UNIT 10: UNIVARIATE AND BIVARIATE ANALYSIS OF DATA

UNIT STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.3 Descriptive vs Inferential Analysis

10.4 Descriptive Analysis of Univariate data

10.5 Measures of Central Tendency

10.6 Measures of Dispersion;

10.7 Graphical Analysis of Nominal scale data

10.8 Descriptive Analysis of Bivariate data

10.9 Analysis of Nominal scale data with multiple category responses

10.10 Let Us Sum Up

10.11 Further Reading

10.12 Answers to Check Your Progress

10.13 Model Questions

10.1 LEARNING OBJECTIVES

After going through this unit, you will be able to

- difference between descriptive and inferential Analysis
- analysis of univariate data
- different measures of central tendency
- different measures of dispersion
- graphical analysis of data
- analysis of bivariate data
- analysis of nominal scale data

10.2 INTRODUCTION

In the previous unit we disccussed the concepts of data and coding of data, in this unit we will discuss the analysis of data. Data analysis is the process of systematically applying statistical and/or logical techniques to describe and evaluate data. Data from various sources is gathered, reviewed, and

then analyzed to form some sort of finding or conclusion. Different data analysis techniques can be used based on the objective of the research or on the type of data gathered, as for example, univariate or bivariate or multivariate. Again, analysis of two kinds, viz. Descriptive and inferential analysis.Let us discuss these concepts in the following sections:

10.3 DESCRIPTIVE VS INFERENTIAL ANALYSIS

Descriptive analysis quantitatively described the features of data sample, while inferential analysis made inferences about the populations from which samples were drawn.

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with the simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

By inferential statistics we make judgements of the probability that an observed difference between groups (samples) is a dependable one or one that might have happened by chance in this study.

10.4 DESCRIPTIVE ANALYSIS OF UNIVARIATE DATA

When one measurement is made on each observation, univariate analysis is applied. Univariate is the simplest form of analyzing data. “Uni” means “one”, so in other words univariate data means data with one variable. The purpose of univariate data is to describe the pattern of the variable. For example, the age of 20 students, the height of 30 children, etc.

Some ways to describe the patterns found in univariate data include central tendency (mean, median and mode); range, variance, quartiles etc. Some other ways to describe the data through chart or graph are: bar charts, pie charts, histograms, frequency polygons, etc. There are three major characteristics of a single variable that we tend to look at:

(i) The distribution
(ii) The central tendency
(iii) The dispersion

Distribution: The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value. For instance, the distribution of college students year wise, or if we describe gender by listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. But if we have large number of possible values, we group the raw scores into categories according to ranges of values. This is known as frequency distribution.

One of the most common ways to describe a single variable is with a ***frequency distribution***. Depending on the particular variable, all of the data values may be represented, or may be grouped the values into categories first. Frequency distributions can be depicted in two ways, as a table or as a graph.

Distributions may also be displayed using percentages. For example, we could use percentages to describe the:

- percentage of people in different income levels
- percentage of people in different age ranges

10.5 MEASURES OF CENTRAL TENDENCY

Central Tendency: The central tendency of a distribution is an estimate of the "centre" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

Mean or average is probably the most commonly used method of describing central tendency. To compute the mean we need to add up all the values and divide by the number of values. For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167, so the mean is 167/8 = 20.875.

Median is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the centre of the sample. For example, if there are 500 scores in the list, score numbered 250 would be the median. If we order the 8 scores shown above, we would get:

15, 15,15,20,20,21,25,36

There are 8 scores and score number 4 and number 5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, then the average between the two scores may be regarded as median.

Mode is the most frequently occurring value in the set of scores. To determine the mode, we might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

Notice that for the same set of 8 scores we got three different values — 20.875, 20, and 15 — for the mean, median and mode respectively. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other.



CHECK YOUR PROGRESS

Q1: Define Central Tendency

……

……

Q2: Fill in the blanks

a. To compute the mean we need to add up all the values and divide by the ……...

b. **Median** is the score found at the exact ……. of the set of values.

10.6 MEASURES OF DISPERSION

Dispersion: Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 - 15 = 21.

