Block 3
Analysis of Quantitative Data

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 7 UNIVARIATE DATA ANALYSIS

Structure

7.0 Objectives

7.1 Introduction

7.2 Exploratory Data Analysis

- 7.2.1 The Five-Number Summary and Boxplots
- 7.2.2 Procedure for Drawing Boxplot and Inferences it reveals
- 7.2.3 Example

7.3 Inferential Statistics: Basic Concepts and Significance of Measures of Central Tendency and Dispersions in Decision Making

- 7.3.1 Parameters and Statistics
- 7.3.2 Random Variables and Probability Distributions
- 7.3.3 Normal Distribution
- 7.3.4 Standard Normal Distribution
- 7.3.5 Application of Standard Normal Distribution in Solving Practical Problems
- 7.3.6 Hypothesis Testing

7.4 Inferential Statistics: Point Estimation and Setting up Confidence Intervals for Population Parameters

- 7.4.1 Point Estimation for Population Mean
- 7.4.2 Confidence Intervals for the Mean When Population Standard Deviation (σ) is Known
- 7.4.3 Confidence Intervals for the Mean When Population Standard Deviation (σ) is Not Known
- 7.4.4 Confidence Intervals for Proportions
- 7.4.5 Confidence Intervals for Variances and Standard Deviations

7.5 Let Us Sum Up

7.6 References

7.7 Answers or Hints to Check Your Progress Exercises

7.0 OBJECTIVES

After going through this unit, you should be able to:

- state the purpose of inferential statistics;
- explain Exploratory Data Analysis;
- discuss the characteristics of normal distribution and its applications in solving the statistical problems;
- state the concept of Hypothesis Testing and the types of errors that emerge in the process and trade off situations between them; and

- appreciate the process of setting the confidence intervals for population parameters.

7.1 INTRODUCTION

In the Unit 2 of this course, you were introduced to the branch of statistics devoted to the summarisation and description of data, called descriptive statistics. You have learned the calculation of various descriptive measures such as averages, measures of variation, and percentiles, etc. It is thus clear that descriptive statistics is much more than just the tabulation of numbers and the graphical presentation of these tabulated numbers. This Unit will basically focus on analysis of data. We begin with exploratory data analysis, where we try to get a feel of prominent features of the data. Then we shall move into more rigorous inferential statistics consisting of methods for drawing and measuring the reliability of conclusions about population.

7.2 EXPLORATORY DATA ANALYSIS

In traditional statistics, data are organised by using a frequency distribution. As discussed in Unit 6, from this distribution, various graphs such as the histogram, frequency polygon, and ogive can be constructed to determine the shape or nature of the distribution. In addition, various statistics such as the mean and standard deviation can be computed to summarise the data. The purpose of traditional analysis is to confirm various conjectures about the nature of the data.

In Exploratory Data Analysis (EDA), the purpose is to examine data to find out what information can be discovered about the data, such as its center and the spread. The measure of central tendency used in EDA is the median. The measure of variation used in EDA is the inter-quartile range $Q_3 - Q_1$. In EDA the data are represented graphically using a boxplot (sometimes called a box-and-whisker plot). Exploratory data analysis was developed by John Tukey and presented in his book Exploratory Data Analysis.

7.2.1 The Five-Number Summary and Boxplots

A boxplot can be used to graphically represent a data set. These plots involve five specific values:

1) The lowest value of the data set (i.e., minimum)
2) Q_1
3) The median (i.e., Q_2)
4) Q_3
5) The highest value of the data set (i.e., maximum)

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

7.2.2 Procedure for Drawing Boxplot and Inferences it reveals

A boxplot is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2. The step by step procedure is outlined below:

1) Find the five-number summary for the data values, that is, the maximum and minimum data values, Q_1 and Q_3, and the median (Q_2).
2) Draw a horizontal axis with a scale such that it includes the maximum and minimum data values.
3) Draw a box whose vertical sides go through Q_1 and Q_3, and draw a vertical line through the median.
4) Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.

