with the mean (also known as average) as the measure of central tendency, but there are others, such as the median and the mode, which are appropriate in some specific situations.

The mean, median, and mode are all valid measures of central tendency, but, under different conditions, some measures of central tendency become more appropriate than other. In the following sections, we will look at the various features of mean, median, and mode and the conditions under which they are most appropriate to be used.

Mean

The mean is the most widely used and popular measure of central tendency. It is also termed as average. It gives an idea as to how an average score looks like. For instance, one might be interested to know that on an average how much is the sale of items per day on the basis of monthly sales figure. The mean is a good measure of central tendency for symmetric distributions but may be misleading in case of skewed distribution. The mean can be computed with both discrete and continuous data. The mean is obtained by dividing the sum of all scores by the number of scores in the data set.

If X_1, X_2, \dots, X_n are the n scores in the data set, then the sample mean, usually denoted by \bar{X} (pronounced X bar), is

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

This formula is usually written by using the Greek capital letter \sum , pronounced “sigma,” which means “sum of...”:

$$\bar{X} = \frac{1}{n} \sum X \quad (2.1)$$

In statistics, sample mean and population mean are represented in different manner, although the formulas for their computations are same. To show that we are calculating the population mean and not the sample mean, we use the Greek lower case letter “mu,” denoted as μ :

$$\mu = \frac{1}{n} \sum X$$

The mean is the model of your data set and explains that on an average, the data set tends to concentrate toward it. You may notice that the mean is not often one of the actual values that you have observed in your data set.

Computation of Mean with Grouped Data

If $X_1, X_2, X_3, \dots, X_n$ are n scores with $f_1, f_2, f_3, \dots, f_n$ frequencies respectively in the data set, then the mean is computed as

$$\bar{X} = \frac{\sum f_i X_i}{\sum f_i} = \frac{\sum f_i X_i}{n} \quad (2.2)$$

where

$\sum f_i X_i$ is the total of all the scores.

In case the data is arranged in class interval format, the X will be the midpoint of the class interval. Let us see how to explain the data shown in the class interval form in Table 2.1. The first class interval shows that the ten articles are in the price range of Rs. 1–50 and that of six articles are in the range of Rs. 51–100 and so on. Here, the exact price of each article is not known because they have been grouped together. Thus, in case of grouped data, the scores lose its own identity. This becomes problematic as it is difficult to add the scores because their magnitudes are not known. In order to overcome this problem, an assumption is made while computing mean and standard deviation from the grouped data. It is assumed that the frequency is concentrated at the midpoint of the class interval. By assuming so, the identity of each and every score can be regained; this helps us to compute the sum of all the scores which is required for computing mean and standard deviation. But by taking this assumption, it is quite likely that the scores may be either underestimated or overestimated. For instance, in Table 2.1, if all the ten items

Table 2.1 Frequency distribution of articles price

Class interval (price range in Rs.)	Frequency (f)	Midpoint (X)	fX
1–50	10	25.5	255
51–100	6	75.5	453
101–150	4	125.5	502
151–200	4	175.5	702
201–250	2	225.5	451
251–300	2	275.5	551
$\sum f = n = 28$			$\sum fX = 2914$

would have had prices in the range of Rs. 1–50 but due to assumption they are assumed to have prices as Rs. 25.5, a negative error may be created which is added in the computation of mean. But it may be quite likely that the prices of other six items may be on the higher side, say Rs. 90, whereas they are assumed to have the price as 75.5 which creates the positive error. Thus, these positive and negative errors add up to zero in the computation of mean.

In Table 2.1, $\sum fX$ represents the sum of all the scores, and therefore,

$$\bar{X} = \frac{\sum f_i X_i}{n} = \frac{2914}{28} = 104.07$$

Effect of Change of Origin and Scale on Mean

If the magnitude of data is large, it may be reduced by using the simple transformation

$$D = \frac{X - A}{i}$$

where “A” and “i” are origin and scale, respectively. Thus, any score which is subtracted from all the scores in the data set is termed as origin, and any score by which all the scores are divided is known as scale. The choice of origin and scale is up to the researcher, but the only criterion which one should always keep in mind is that the very purpose of using the transformation is to simplify the data and computation.

Let us see what is the effect of change of origin and scale on the computation of mean? If all the X scores are transformed into D by using the above-mentioned transformation, then taking summation on both sides,

$$\begin{aligned} \sum D &= \sum \left(\frac{X-A}{i} \right) \\ \Rightarrow \sum (X - A) &= i \times \sum D \end{aligned}$$

Table 2.2 Showing computation for mean

Class interval (price range in Rs.)	Frequency (f)	Midpoint (X)	$D = \frac{X-175.5}{50}$	fD
1-50	10	25.5	-3	-30
51-100	6	75.5	-2	-12
101-150	4	125.5	-1	-4
151-200	4	175.5	0	0
201-250	2	225.5	1	2
251-300	2	275.5	2	4
$\sum f = n = 28$				$\sum fD = -40$

Dividing both side by n ,

$$\begin{aligned} \frac{\sum (X - A)}{n} &= \frac{i \times \sum D}{n} \\ \Rightarrow \frac{1}{n} \sum X - \frac{nA}{n} &= i \times \frac{1}{n} \sum D \\ \Rightarrow \bar{X} &= A + i \times \bar{D} \end{aligned}$$

Thus, we have seen that if all the scores X are transformed into D by changing the origin and scales as A and i , respectively, then the original mean can be obtained by multiplying the new mean \bar{D} by the scale i and adding the origin value into it. Thus, it may be concluded that the mean is not independent of change of origin and scale.

Computation of Mean with Deviation Method

In case of grouped data, the mean can be computed by transforming the scores so obtained by taking the midpoint of the class intervals. Consider the data shown in Table 2.1 once again. After computing the midpoint of the class intervals, let us transform the scores by changing the origin and scale as 175.5 and 50, respectively. Usually, origin (A) is taken as the midpoint of the middlemost class interval, and the scale (i) is taken as the width of the class interval. The origin A is also known as assumed mean (Table 2.2).

Here, width of the class interval = $i = 50$ and assumed mean $A = 175.5$

Since we know that

$$\begin{aligned} \bar{X} &= A + i \times \bar{D} = A + i \times \frac{1}{n} \sum fD \\ \Rightarrow \bar{X} &= 175.5 + 50 \times \frac{1}{28} \times (-40) \\ &= 175.5 - 71.43 = 104.07 \end{aligned}$$

In computing the mean, the factor $i \times (1/n) \sum fD$ can be considered as the correction factor. If the assumed mean is taken higher than the actual mean, the correction factor shall be negative, and, if it is taken as lower than the actual mean, the correction factor will become positive. One may take assumed mean as the midpoint of the even lowest or highest class interval. But in that case, the magnitude of the correction factor shall be higher and the very purpose of simplifying the computation process shall be defeated. Thus, the correct strategy is to take the midpoint of the middlemost class interval as the assumed mean. However, in case the number of class intervals is even, midpoint of any of the two middle class intervals may be taken as the assumed mean.

Properties of Mean

1. The mean is the most reliable measure of central tendency as it is computed by using all the data in the data set.
2. Mean is more stable than any other measures of central tendency because its standard error is least in comparison to median and mode. It simply means that if you compute mean of different samples that are drawn from the same population, then the fluctuation among these means shall be least in comparison to that of other measures of central tendencies like median and mode.
3. If \bar{X}_1 and \bar{X}_2 are the means of the two groups computed from the two sets of values n_1 and n_2 , then the combined mean \bar{X} is given by the following formula:

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

4. The sum of the deviation of a set of values from their arithmetic mean is always 0. In other words,

$$\sum (X - \bar{X}) = 0$$

To prove this, expand left-hand side of this expression

$$\begin{aligned} \sum (X - \bar{X}) &= \sum X - \sum \bar{X} \\ &= N\bar{X} - N\bar{X} = 0 \end{aligned}$$

5. The mean is highly affected by the outliers.
6. In the absence of even one observation, it is impossible to compute the mean correctly.
7. In case of open-ended class interval, the mean cannot be computed.

Median

Median is the middlemost score in the data set arranged in order of magnitude. It is a positional average and is not affected by the extreme scores. If X_1, X_2, \dots, X_n are the n scores in a data set arranged in the ascending or descending order, then its median is obtained by

$$M_d = \left(\frac{N+1}{2} \right)^{\text{th}} \text{ score} \quad (2.4)$$

One should note that $(n+1)/2$ is not the median, but the score lying in that position is the median. Consider the weight of the following ten subjects: 56, 45, 53, 41, 48, 53, 52, 65, 38, 42.

After arranging the scores

S.N.:	1	2	3	4	5	6	7	8	9	10
Weight:	38	41	42	45	48	52	53	53	56	65

Here, $n = 10$.

Thus, $M_d = \left(\frac{10+1}{2} \right)^{\text{th}} = 5.5^{\text{th}} \text{ score} = \frac{(48+52)}{2} = 50$

In case of odd number of scores you will get a single score lying in the middle, but in case of even number of scores, the middlemost score is obtained by taking the average of the two middle scores as in that case there are two middle scores.