Standard Deviation is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The Standard Deviation shows the relation that set of scores has to the mean of the sample. It is denoted by ó and is given by the formula:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Again let us take the set of scores: **15, 15,15,20,20,21,25,36**

Calculation of Standard Deviation

Sl. No.	x	$(x - \bar{x})^2$
1	15	34.516
2	15	34.516
3	15	34.516
4	20	0.766
5	20	0.766
6	21	0.016
7	25	17.016
8	36	228.766
Total	167	350.875

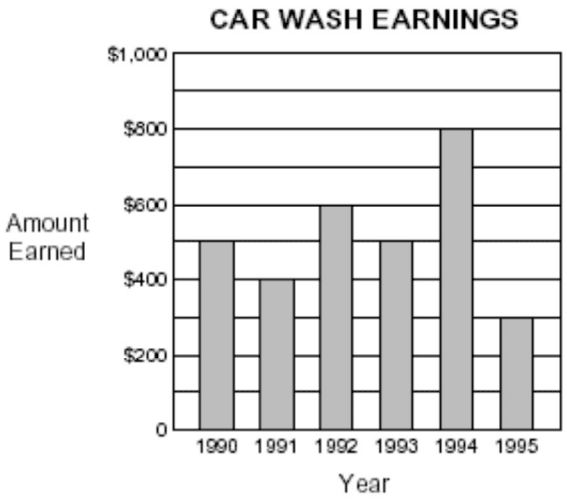
$\bar{x} = 20.875$

ó = 6.622

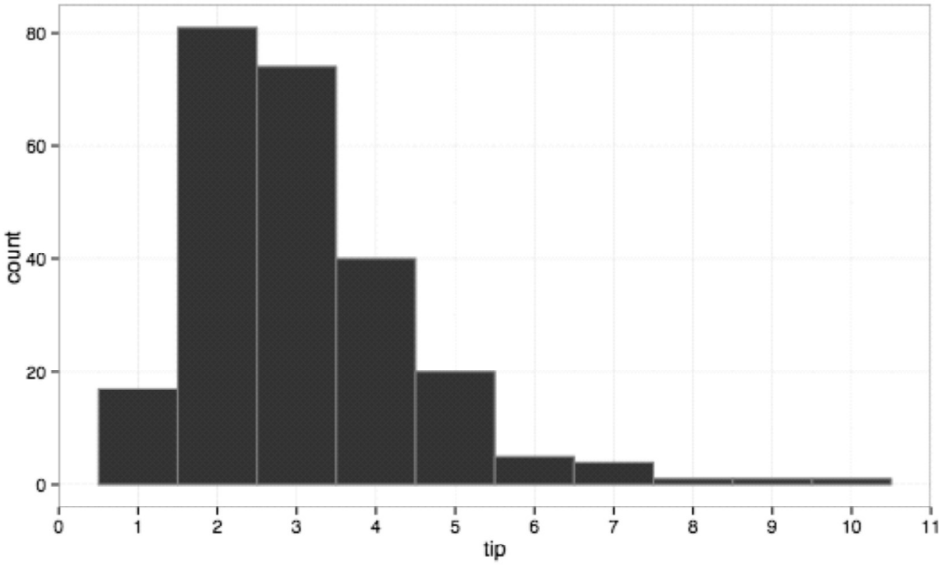
10.7 GRAPHICAL ANALYSIS OF NOMINAL SCALE DATA

Under graphical analysis following charts are used: bar charts, pie charts, histograms, frequency polygons:

Bar Charts: A bar chart or bar graph is a chart that presents grouped data with rectangular bars with lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a Line graph if bars are replaced by lines.



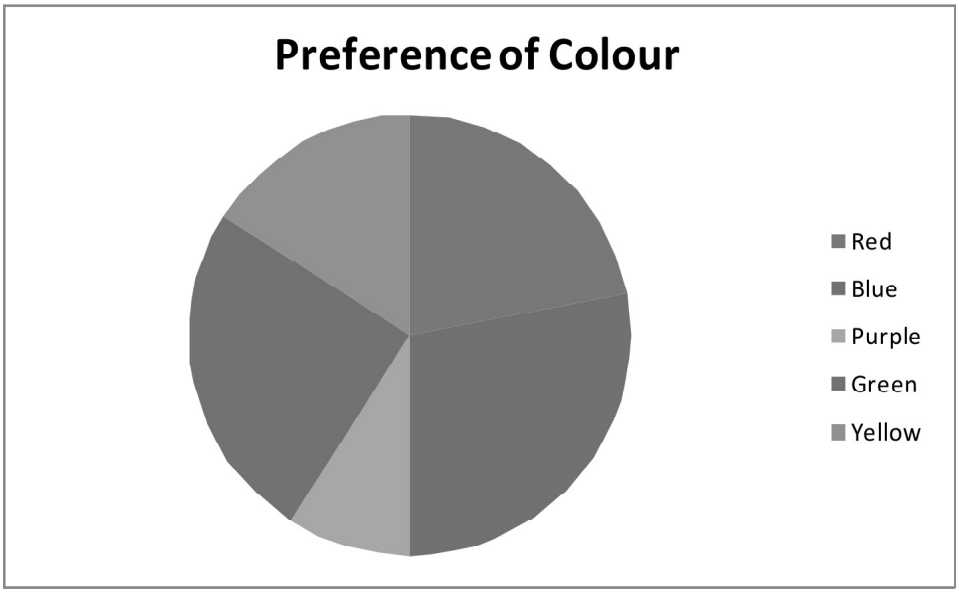
Histogram: Histogram is a graphical display of data using bars of different heights. It is drawn by placing class intervals on the x–axis and corresponding frequencies on y–axis. Thus we get a series of rectangles one attached to other. Histograms are commonly used in statistics to demonstrate how many of a certain type of variable occurs within a specific range. For example, a census focused on the demography of a country may use a histogram of how many people there are between the ages of 0 and 10, 11 and 20, 21 and 30, 31 and 40, 41 and 50 etc.



Pie Charts: Pie charts or circle graphs as they are sometimes known are different from the other types of graphs. This graph does not use a set of axes to plot points. Pie charts display percentages. A Pie Chart (or Pie Graph) is a circular chart divided into sectors; each sector shows the relative size of each value.

Example:

Preference of Favourite Colour		
Colour	**Frequency**	**Percentage**
Red	22	79.2
Blue	28	100.8
Purple	9	32.4
Green	25	90
Yellow	16	57.6





CHECK YOUR PROGRESS

Q3: Define Histogram.

.....

.....

.....

Q4: Compute standard deviation for the following set of data:

25, 36, 46, 58, 39, 43, 60, 40, 55, 45

10.8 DESCRIPTIVE ANALYSIS OF BIVARIATE DATA

Bivariate data is that where exactly two measurements are made on each observation. The two measurements will be called X and Y. Since X and Y are obtained for each observation, the data for one observation is the air (X, Y). For example height and weight of 20 college students, price and demand of a good for 10 weeks, Stock market valuation (X) and quarterly corporate earnings (Y) are recorded for each company in a sample etc.

Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables for the purpose of

determining the empirical relationship between them. There is a difference between bivariate data and two sample data. In two sample data, the X and Y values are not paired, and there aren't necessarily the same number of X and Y values.