Median is used in case the effect of extreme scores needs to be avoided. For example, consider the marks of the students in a college as shown below:

Student:	1	2	3	4	5	6	7	8	9	10
Marks:	35	40	30	32	35	39	33	32	91	93

The mean score for these ten students is 46. However, the raw data suggests that this mean value might not be the best way to accurately reflect the typical performance of a student, as most students have marks in between 30 and 40. Here, the mean is being skewed by the two large scores. Therefore, in this situation, median gives better estimate of average instead of mean. Thus, in a situation where the effect of extreme scores needs to be avoided, median should be preferred over mean. In case the data is normally distributed, the values of mean, median, and mode are same. Moreover, they all represent the most typical value in the data set. However, as the data becomes skewed, the mean loses its ability to provide the best central location as the mean is being dragged in the direction of skew. In that case, the median best retains this position and is not influenced much by the skewed values. As a rule of thumb if the data is non-normal, then it is customary to use the median instead of the mean.

Computation of Median for Grouped Data

While computing the median for grouped data, it is assumed that the frequencies are equally distributed in the class interval. This assumption is also used in computing the quartile deviation because median and quartile deviation both are nonparametric statistics and depend upon positional score. In case of grouped data, the median is computed by the following formula:

$$M_d = || + \frac{\frac{n}{2} - F}{f_m} \times i \quad (2.5)$$

where

$||$: lower limit of the median class

n : total of all the frequencies

F : cumulative frequency of the class just lower than the median class

f_m : frequency of the median class

i : width of the class interval

The computation of the median shall be shown by means of an example. Consider the marks in mathematics obtained by the students as shown in Table 2.3.

In computing median, first of all we need to find the median class. Median class is the one in which the median is supposed to lie. To obtain the median class, we compute $n/2$ and then we look for this value in the column of cumulative frequency. The class interval for which the cumulative frequency includes the value $n/2$ is taken as median class.

Here, $n = 70$
and therefore, $\frac{n}{2} = \frac{70}{2} = 35$

Now, we look for 35 in the column of cumulative frequency. You can see that the class interval 31–35 has a cumulative frequency 48 which includes the value $n/2 = 35$. Thus, 31–35 is the median class. After deciding the median class, the median can be computed by using the formula (2.5).

Here, $||$ = Lower limit of the median class = 30.5

f_m = Frequency of the median class = 20

F = Cumulative frequency of the class just lower than the median class = 28

i = Width of the class interval = 5

Substituting these values in the formula (2.5),

$$\begin{aligned} M_d &= || + \frac{\frac{n}{2} - F}{f_m} \times i \\ &= 30.5 + \frac{35 - 28}{20} \times 5 = 30.50 + 1.75 = 32.25 \end{aligned}$$

In computing the lower limit of the median class, 0.5 has been subtracted from the lower limit because the class interval is discrete. Any value which is equal or

Table 2.3 Frequency distribution of marks in mathematics

	Class interval (marks range)	Frequency (<i>f</i>)	Cumulative frequency (<i>F</i>)
	10 or less	2	2
	11–15	4	6
	16–20	5	11
	21–25	6	17
	26–30	11	28
Median class	31–35	20	48
	36–40	15	63
	41–45	4	67
	46–50	3	70
	$\sum f = n = 70$		

greater than 30.5 shall fall in the class interval 31–35, and that is why actual lower limit is taken as 30.5 instead of 31. But in case of continuous class intervals, lower limit of the class interval is the actual lower limit, and we do not subtract 0.5 from it. In case of continuous class interval, it is further assumed that the upper limit is excluded from the class interval. This make the class intervals mutually exclusive.

In Table 2.3, the lowest class interval is truncated, and therefore, its midpoint can be computed; hence, the mean can not be computed in this situation. Thus, if the class intervals are truncated at one or both the ends, median is the best choice as a measure of central tendency.

Mode

Mode can be defined as the score that occurs most frequently in a set of data. If the scores are plotted, then the mode is represented by the highest bar in a bar chart or histogram. Therefore, mode can be considered as the most popular option in the set of responses. Usually, mode is computed for categorical data where we wish to know as to which the most common category is. The advantage of mode is that it is not affected by the extreme scores (outliers). Sometime, there could be two scores having equal or nearly equal frequencies in the data set. In that case, the data set will have two modes and the distribution shall be known as bimodal. Thus, on the basis of the number of modes, the distribution of the scores may be unimodal, bimodal, or multimodal. Consider the following data set: 2, 5, 4, 7, 6, 3, 7, 8, 7, 9, 1, 7. Here, the score 7 is being repeated maximum number of times; hence, the mode of this data set is 7.

The mode can be used in variety of situations. For example, if a pizza shop sells 12 different varieties of pizzas, the mode would represent the most popular pizza. Mode may be computed to know as to which of the text book is more popular

than others, and accordingly, the publisher would print more copy of that book instead of printing equal number of all books.

Similarly, it is important for the manufacturer to produce more of the most popular shoes because manufacturing different shoes in equal numbers would cause a shortage of some shoes and an oversupply of others. Other applications of the mode may be to find the most popular brand of soft drink or biscuits to take the manufacturing decision accordingly.

Drawbacks of Mode

1. Computing mode becomes problematic if the data set consists of continuous data, as we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 40 persons' height (to the nearest 1 cm). It will be very unlikely that any two or more people will have the same height. This is why the mode is very rarely used with continuous data.
2. Mode need not necessarily be unique. There may be more than one mode present in the data set. In that case, it is difficult to interpret and compare the mode.
3. If no value in the data set is repeated, then every score is a mode which is useless.

Computation of Mode for Grouped Data

In computing the mode with grouped data first of all one needs to identify the modal class. The class interval, for which the frequency is maximum, is taken as modal class. The frequency of the modal class is denoted by f_0 , and that of frequencies before and after the modal class are represented by f_1 and f_2 , respectively. Once these frequencies are identified, they can be used to compute the value of the mode. The formula for computing mode with the grouped data is given by

$$M_0 = ll + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i \quad (2.6)$$

where

ll : lower limit of the modal class

f_m : frequency of the modal class

f_1 : frequency of the class just lower than the modal class

f_2 : frequency of the class just higher than the modal class

i : width of the class interval

Table 2.4 shows the distribution of age of bank employees. Let us compute the value of mode in order to find as to what is the most frequent age of employees in the bank.

Since the maximum frequency is 50 for the class interval 26–30, hence this will be the modal class here.

Table 2.4 Frequency distribution of age

Class interval C.I.	Frequency (<i>f</i>)
21–25	25
26–30	50
31–35	10
36–40	5
41–45	4
46–50	2

Now, ll = lower limit of the modal class = 25.5
 f_m : frequency of the modal class = 50
 f_1 : frequency of the class just lower than the modal class = 25
 f_2 : frequency of the class just higher than the modal class = 10
 i : width of the class interval = 5
After substituting these values in the formula (2.6), we get

$$\begin{aligned} M_0 &= ll + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i \\ &= 25.5 + \frac{50 - 25}{2 \times 50 - 25 - 10} \times 5 \\ &= 25.5 + 1.92 = 27.42 \end{aligned}$$

Thus, one may conclude that mostly employees in the bank are of around 27 years of age.

Summary of When to Use the Mean, Median, and Mode

Following summary table shows the suitability of different measures of central tendency for different types of data.

Nature of variable	Suitable measure of central tendency
Nominal data (categorical)	Mode
Ordinal data	Median
Interval/ratio (symmetrical or nearly symmetrical)	Mean
Interval/ratio (skewed)	Median

Measures of Variability

Variability refers to the extent of scores that vary from each other. The data set is said to have high variability when it contains values which are considerably higher and lower than the mean value. The terms variability, dispersion, and spread are all synonyms and refer as to how much the distribution is spread out. Measure of central tendency refers to the central location in the data set, but the central location itself is not sufficient to define the characteristics of the data set. It may happen that the two data sets are similar in their central location but might differ in their variability. Thus, measure of central tendency and measure of variability both are required to describe the nature of the data correctly. There are four measures of variability that are frequently used, the range: interquartile range, variance, and standard deviation. In the following paragraphs, we will look at each of these four measures of variability in more detail.

The Range

The range is the crudest measure of variability and is obtained by subtracting the lowest score from the highest score in the data set. It is rarely used because it is based on only two extreme scores. The range is simple to compute and is useful when it is required to evaluate the whole of a data set. The range is useful in showing the maximum spread within a data set. It can be used to compare the spread between similar data sets.

Using range becomes problematic if one of the extreme score is exceptionally high or low (referred to as outlier). In that case, the range so computed may not represent the true variability within the data set. Consider a situation where scores obtained by the students on a test were recorded and the minimum and maximum scores were 25 and 72, respectively. If a particular student did not appear in the exam due to some reason and his score was posted as zero, then the range becomes $72(72-0)$ instead of $47(72-25)$. Thus, in the presence of an outlier, the range provides the wrong picture about the variability within the data set. To overcome the problem of outlier in a data set, the interquartile range is often calculated instead of the range.

The Interquartile Range

The interquartile range is a measure that indicates the maximum variability of the central 50% of values within the data set. The interquartile range can further be divided into quarters by identifying the upper and lower quartiles. The lower quartile (Q_1) is equivalent to the 25th percentile in the data set which is arranged in order of magnitude, whereas the upper quartile (Q_3) is equivalent to the 75th

percentile. Thus, Q_1 is a point below which 25% scores lie, and Q_3 refers to a score below which 75% scores lie. Since the median is a score below which 50% scores lie, hence, the upper quartile lies halfway between the median and the highest value in the data set, whereas the lower quartile lies halfway between the median and the lowest value in the data set. The interquartile range is computed by subtracting the lower quartile from the upper quartile and is given by

$$Q = Q_3 - Q_1 \quad (2.7)$$

The interquartile range provides a better picture of the overall data set by ignoring the outliers. Just like range, interquartile range also depends upon the two values. Statistically, the standard deviation is more powerful measure of variability as it is computed with all the values in the data set.

The Standard Deviation

The standard deviation is the most widely used measure of variability, the value of which depends upon how closely the scores cluster around the mean value. It can be computed only for interval or ratio data. The standard deviation is the square root of the average squared deviation of the scores from its mean value and is represented by σ (termed as sigma):

$$\sigma = \sqrt{\frac{1}{N} \sum (X - \mu)^2}$$

After simplification,

$$\sigma = \sqrt{\frac{1}{N} \sum X^2 - \left(\frac{\sum X}{N}\right)^2} \quad (2.8)$$

where μ is the population mean. The term σ is used for population standard deviation, whereas S is used for sample standard deviation. The population standard deviation σ can be estimated from the sample data by the following formula:

$$S = \sqrt{\frac{1}{n-1} \sum (X - \bar{X})^2}$$

After simplifying,

$$S = \sqrt{\frac{1}{n-1} \sum X^2 - \frac{(\sum X)^2}{n(n-1)}} \quad (2.9)$$

If $X_1, X_2, X_3, \dots, X_n$ are the n scores with $f_1, f_2, f_3, \dots, f_n$ frequencies respectively the data set, then the standard deviation shall be given as

$$S = \sqrt{\frac{1}{n-1} \sum f(X - \bar{X})^2}$$

After simplification,

$$S = \sqrt{\frac{1}{n-1} \sum fX^2 - \frac{(\sum fX)^2}{n(n-1)}} \tag{2.10}$$

where \bar{X} refers to the sample mean. The standard deviation measures the aggregate variation of every value within a data set from the mean. It is the most robust and widely used measure of variability because it takes into account every score in the data set.

When the scores in a data set are tightly bunched together, the standard deviation is small. When the scores are widely apart, the standard deviation will be relatively large. The standard deviation is usually presented in conjunction with the mean and is measured in the same units.

Computation of Standard Deviation with Raw Data

The sample standard deviation of a series of scores can be computed by using the formula (2.9). Following are the data on memory retention test obtained on 10 individuals. The scores are the number of items recollected by individuals (Table 2.5).

Table 2.5 Computation for standard deviation

(X)	(X ²)
4	16
5	25
3	09
2	04
6	36
8	64
4	16
5	25
6	36
4	16
$\sum X = 47$	$\sum X^2 = 247$

Here $n = 10$, $\sum X = 47$, and $\sum X^2 = 247$.

Substituting these values in the formula (2.9),

$$\begin{aligned} S &= \sqrt{\frac{1}{n-1} \sum X^2 - \frac{(\sum X)^2}{n(n-1)}} \\ &= \sqrt{\frac{1}{10-1} \times 247 - \frac{(47)^2}{10 \times 9}} \\ &= \sqrt{27.44 - 24.54} = 1.7 \end{aligned}$$

Thus, the standard deviation of the test scores on memory retention is 1.7. Looking to this value of standard deviation, no conclusion can be drawn as to whether the variation is less or more. It is so because standard deviation is considered to be the absolute variability. This problem can be solved by computing coefficient of variability. It will be discussed later in this chapter

Effect of Change of Origin and Scale on Standard Deviation

Let us see what happens to the standard deviation if the origin and scale of the scores are changed in the data set. Let the scores transformed by using the following transformation:

$$\begin{aligned} D &= \frac{X - A}{i} \\ \Rightarrow X &= A + i \times D \end{aligned}$$

where “A” is origin and “i” is the scale. One can choose any value of origin, but the value of scale is usually the width of the class interval.

Taking summation on both side and dividing both sides by n , we get

$$\bar{X} = A + i \times \bar{D}$$

(This has been proved above in (2.3))

$$S_X = \sqrt{\frac{1}{n-1} \sum f(X - \bar{X})^2}$$

Since $X - \bar{X} = A + iD - (A + i\bar{D}) = i(D - \bar{D})$

Substituting the value of $X - \bar{X}$, we get

$$\begin{aligned} S_X &= \sqrt{\frac{1}{n-1} \sum i^2 f(D - \bar{D})^2} = i \times \sqrt{\frac{1}{n-1} \sum f(D - \bar{D})^2} \\ \Rightarrow S_X &= i \times S_D \end{aligned} \tag{2.11}$$

Table 2.6 Computation of standard deviation

Class interval (price range in Rs.)	Frequency (<i>f</i>)	Midpoint (<i>X</i>)	$D = \frac{X-175.5}{50}$	fD	fD^2
1–50	10	25.5	–3	–30	90
51–100	6	75.5	–2	–12	24
101–150	4	125.5	–1	–4	4
151–200	4	175.5	0	0	0
201–250	2	225.5	1	2	2
251–300	2	275.5	2	4	8
$n = 28$				–40	128

Thus, it may be concluded that the standard deviation is free from change of origin but is affected by the change scale.

Let us compute the standard deviation for the data shown in Table 2.1. Consider the same data in Table 2.6 once again. After computing the midpoints of the class intervals, let us transform the scores by taking the origin and scale as 175.5 and 50, respectively. Usually, origin (*A*) is taken as the midpoint of the middlemost class interval, and the scale (*h*) is taken as the width of the class interval. The origin *A* is also known as assumed mean.

Here, width of the class interval = $h = 50$ and assumed mean $A = 175.5$.

From the equation (2.11), $S_X = i \times S_D = i \times \sqrt{\frac{1}{n-1} \sum f(D - \bar{D})^2}$

After simplification,

$$S_X = i \times \sqrt{\frac{1}{n-1} \sum fD^2 - \frac{(\sum fD)^2}{n(n-1)}}$$

Substituting the values of n , $\sum fD$ and $\sum fD^2$, we get

$$\begin{aligned} S_X &= 50 \times \sqrt{\frac{1}{28-1} \times 128 - \frac{(-40)^2}{28 \times 27}} \\ &= 80.93 \end{aligned}$$

Variance

The variance is the square of standard deviation. It can be defined as the average of the squared deviations of scores from their mean value. It also measures variation of

the scores in the distribution. It shows the magnitude of variation among the scores around its mean value. In other words, it measures the consistency of data. Higher variance indicates more heterogeneity, whereas lower variance represents more homogeneity in the data.

Like standard deviation, it also measures the variability of scores that are measured in interval or ratio scale. The variance is usually represented by σ^2 and is computed as

$$\sigma^2 = \frac{1}{N} \sum (X - \mu)^2 \quad (2.12)$$

The variance can be estimated from the sample by using the following formula:

$$\begin{aligned} \sigma^2 &= \frac{1}{n-1} \sum (X - \bar{X})^2 \\ &= \frac{1}{n-1} \sum X^2 - \frac{(\sum X)^2}{n(n-1)} \end{aligned}$$

Remark Population mean and population standard deviation are represented by μ and σ , respectively, whereas sample mean and sample standard deviation are represented by \bar{X} and S , respectively.

The Index of Qualitative Variation

Measures of variability like range, standard deviation, or variance are computed for interval or ratio data. What if the data is in nominal form? In social research, one may encounter many situations where it is required to measure the variability of the data based on nominal scale. For example, one may like to find the variability of ethnic population in a city, variation in the responses on different monuments, variability in the liking of different sports in an institution, etc. In all these situations, an index of qualitative variation (IQV) may be computed by the following formula to find the magnitude of variability:

$$IQV = \frac{K(100^2 - \sum P^2)}{100^2(K-1)} \quad (2.13)$$

where

K = The number of categories

$\sum P^2$ = Sum of squared percentages of frequencies in all the groups

Table 2.7 Frequency distribution of the students in different community

S.N.	Community	No. of students	% of students (P)	P^2
1	Hindu	218	68.1	4637.61
2	Muslim	55	17.2	295.84
3	Christian	25	7.8	60.84
4	Sikh	10	3.1	9.61
5	Others	12	3.8	14.44
				$\sum P^2 = 5018.34$

The IQV is based on the ratio of the total number of differences in the distribution to the maximum number of possible differences within the same distribution. This IQV can vary from 0.00 to 1.00. When all the cases are in one category, there is no variation and the IQV is 0.00. On the other hand, if all the cases are equally distributed across the categories, there is maximum variation and the IQV is 1.00.