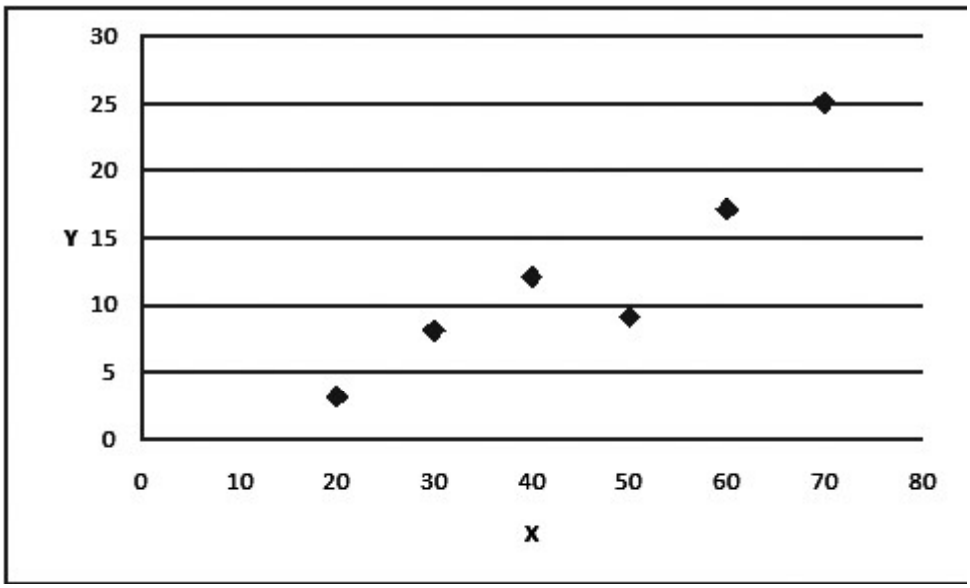
Two-sample data:
Sample 1: 3, 2, 5, 1, 3, 4, 2 ,3
Sample 2: 4, 4, 3, 6, 5

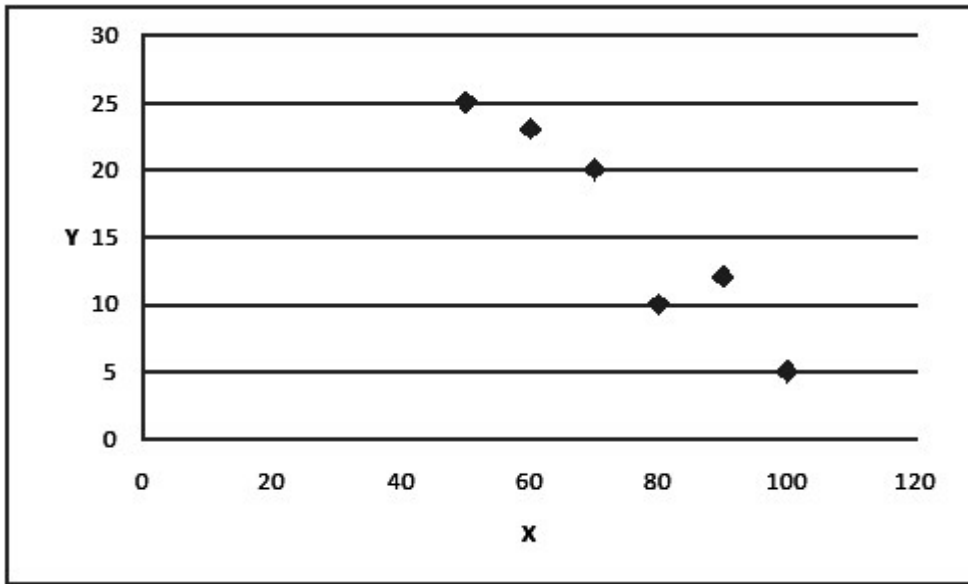
Bivariate data can be written $(X_1, Y_1), (X_2, Y_2), \ldots, (Xn, Yn)$. Each observation is a pair of values, for example (X_3, Y_3) is the third observation. In a bivariate data, the observations are independent of each other, but the two measurements within an observation may not be independent.

Types of Bivariate Analysis

Common types of bivariate analysis include:

1. Scatter Plot: The most important graphical summary of bivariate data is the scatter plot. This is simply a plot of the points (X_i, Yi) in the plane. A simple scatter plot gives an visual idea of the pattern of the variables of interest. It reflects the association between the variables.





2. Correlation Coefficient: A measure that determines the degree to which two variable's movements are associated. Correlation is a term that refers to the strength of a relationship between two variables. A strong, or high, correlation means that two or more variables have a strong relationship with each other, while a weak or low correlation means that the variables are hardly related. Correlation analysis is the process of studying the strength of that relationship with available statistical data.

The most widely used type of correlation coefficient is the Pearson r. Correlation coefficients can range from -1.00 to +1.00. The value of -1.00 represents a perfect negative correlation, which means that as the value of one variable increases, the other decreases. While a value of +1.00 represents a perfect positive relationship, meaning that as one variable increases in value, so does the other. Values like these—of + or - 1.00—signal a perfectly linear relationship between the two variables, so that if you plot the results on a graph it would make a straight line. A value of 0.00 means that there is no relationship between the variables being tested.