To show the computation process, consider an example where the number of students belonging to different communities were recorded as shown in Table 2.7

Here, we have K = number of categories = 5:

$$\begin{aligned}
 \text{IQV} &= \frac{K(100^2 - \sum P^2)}{100^2(K - 1)} \\
 &= \frac{5 \times (100^2 - 5,018.34)}{100^2 \times (5 - 1)} = \frac{24,908.3}{40,000} \\
 &= 0.62
 \end{aligned}$$

By looking to the formula (2.13), you can see that the IQV is partially a function of the number of categories. Here, we used five categories of communities. Had we used more number of categories, the IQV would have been quite less, and, on the other hand, if the number of categories would have been less than the value of IQV, it would have been higher than what we are getting.

Standard Error

If we draw n samples from the same population and compute their means, then these means will not be the same but will differ with each other. The variation among these means is referred as the standard error of mean. Thus, the standard error of any statistic is the standard deviation of that statistic in the sampling distribution. Standard error measures the sampling fluctuation of any statistic and is widely used in statistical inference. The standard error gives a measure of how well a sample is true

representative of the population. When the sample is truly representing the population, the standard error will be small.

Constructing confidence intervals and testing of significance are based on standard errors. The standard error of mean can be used to compare the observed mean to a hypothesized value. The two values may be different at 5% level if the ratio of the difference to the standard error is less than -2 or greater than $+2$.

The standard error of any statistics is affected by the sample size. In general, the standard error decreases with the increase in sample size. It is denoted by σ with a subscript of a statistic for which it is computed.

Let $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_n$ are the means of n samples drawn from the same population. Then the standard deviation of these n mean scores is said to be standard error of mean. The standard error of sample mean can be estimated by even one sample. If any sample consists of n scores with population standard deviation σ , then the standard error of the mean is given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (2.14)$$

Whereas the standard error of the standard deviation is given by

$$\sigma_{\sigma} = \frac{\sigma}{\sqrt{2n}} \quad (2.15)$$

Like standard error of the mean, the standard error of the standard deviation also measures the fluctuation of standard deviations among the samples.

Remark If population standard deviation σ is unknown, it may be estimated by the sample standard deviation S .

Coefficient of Variation (CV)

Coefficient of variation is an index which measures the extent of variability in the data set in relation to its mean value. It is free from unit and compensates with the value of mean in the data set. Coefficient of variation is also known as relative variability and is denoted by CV

$$CV = \frac{S}{\bar{X}} \times 100 \quad (2.16)$$

where S and \bar{X} represent sample standard deviation and sample mean respectively. Since coefficient of variation measures the relative variability and computes the variability in percentage, it can be used to know whether a particular parameter is more variable or less variable. Coefficient of variation can be used for comparing the variability of two groups in a situation when their mean values are not equal.

It may also be used to compare the variability of two groups of data having different units.

On the other hand, standard deviation is a measure of absolute variability, and therefore, it cannot be used to assess the variability of any data set without knowing its mean value. Further, standard deviation cannot be used to compare the variability of two sets of scores if their mean value differs.

Consider the following statistics obtained on the number of customers visiting the retail outlets of a company in two different locations in a month. Let us see what conclusions can be drawn with this information.

Location	A	B
Mean	40	20
SD	8	6
CV	20%	30%

The standard deviation of the number of customers in location A is larger in comparison to location B, whereas coefficient of variation is larger in location B in comparison to location A. Thus, it may be inferred that the variation among the number of customers visiting the outlet in location B is higher than that of location A.

Moments

A moment is a quantitative value that tells us the shape of a set of points. The moment can be central or noncentral. Central moment is represented by μ_r , whereas noncentral moment is denoted by μ'_r . If the deviation of scores is taken around mean, then the moment becomes central, and if it is taken around zero or any other arbitrary value, it is known as noncentral moment. The r th central moment is given by

$$\mu_r = \frac{1}{n} \sum (X - \bar{X})^r \quad (2.17)$$

Different moments convey different meanings. For instance, second central moment μ_2 is always equal to variance of a distribution. Similarly second, third, and fourth moments are used to compute skewness and kurtosis of the data set. On the other hand, r th noncentral moment around the origin zero is denoted by

$$\mu'_r = \frac{1}{n} \sum X^r \quad (2.18)$$

The first noncentral moment μ'_1 about zero always represents mean of the distribution. These noncentral moments are used to compute central moments by means of a recurrence relation.

Skewness

Skewness gives an idea about the symmetricity of the data. In symmetrical distribution if the curve is divided in the middle, the two parts become the mirror image of each other. If the curve is not symmetrical, it is said to be skewed. The skewness of the distribution is represented by β_1 and is given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \quad (2.19)$$

where μ_2 and μ_3 are the second and third central moments. For a symmetric distribution, β_1 is 0. A distribution is positively skewed if β_1 is positive and negatively skewed if it is negative. In a positively skewed distribution, the tail is heavy toward the right side of the curve, whereas in a negatively skewed curve, the tail is heavy toward the left side of the curve. Further, in positively skewed curve, median is greater than mode, whereas in negatively skewed curve, the median is less than mode. Graphically both these curves can be shown by Fig. 2.1a, b:

The standard error of the skewness is given by

$$\text{SE}(\text{Skewness}) = \text{SE}(\beta_1) = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \quad (2.20)$$

where n is the sample size. Some authors use $\sqrt{6/n}$ for computing standard error of the skewness, but it is a poor approximation for the small sample.

The standard error of skewness can be used to test its significance. In testing the significance of skewness, the following Z statistic is used which follows a normal distribution.

$$Z = \frac{\sqrt{n(n-1)}}{n-2} \times \frac{\beta_1}{\text{SE}(\beta_1)} \quad (2.21)$$

The critical value of Z is approximately 2 (for a two-tailed test with roughly at 5% level). Thus, if calculated value of $Z < -2$, we may interpret that the population is

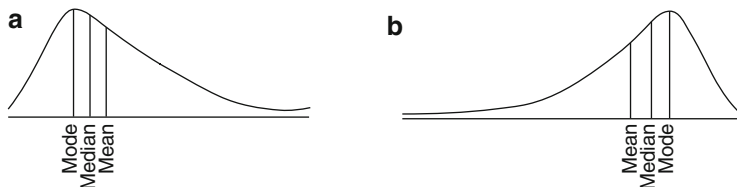


Fig. 2.1 (a and b) Showing positively and negatively skewed curve

very likely to be skewed negatively. On the other hand, if calculated $Z > +2$, it may be concluded that the population is positively skewed.

However, in general, skewness values more than twice its standard error indicates a departure from symmetry. This gives a criterion to test whether skewness (positive or negative) in the distribution is significant or not. Thus, if the data is positively skewed, it simply means that majority of the scores are less than its mean value, and in case of negative skewness, most of the scores are more than its mean value.

Kurtosis

Kurtosis is a statistical measure used for describing the distribution of observed data around the mean. It measures the extent to which the observations cluster around the mean value. It is measured by (Gamma) and is computed as

$$\gamma = \beta_2 - 3 \quad (2.22)$$

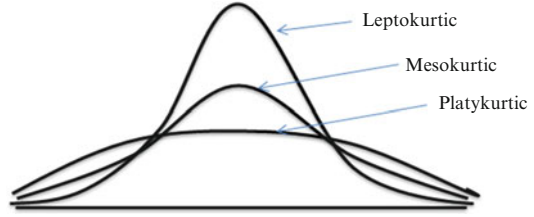
where $\beta_2 = \frac{\mu_4}{\mu_2^2}$, μ_2 and μ_4 represent the second and fourth central moments respectively.

For a normal distribution, the value of kurtosis (γ) is zero. Positive value of kurtosis in a distribution indicates that the observations cluster more around its mean value and have longer tails in comparison to that of normal distribution, whereas a distribution with negative kurtosis indicates that the observations cluster less around its mean and have shorter tails.

Depending upon the value of kurtosis, the distribution of scores can be classified into any one of the three categories: leptokurtic, mesokurtic, and platykurtic.

If for any variable the kurtosis is positive, the curve is known as leptokurtic and it represents a low level of data fluctuation, as the observations cluster around the mean. On the other hand, if the kurtosis is negative, the curve is known as platykurtic and it means that the data has a larger degree of variance. In other words, if the value of kurtosis is significant and positive, it signifies less variability in the data set or we may say that the data is more homogenous. On the other hand, significant negative kurtosis indicates that there is more variability in the data set or we may conclude that the data is more heterogeneous. Further, if the kurtosis is 0,

Fig. 2.2 Classification of curve on the basis of kurtosis



the curve is classified as mesokurtic. Its flatness is equivalent to normal curve. Thus, a normal curve is always a mesokurtic curve. The three types of the curves are shown in Fig. 2.2

The standard error of kurtosis can be given by

$$SE(\text{Kurtosis}) = SE(\gamma) = 2SE(\beta_1) \sqrt{\frac{n^2 - 1}{(n - 3)(n + 5)}} \quad (2.23)$$

where n is the sample size. Some author suggests the approximated formula for standard error of kurtosis as

$$SE(\text{Kurtosis}) = \sqrt{\frac{24}{n}} \quad (2.24)$$

but this formula is poor approximation for small samples.

The standard error of the kurtosis is used to test its significance. The test statistics Z can be computed as follows:

$$Z = \frac{\gamma}{SE(\gamma)} \quad (2.25)$$

This Z follows normal distribution. The critical value for Z is approximately 2 for two-tailed test in testing the hypothesis that kurtosis = 0 at approximately 5% level.

If the value of calculated Z is < -2 , then the population is very likely to have negative kurtosis and the distribution may be considered as platykurtic. On the other hand, if the value of calculated Z is $> +2$, then the population is very likely to have positive kurtosis and the distribution may be considered as leptokurtic.

Percentiles

Percentiles are used to develop norms based on the performance of the subjects. A given percentile indicates the percentage of scores below it and is denoted by P_X . For example, P_{40} is a score below which 40% scores lie. Median is also known as P_{50} , and it indicates that 50% scores lie below it. Percentiles can be computed to know the position of an individual on any parameter. For instance, 95th percentile obtained by a student in GMAT examination indicates that his performance is better than 95% of the students appearing in that examination.

Since 25th percentile P_{25} , 50th percentile P_{50} , and 75th percentile P_{75} are also known as first, second, and third quartiles, respectively, hence procedure of computing other percentiles will be same as the procedure adopted in computing quartiles. Quartiles (the 25th, 50th, and 75th percentiles) divide the data into four groups of equal size. Percentiles at decile points and quartiles can be computed by using SPSS.

Percentile Rank

A percentile rank can be defined as the percentage of scores that fall at or below a given score. Thus, if the percentile rank of a score A is X , it indicates that X percentage of scores lies below the score A . The percentile rank can be computed from the following formula:

$$\text{Percentile rank of the score } X = \frac{CF - 0.5 \times f_s}{n} \times 100 \quad (2.26)$$

where

CF: number of scores below X

f_s : number of times the score X occurs in the data set

n : number of scores in the data set

Situation for Using Descriptive Study

There may be varieties of situation where a descriptive study may be planned. One such situation has been discussed below to narrate the use of such study.

Nowadays Industries are also assuming social responsibilities toward society. They keep engage themselves in many of the social activities like adult literacy, slum development, HIV and AIDS program, community development, energy conservation drive, and go green campaign. One such organization has started its HIV and AIDS program in which it not only promotes the awareness but also provides treatments. This company provides antiretroviral therapy to anyone in the community who is a HIV-positive irrespective of whether that person is an employee of the company or not. The company also provides counseling, education, and training and disseminates information on nutrition, health, and hygiene. The target population of the company for this program is truck drivers, contract and migrant workers, employees of local organizations, and members of the local community. Descriptive study may be conducted to investigate the following issues:

- (a) Number of programs organized in different sections of the society
- (b) Number of people who attended the awareness program in different sections of the society

- (c) Number of people who are affected with HIV/AIDS in different sections of the society
- (d) The most vulnerable group affected by the HIV
- (e) Details of population affected from HIV in different age and sex categories

To cater the above objectives, data may be processed as follows:

- (i) Classify the data on HIV/AIDS-infected persons in different sections of the society like truck drivers, contract laborers, migrant's laborers, and local establishment members of the local community month wise in the last 5 years.
- (ii) Classify the number of participants attending the HIV/AIDS awareness program in different sections of the society month wise in the last 5 years.
- (iii) Compute the largest and smallest scores, mean, SD, coefficient of variation, standard error, skewness, kurtosis, and quartile deviation for the data in all the groups.

All these computations can be done by using SPSS, the procedure of which shall be explained later in this chapter by using the following example:

Solved Example of Descriptive Statistics using SPSS

The procedure of computing various descriptive statistics including central tendency, dispersion, percentile values, and distribution parameters through SPSS has been explained in the solved Example 2.1.

Example 2.1 In a study conducted by response of customers were obtained on various attributed of a company along with their satisfaction level. Apply descriptive analysis to compute various statistics and explain the findings (Table 2.8).

Solution In order to compute various descriptive statistics like mean, median, mode, SD, variance, skewness, SE of skewness, kurtosis, SE of kurtosis, range, minimum and maximum scores, and percentiles, a data file shall be made in SPSS and then steps shown below shall be followed to get the output. After getting the output, its interpretation shall be made.

Computation of Descriptive Statistics Using SPSS

(a) *Preparing Data File*

In order to use SPSS for computing descriptive statistics, a data file needs to be prepared. The data file can also be imported in SPSS from the ASCII or Excel files. The readers are advised to go through the first chapter of the book to learn

for starting the SPSS on the system, preparing the data file, and importing the file in SPSS from other sources. The following steps will help you to prepare the data file:

- (i) *Starting the SPSS*: Use the following command sequence to start SPSS on your system:
Start → All Programs → IBM SPSS Statistics → IBM SPSS Statistics 20
 After clicking **Type in Data** option, you will be taken to the **Variable View** option for defining the variables in the study.
- (ii) *Defining variables*: In this example, there are eight variables that need to be defined along with their properties. Do the following:
 1. Click **Variable View** in the left corner of the bottom of the screen to define variables and their properties.
 2. Write short name of all the eight variables as *Del_Speed*, *Price_Lev*, *Price_Fle*, *Manu_Ima*, *Service*, *Salfor_Ima*, *Prod_Qua*, and *Sat_Lev* under the column heading **Name**.
 3. Under the column heading **Label**, full name of these variables may be defined as *Delivery Speed*, *Price Level*, *Price Flexibility*, *Manufacturer Image*, *Service*, *Salesforce Image*, *Product Quality*, and *Satisfaction Level*.
 4. Since all the variables were measured on an interval scale, hence select the option “Scale” under the heading **Measure** for each variable.
 5. Use default entries in rest of the columns.

After defining variables in **Variable View**, the screen shall look like Fig. 2.3.

- (iii) *Entering data*: After defining all the eight variables in the **Variable View**, click **Data View** on the left bottom of the screen to open the format for entering the data column wise. For each variable, enter the data column wise. After entering the data, the screen will look like Fig. 2.4. Save the data file in the desired location before further processing.
- (b) **SPSS Commands for Descriptive Analysis**
 After entering the data in data view, do the following steps for computing desired descriptive statistics:
 - (i) *SPSS commands for descriptive statistics*: In data view, click the following commands in sequence:
Analyze ⇒ Descriptive Statistics ⇒ Frequencies
 The screen shall look like as shown in Fig. 2.5.
 - (ii) *Selecting variables for computing descriptive statistics*: After clicking the **Frequencies** tag, you will be taken to the next screen for selecting variables

Table 2.8 Response of customers on company's attributes

	Delivery S. speed N. (X_1)	Price level (X_2)	Price flexibility (X_3)	Manufacturer image (X_4)	Service (X_5)	Salesforce image (X_6)	Product quality (X_7)	Satisfaction level (X_8)
1	4.1	0.6	6.9	4.7	2.4	2.3	5.2	4.2
2	1.8	3.0	6.3	6.6	2.5	4.0	8.4	4.3
3	3.4	5.2	5.7	6.0	4.3	2.7	8.2	5.2
4	2.7	1.0	7.1	5.9	1.8	2.3	7.8	3.9
5	6.0	0.9	9.6	7.8	3.4	4.6	4.5	6.8
6	1.9	3.3	7.9	4.8	2.6	1.9	9.7	4.4
7	4.6	2.4	9.5	6.6	3.5	4.5	7.6	5.8
8	1.3	4.2	6.2	5.1	2.8	2.2	6.9	4.3
9	5.5	1.6	9.4	4.7	3.5	3.0	7.6	5.4
10	4.0	3.5	6.5	6.0	3.7	3.2	8.7	5.4
11	2.4	1.6	8.8	4.8	2.0	2.8	5.8	4.3
12	3.9	2.2	9.1	4.6	3.0	2.5	8.3	5.0
13	2.8	1.4	8.1	3.8	2.1	1.4	6.6	4.4
14	3.7	1.5	8.6	5.7	2.7	3.7	6.7	5.0
15	4.7	1.3	9.9	6.7	3.0	2.6	6.8	5.9
16	3.4	2.0	9.7	4.7	2.7	1.7	4.8	4.7
17	3.2	4.1	5.7	5.1	3.6	2.9	6.2	4.4
18	4.9	1.8	7.7	4.3	3.4	1.5	5.9	5.6
19	5.3	1.4	9.7	6.1	3.3	3.9	6.8	5.9
20	4.7	1.3	9.9	6.7	3.0	2.6	6.8	6.0

for which descriptive statistics need to be computed. The screen shall look like as shown in Fig. 2.6. Do the following:

- Select the variables *Del_Speed*, *Price_Lev*, *Price_Fle*, *Manu_Ima*, *Service*, *Salfor_Ima*, *Prod_Qua*, and *Sat_Lev* from the left panel to the “Variable(s)” section of the right panel.

Here, all the eight variables can be selected one by one or all at once. To do so, the variable(s) needs to be selected from the left panel, and by arrow command, it may be brought to the right panel. The screen shall look like Fig. 2.6.

- (iii) *Selecting option for computation*: After selecting the variables, options need to be defined for the computation of desired statistics. Do the following:

- Click the option **Statistics** on the screen as shown in Fig. 2.6. This will take you to the next screen that is shown in Fig. 2.7. Do the following:
 - Check the options “Quartiles” and “Cut points for 10 equal groups” in “Percentile Values” section.
 - Check the option “Mean,” “Median,” and “Mode” under “Central Tendency” section.
 - Check the option “Std. Deviation,” “Variance,” “Range,” “Minimum,” “Maximum,” “Range,” and “S.E. mean” under “Dispersion” section.