Example: Calculate and analyze the correlation coefficient between the number of study hours and the number of sleeping hours of different students.

Number of Study hours	2	4	6	8	10
Number of Sleeping hours	10	9	8	7	6

Solution: Pearson's correlation coefficient is given by

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2}\sqrt{\sum (y - \bar{y})^2}}$$

Calculation:

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-x)^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
2	10	-4	2	16	4	-8
4	9	-2	1	4	1	-2
6	8	0	0	0	0	0
8	7	2	-1	4	1	-2
10	6	4	-2	16	4	-8
$\sum x$ =30	$\sum y$ =40			$\sum (x-\bar{x})^2$ =40	$\sum (y-\bar{y})^2$ =10	$\sum (x-\bar{x})(y-\bar{y})$ =-20

$$\therefore r = \frac{-20}{\sqrt{10X40}} = -1$$

3. Regression Analysis: Scatter plot is giving the pattern of relationship between bivariate data whereas regression analysis is a statistical process for estimating the relationships among variables. Its main focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors').

How to Find the Regression Equation?

In the table below, the x_i column shows scores on the aptitude test. Similarly, the y_i column shows statistics grades. The last two rows show sums and mean scores that we will use to conduct the regression analysis.

	Student	x_i	y_i	$(x_i - x)$	$(y_i - y)$	$(x_i - x)^2$	$(y_i - y)^2$	$(x_i - x)(y_i - y)$
	1	95	85	17	8	289	64	136
	2	85	95	7	18	49	324	126
	3	80	70	2	-7	4	49	-14
	4	70	65	-8	-12	64	144	96
	5	60	70	-18	-7	324	49	126
Sum		390	385			730	630	470
Mean		78	77					

Let us consider the regression equation be a linear equation of the form:

$$\hat{y} = b_0 + b_1 x$$

To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below.

$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$ $b_1 = 470/730 = 0.644$		$b_0 = y - b_1 * x$ $b_0 = 77 - (0.644)(78) =$ 26.768

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.



CHECK YOUR PROGRESS

Q5: What are the different ways of analysing bivariate data?

.....

.....

Q6: Plot a scatter diagram using following data:

X: 29	35	10	37	45	39	57	68	75
Y: 44	53	18	42	55	60	82	85	91

10.9 ANALYSIS OF NOMINAL SCALE DATA WITH MULTIPLE CATEGORY RESPONSES

Multiple response data routinely arise in the study of consumer behaviour. The use of specific products, brands, and distribution channels (e.g., the Internet); magazine readership; product consumption contexts; activities engaged in outside the marketplace (e.g., hobbies); and perceptions about brand attributes are examples of behavioural constructs that are commonly studied by means of multiple response questions. Respondents are directed to consider the response options within a common universe (e.g., brands within a product class, contexts of use for a specific product) that focuses cognitive effort to a particular domain. The use of a nominal scale to capture discrete behaviour (e.g., use, do not use) and perceptions (e.g., provides

benefit, does not provide benefit) further reduces the response effort, there by resulting in a question format that is easy for respondents to use. Surveys containing multiple response questions provide a useful format for collecting behavioural and perceptual data across many response options.

Multiple response data are often collected as part of a descriptive research study in which the goal is to move inductively from data to theory, not from theory to data. When a well-developed theory is available to guide a research effort, richer response scales are often developed, and the domain of the inquiry is focused on specific hypotheses. In contrast, descriptive research typically employs less developed scales and attempts to be more encompassing in its domain. A goal of analyzing multiple response data in descriptive research is to uncover structure in the binary responses that can ultimately suggest new theory. Many techniques are available to marketing researchers for the exploratory analysis of data, particularly for continuous data from elliptically symmetric distributions (e.g., the multivariate normal). Principal component, canonical correlation, and cluster analysis are examples of multivariate data–reduction techniques that can provide insights into the structure underlying high-dimensional data.