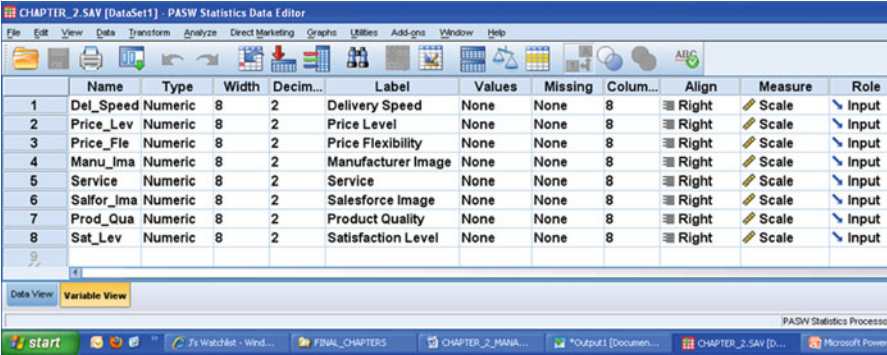
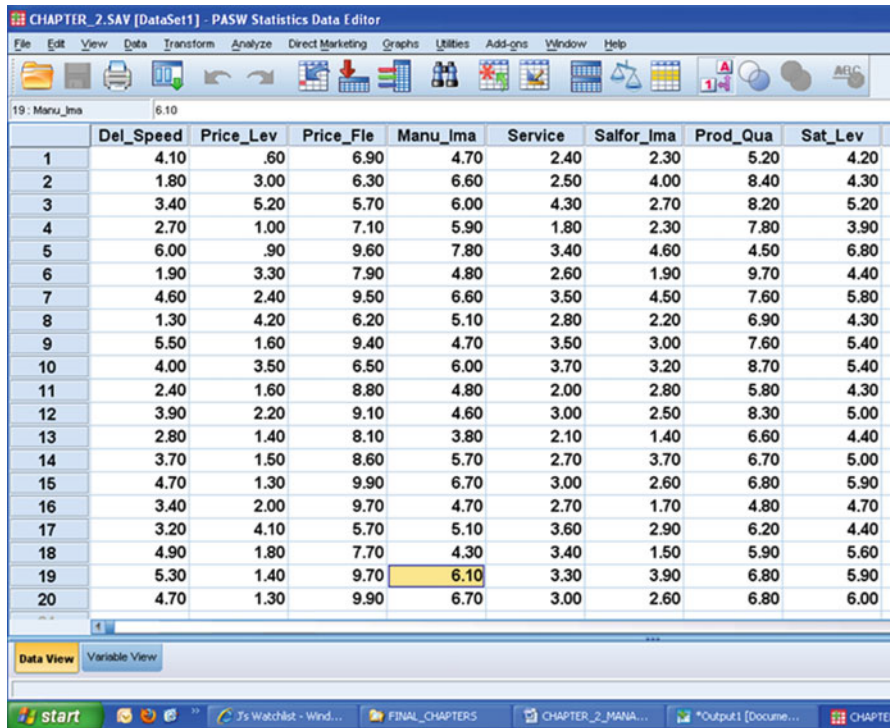


Fig. 2.3 Defining variables along with their characteristics

- Check the option “Skewness” and “Kurtosis” under “Distribution” section.
- Click **Continue** for getting back to the screen shown in Fig. 2.6.

Remarks

- (a) You have four different classes of statistics like “Percentile Value,” “Central Tendency,” “Dispersion,” and “Distribution” that can be computed. Any or all the options may be selected under these categories. Under the category “Percentile Values,” quartiles can be checked (✓) for computing Q_1 and Q_3 . For computing percentiles at deciles points, cut points can be selected for 10 equal groups. Similarly, if the percentiles are required to be computed in the interval of 5, cut points may be selected as 5.
 - (b) In using the option cut points for the percentiles, output contains some additional information on frequency in different segments. If the researcher is interested, the same may be incorporated in the findings; otherwise, it may be ignored.
 - (c) “Percentile” option is selected if percentile values at different intervals are required to be computed. For example, if we are interested in computing P_4, P_{16}, P_{27} , and P_{39} , then these numbers are added in the “Percentile(s)” option.
 - (d) In this problem, only quartiles and cut points for “10” options have been checked under the heading “Percentile Values,” whereas under the heading “Central Tendency,” “Dispersion,” and “Distribution,” all the options have been checked.
- (iv) *Option for graph:* The option **Chart** can be clicked in Fig. 2.6 if graph is required to be constructed. Any one of the option under this tag like bar charts, pie charts, or histograms may be selected. If no chart is required, then option “None” may be selected.



	Del_Speed	Price_Lev	Price_Fle	Manu_Ima	Service	Salfor_Ima	Prod_Qua	Sat_Lev
1	4.10	.60	6.90	4.70	2.40	2.30	5.20	4.20
2	1.80	3.00	6.30	6.60	2.50	4.00	8.40	4.30
3	3.40	5.20	5.70	6.00	4.30	2.70	8.20	5.20
4	2.70	1.00	7.10	5.90	1.80	2.30	7.80	3.90
5	6.00	.90	9.60	7.80	3.40	4.60	4.50	6.80
6	1.90	3.30	7.90	4.80	2.60	1.90	9.70	4.40
7	4.60	2.40	9.50	6.60	3.50	4.50	7.60	5.80
8	1.30	4.20	6.20	5.10	2.80	2.20	6.90	4.30
9	5.50	1.60	9.40	4.70	3.50	3.00	7.60	5.40
10	4.00	3.50	6.50	6.00	3.70	3.20	8.70	5.40
11	2.40	1.60	8.80	4.80	2.00	2.80	5.80	4.30
12	3.90	2.20	9.10	4.60	3.00	2.50	8.30	5.00
13	2.80	1.40	8.10	3.80	2.10	1.40	6.60	4.40
14	3.70	1.50	8.60	5.70	2.70	3.70	6.70	5.00
15	4.70	1.30	9.90	6.70	3.00	2.60	6.80	5.90
16	3.40	2.00	9.70	4.70	2.70	1.70	4.80	4.70
17	3.20	4.10	5.70	5.10	3.60	2.90	6.20	4.40
18	4.90	1.80	7.70	4.30	3.40	1.50	5.90	5.60
19	5.30	1.40	9.70	6.10	3.30	3.90	6.80	5.90
20	4.70	1.30	9.90	6.70	3.00	2.60	6.80	6.00

Fig. 2.4 Screen showing entered data for all the variables in the data view

– Press **O.K.** for output.

(c) *Getting the Output*

Clicking the option **OK** will lead you to the output window. The output panel shall have lots of results. It is up to the researcher to select the relevant outputs in their results. In the output window of the SPSS, the relevant output can be selected by pressing the right click of the mouse over it and may be copied in the word file. In this example, the output generated will look like as shown in Table 2.9

Interpretation of the Outputs

Different interpretations can be made from the results in Table 2.9. However, some of the important findings that can be drawn are as follows:

1. Except price level, mean and median of all the variables are nearly equal.

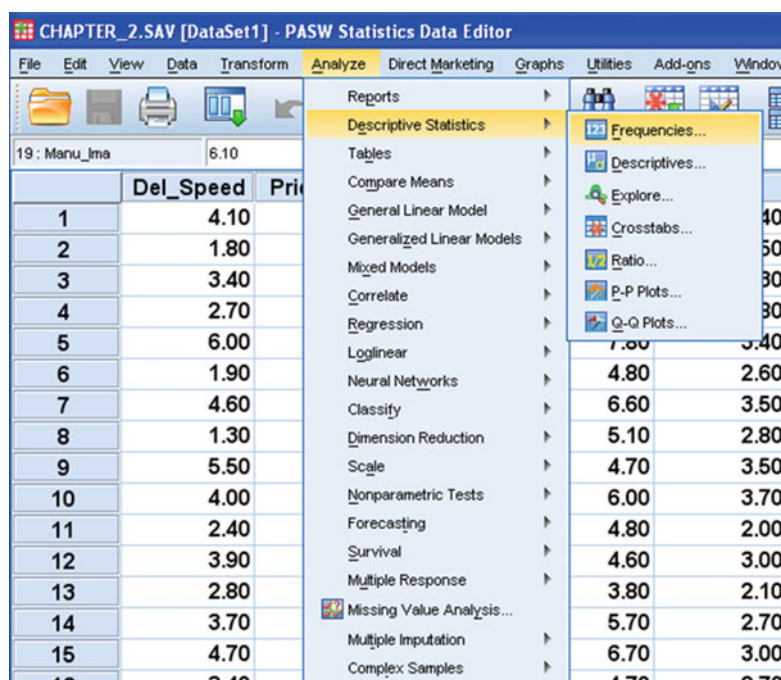


Fig. 2.5 Screen showing the SPSS commands for computing descriptive statistics

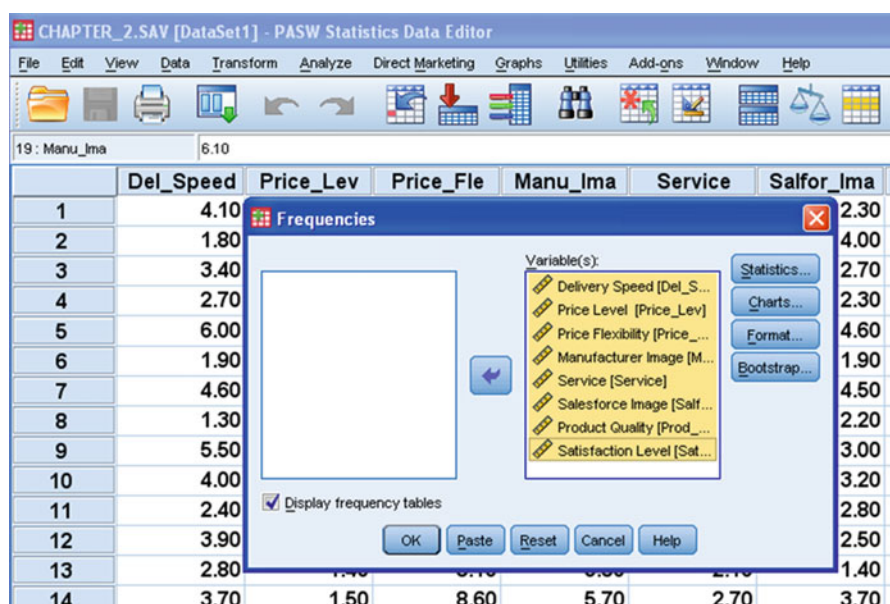


Fig. 2.6 Screen showing selection of variables for descriptive analysis

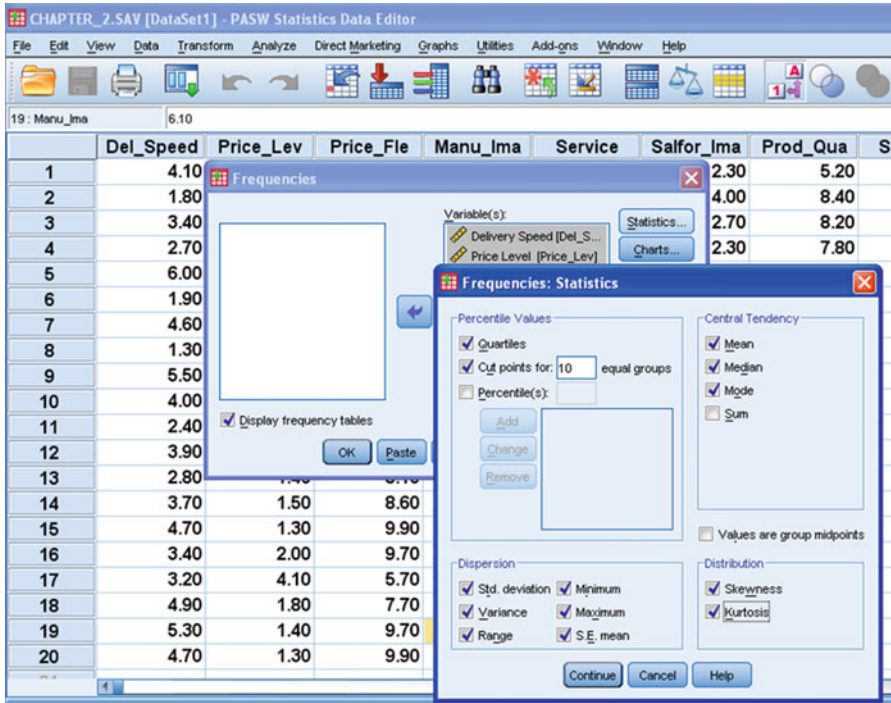


Fig. 2.7 Screen showing option for different statistics to be computed

- Standard error of mean is least for the service whereas maximum for the price flexibility.
- As a guideline, a skewness value more than twice its standard error indicates a departure from symmetry. Since none of the variable's skewness is greater than twice its standard error ($2 \times .512$) hence all the variables are symmetrically distributed.
- SPSS uses the statistic $\beta_2 - 3$ for kurtosis. Thus, for a normal distribution, kurtosis value is 0. If for any variable the value of kurtosis is positive, its distribution is known as leptokurtic, which indicates low level of data fluctuation around its mean value, whereas negative value of kurtosis indicates large degree of variance among the data and the distribution is known as platykurtic.

Since the value of kurtosis for any of the variable is not more than twice its standard error of kurtosis hence none of the kurtosis values are significant. In other words the distribution of all the variables is mesokurtic.

- Minimum and maximum values of the parameter can give some interesting facts and provide the range of variation. For instance, delivery speed of the products is in the range of 1.3–6 days. Thus, one can expect the delivery of any product in at the most 6 days time, and accordingly, one may try to place the order.

Table 2.9 Output showing various statistics for different attributes of the company

	Del_Speed	Price_Lev	Price_Fle	Manu_Ima	Service	Salfor_Ima	Prod_Qua	Sat_Lev
N Valid	20	20	20	20	20	20	20	20
Missing	0	0	0	0	0	0	0	0
Mean	3.7150	2.2150	8.1150	5.5350	2.9650	2.8150	6.9650	5.0450
SE of mean	.29094	.28184	.33563	.22933	.14203	.20841	.30177	.17524
Median	3.8000	1.7000	8.3500	5.4000	3.0000	2.6500	6.8000	5.0000
Mode	3.40 ^a	1.30 ^a	5.70 ^a	4.70	3.00	2.30 ^a	6.80	4.30 ^a
Std. deviation	1.30112	1.26044	1.50097	1.02561	.63518	.93205	1.34957	.78370
Variance	1.693	1.589	2.253	1.052	.403	.869	1.821	.614
Skewness	-.144	.970	-.338	.380	.010	.459	.001	.490
SE of skewness	.512	.512	.512	.512	.512	.512	.512	.512
Kurtosis	-.732	.092	-1.435	-.467	-.288	-.500	-.324	-.566
SE of kurtosis	.992	.992	.992	.992	.992	.992	.992	.992
Range	4.70	4.60	4.20	4.00	2.50	3.20	5.20	2.90
Minimum	1.30	.60	5.70	3.80	1.80	1.40	4.50	3.90
Maximum	6.00	5.20	9.90	7.80	4.30	4.60	9.70	6.80
Sum	74.30	44.30	162.30	110.70	59.30	56.30	139.30	100.90
Percentiles								
10	1.8100	.9100	5.7500	4.3300	2.0100	1.5200	4.8400	4.2100
20	2.4600	1.3000	6.3400	4.7000	2.4200	1.9600	5.8200	4.3000
25	2.7250	1.3250	6.6000	4.7000	2.5250	2.2250	5.9750	4.3250
30	2.9200	1.4000	6.9600	4.7300	2.6300	2.3000	6.3200	4.4000
40	3.4000	1.5400	7.7800	4.9200	2.7400	2.5400	6.7400	4.5200
50	3.8000	1.7000	8.3500	5.4000	3.0000	2.6500	6.8000	5.0000
60	4.0600	2.1200	8.9800	5.9600	3.1800	2.8600	7.3200	5.3200
70	4.6700	2.8200	9.4700	6.0700	3.4000	3.1400	7.7400	5.5400
75	4.7000	3.2250	9.5750	6.4750	3.4750	3.5750	8.1000	5.7500
80	4.8600	3.4600	9.6800	6.6000	3.5000	3.8600	8.2800	5.8800
90	5.4800	4.1900	9.8800	6.7000	3.6900	4.4500	8.6700	5.9900

^aMultiple modes exist. The smallest value is shown
Del_Speed delivery speed, Price_Lev price level, Price_Fle price flexibility, Manu_Ima manufacturer image, Service service, Salfor_Ima salesforce image, Prod_Qua product quality, Sat_Lev satisfaction level

- 6. Similarly, price flexibility of any product is in the range of 5.7–9.9%. This provides a feedback to the customers in taking a decision of buying an article in case of urgency.
- 7. Percentile scales can be used to draw various conclusions about different parameters. For instance, P_{40} for the delivery speed is 3.40, which indicates that 40% customers get their product delivered in less than 3.4 days.

Developing Profile Chart

In a descriptive study, a researcher generally computes different statistics that are described in Table 2.9. Based on these computations, meaningful interpretations can be made as shown above in the example. However, it would be more interesting to prepare a profile of the company using all its parameters investigated in the survey. The procedure of making a profile chart shall be explained by using the minimum score, maximum score, mean, and standard deviation of all the parameters shown in Table 2.9.

After manipulating data as per the following steps, the graphical functionality of Excel can be used to prepare the graphical profile of the company’s parameters:

- Step 1: Segregate the statistics like minimum score, maximum score, mean, and standard deviation of all the parameters in Table 2.9. The same has been shown in Table 2.10.
- Step 2: Convert minimum and maximum scores for each of the variables into its standard scores by using the following transformation:

$$Z = \frac{X - \bar{X}}{S}$$

Thus, mean of all the variables will become same. The values so obtained are shown in Table 2.11.