10.10 LET US SUM UP

After going through the unit we have come through the following:

- Descriptive analysis quantitatively described the features of data sample, while inferential analysis made inferences about the populations from which samples were drawn.
- Data may be of the type univariate, bivariate or multivariate.
- Some ways to describe the patterns found in univariate data include central tendency (mean, median and mode) and dispersion (range, variance, quartiles etc.) Some other ways to describe the data through chart or graph are: bar charts, pie charts, histograms, frequency polygons, etc.
- Bivariate data is that where exactly two measurements are made on each observation.
- Scatter plot, Correlation and regression analysis are some common form of analysis of bivariate data.

- Multiple response data are often collected as part of a descriptive research study in which the goal is to move inductively from data to theory, not from theory to data.



10.11 FURTHER READING

1. Malhotra. N.K.(2011), *Marketing Research -An Applied Orientation*, Pearson Education



10.12 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q.1: **Central Tendency:** The central tendency of a distribution is an estimate of the "centre" of a distribution of values.

Ans to Q.2: **(a)** To compute the mean we need to add up all the values and divide by the number of values.

(b) Median is the score found at the exact middle of the set of values.

Ans to Q.3: Histogram is a graphical display of data using bars of different heights. It is drawn by placing class intervals on the x–axis and corresponding frequencies on y–axis. Thus we get a series of rectangles one attached to other.

Ans to Q.4: Calculation of Standard Deviation

Sl. No.	x	$(x-\bar{x})^2$
1	25	388.09
2	46	1.69
3	39	32.49
4	60	234.09
5	55	106.09
6	36	75.69
7	58	176.89
8	43	2.89
9	40	22.09
10	45	0.09
Total	447	1040.1

$$\bar{x} = 44.7$$

$$\sigma = \sqrt{(\frac{1}{10} X 1040.1)} = 10.20$$

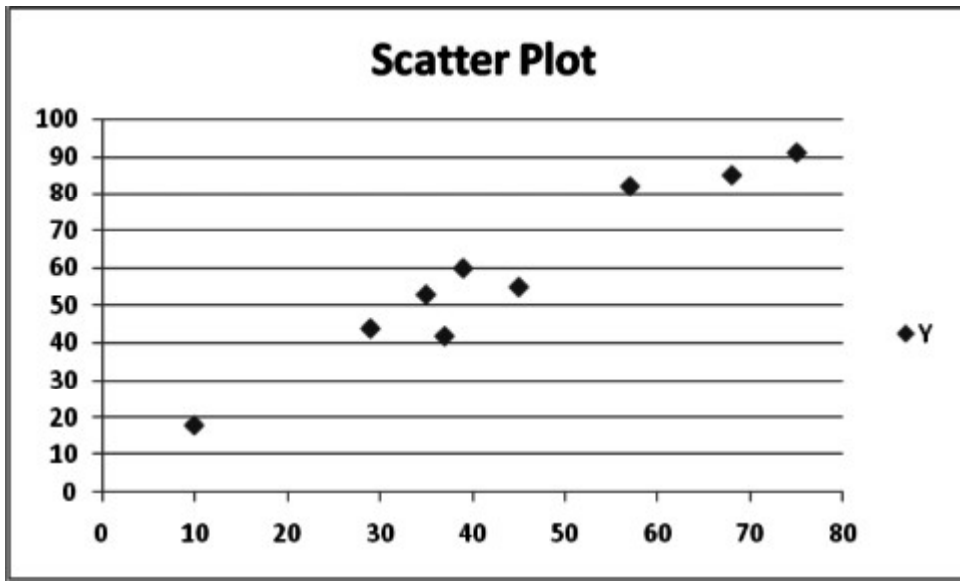
Ans to Q.5: Bivariate analysis is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables for the purpose of determining the empirical relationship between them. There are different methods for analysing bivariate data, viz.

(a) Scatter Plot

(b) Correlation and

(c) Regression Analysis

Ans to Q.6:





10.13 MODEL QUESTIONS

Q.1. What are the measures of central tendency?

Q.2. What is the difference between descriptive analysis and inferential analysis?

Q.3. Explain the requirement of a measure of dispersion?

Q.4. What is regression and what are its importance in the analysis of bivariate data?

Q.5. Write short notes on correlation.

*****