- Step 3: Convert these Z values into its linear transformed scores by using the transformation $Z_1 = 50 + 10 \times Z$. By using this transformation, the negative values of Z -scores can be converted into positive scores. Descriptive statistics shown in the form of linearly transformed scores are shown Table 2.12.

Table 2.10 Selected descriptive statistics of all the variables

Variables	Min	Max	Mean	S.D.
Delivery speed	1.3	6.00	3.715	1.30
Price level	0.60	5.20	2.21	1.26
Price flexibility	5.70	9.90	8.12	1.50
Manufacturer image	3.80	7.80	5.54	1.03
Service	1.80	4.30	2.97	0.64
Salesforce image	1.40	4.60	2.82	0.93
Product quality	4.50	9.70	6.97	1.35
Satisfaction level	3.90	6.80	5.05	0.78

Table 2.11 Standard scores of minimum, maximum, and average of all the variables

	Min (Z)	Mean (Z)	Max (Z)
Delivery speed	−1.86	0	1.76
Price level	−1.28	0	2.37
Price flexibility	−1.61	0	1.19
Manufacturer image	−1.69	0	2.19
Service	−1.83	0	2.08
Salesforce image	−1.53	0	1.91
Product quality	−1.83	0	2.02
Satisfaction level	−1.47	0	2.24

Table 2.12 Transformed standard scores of minimum, maximum, and average of all the variables

	Min	Mean	Max
Delivery speed	31.4	50	67.6
Price level	37.2	50	73.7
Price flexibility	33.9	50	61.9
Manufacturer image	33.1	50	71.9
Service	31.7	50	70.8
Salesforce image	34.7	50	69.1
Product quality	31.7	50	70.2
Satisfaction level	45.3	50	72.4

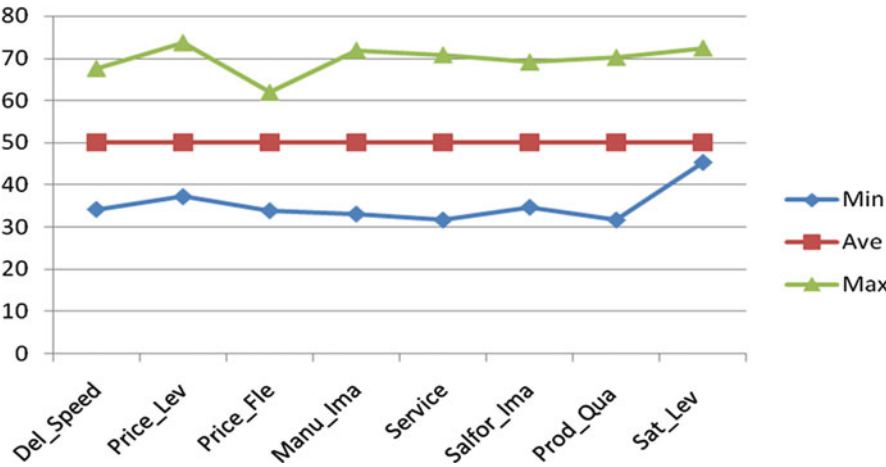


Fig. 2.8 Profile chart of the company's performance indicators

Step 4: Use Excel graphic functionality for developing line diagram to show the company's profile on its various parameters. The profile chart so prepared is shown in Fig. 2.8.

Summary of the SPSS Commands

1. Start SPSS by using the following commands:

Start → All Programs → IBM SPSS Statistics → IBM SPSS Statistics 20

2. Click **Variable View** tag and define the variables *Del_Speed*, *Price_Lev*, *Price_Fle*, *Manu_Ima*, *Service*, *Salfor_Ima*, *Prod_Qua*, and *Sat_Lev* as a scale variable.
3. Once the variables are defined, then type the data for these variables by clicking **Data View**.
4. In the data view, follow the below-mentioned command sequence for computing descriptive statistics:
Analyze → **Descriptive Statistics** → **Frequencies**
5. Select all the variables from left panel to the right panel for computing various descriptive statistics.
6. Click the tag **Statistics** and check the options under the headings “Percentile Values,” “Central Tendency,” “Dispersion,” and “Distribution.” Press **Continue**.
7. Click the **Charts** option and select the required chart, if graph is required for all the variables.
8. Click **OK** to get the output for descriptive statistics.

Exercise

Short-Answer Questions

Note: Write answer to each of the following questions in not more than 200 words:

- Q.1. If average performance of two groups is equal, can it be said that both the groups are equally good?
- Q.2. What do you mean by absolute and relative variability? Explain by means of examples.
- Q.3. What is coefficient of variation? In what situation it should be computed? With the help of the following data on BSE quote during last trading sessions, can it be concluded that the group WIPRO's quotes were more variable than GAIL?

	Group WIPRO	Group GAIL
Mean	5400	170
SD	200	40

- Q.4. Is there any difference between standard error of mean and error in computing the mean? Explain your answer.
- Q.5. If skewness of a set of data is zero, can it be said to be normally distributed? If yes, how? And if no, how it can be checked for its normality?
- Q.6. If performance of a student is 96th percentile in a particular subject, can it be concluded that he is very intelligent in that subject? Explain your answer.
- Q.7. What is a quartile measure? In what situation it should be used?

Multiple-Choice Questions

Note: Question no. 1–10 has four alternative answers for each question. Tick mark the one that you consider the closest to the correct answer.

1. If a researcher is interested to know the number of employees in an organization belonging to different regions and how many of them have opted for club memberships, the study may be categorized as
 - (a) Descriptive
 - (b) Inferential
 - (c) Philosophical
 - (d) Descriptive and inferential both
2. Choose the correct sequence of commands to compute descriptive statistics.
 - (a) Analyze -> Descriptive Statistics -> Frequencies
 - (b) Analyze -> Frequencies -> Descriptive Statistics
 - (c) Analyze -> Frequencies
 - (d) Analyze -> Descriptive Statistics
3. Which pair of statistics are nonparametric statistics?
 - (a) Mean and median
 - (b) Mean and SD
 - (c) Median and SD
 - (d) Median and Q.D.
4. Standard error of mean can be defined as
 - (a) Error in computing mean
 - (b) Difference in sample and population mean
 - (c) Variation in the mean values among the samples drawn from the same population
 - (d) Error in measuring the data on which mean is computed
5. The value of skewness for a given set of data shall be significant if
 - (a) Skewness is more than twice its standard error.
 - (b) Skewness is more than its standard error.
 - (c) Skewness and standard error are equal.
 - (d) Skewness is less than its standard error.
6. Kurtosis in SPSS is assessed by
 - (a) β_2
 - (b) $\beta_2 + 3$
 - (c) $\beta_2 - 3$
 - (d) $2 + \beta_2$

7. In order to prepare the profile chart, minimum scores for each variable are converted into
 - (a) Percentage
 - (b) Standard score
 - (c) Percentile score
 - (d) Rank
8. While selecting option for percentile in SPSS, cut points are used for
 - (a) Computing Q_1 and Q_3
 - (b) Preparing the percentile at deciles points only
 - (c) Cutting Q_1 and Q_3
 - (d) Computing the percentiles at fixed interval points
9. If IQ of a group of students is positively skewed, what conclusions could be drawn?
 - (a) Most of the students are less intelligent.
 - (b) Most of the students are more intelligent.
 - (c) There are equal number of high and low intelligent students.
 - (d) Nothing can be said about the intelligence of the students.
10. If the data is platykurtic, what can be said about its variability?
 - (a) More variability exists.
 - (b) Less variability exists.
 - (c) Variability is equivalent to normal distribution.
 - (d) Nothing can be said about the variability.

Assignment

1. Following table shows the data on different abilities of employees in an organization. Compute various descriptive statistics and interpret its findings.

Data on different abilities of employees

Define problems	Supervise others	Make decisions	Build consensus	Facilitate decision-making	Work on a team
.81	.84	.80	.89	.79	.72
.45	.31	.29	.37	.21	.12
.87	.79	.90	.88	.67	.50
.78	.71	.84	.92	.82	.62
.65	.59	.72	.85	.81	.56
.56	.55	.62	.71	.73	.61

2. Following are the grades of ten MBA students in 10 courses. Compute various descriptive statistics and interpret your findings.

Data on grades of MBA students in ten courses

S.N.	FACTG	MACTG	ECON	FIN	MKTG	ENVIR	MIS	QM	OPSM	OB
1	7	1	7	6	6	6	7	5	5	6
2	7	6	7	7	6	5	6	7	4	7
3	3	6	6	6	6	4	5	4	6	7
4	8	7	8	6	7	8	7	8	9	6
5	5	3	5	7	5	8	6	6	4	8
6	3	3	3	3	3	5	6	7	7	7
7	4	7	6	5	8	6	5	4	4	6
8	5	6	8	7	6	7	8	7	5	5
9	6	5	7	8	5	6	5	8	7	7
10	7	6	5	8	6	4	8	6	7	8

FACTG financial accounting for managers, *MACTG* management accounting, *ECON* economic environment of business, *FIN* managerial finance, *MKTG* marketing management, *ENVIR* business environment, *MIS* management information systems, *QM* quantitative methods, *OPSM* operations management, *OB* organizational behavior

Answers of Multiple-Choice Questions

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