The kind of information provided by Boxplot is summarised as below:

- If the median is near the center of the box, the distribution is approximately symmetric.
- If the median falls to the left of the center of the box, the distribution is positively skewed.
- If the median falls to the right of the center, the distribution is negatively skewed.
- If the lines are about the same length, the distribution is approximately symmetric.
- If the right line is larger than the left line, the distribution is positively skewed.
- If the left line is larger than the right line, the distribution is negatively skewed.

If the boxplots for two or more data sets are graphed on the same axis, the distributions can be compared. To compare the averages, use the location of the medians. To compare the variability, use the inter-quartile range (Q_3–Q_1), i.e., the length of the boxes.

7.2.3 Example

A dietitian is interested in comparing the sodium content of real cheese with the sodium content of a cheese substitute. The data for two random samples are shown. Compare the distributions, using boxplots.

Real Cheese				Cheese Substitute			
310	420	45	40	270	180	250	290
220	240	180	90	130	260	340	310

Solution

Step 1: Find Q_1, Median (MD) and Q_3 for the real cheese data.

40 45 90 180 220 240 310 420

Q_1 MD Q_3

$Q_1 = \frac{45+90}{2} = 67.5$ $\quad$ $MD = \frac{180+220}{2} = 200$

$Q_3 = \frac{240+310}{2} = 275$

Step 2: Find Q_1, MD, and Q_3 for the cheese substitute data.

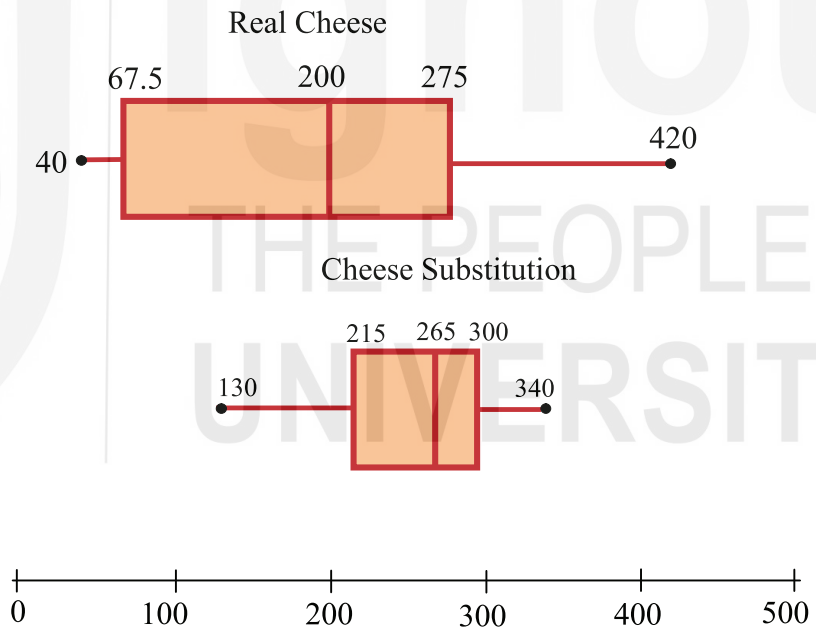
130 180 250 260 270 290 310 340

Q_1 MD Q_3

$Q_1 = \frac{180+250}{2} = 215$ $\quad$ $MD = \frac{260+270}{2} = 265$

$Q_3 = \frac{290+310}{2} = 300$

Step 3: Draw the boxplots for each distribution on the same graph.



Step 4: Compare the plots. It is quite apparent that the distribution for the cheese substitute data has a higher median than the median (265) for the distribution for the real cheese data (200). The variation or spread for the distribution of the real cheese data is larger than the variation for the distribution of the cheese substitute data.

Check Your Progress 1

1) Define the terms 'The Five-Number Summary' and 'Boxplot'.

.....

.....

2) Explain the characteristics of a distribution with the help of boxplot.

.....

.....

.....

.....

7.3 INFERENTIAL STATISTICS: BASIC CONCEPTS AND SIGNIFICANCE OF MEASURES OF CENTRAL TENDENCY AND DISPERSIONS IN DECISION MAKING

Descriptive and inferential statistics are interrelated. ***Descriptive statistics*** consists of collection, organisation, summarisation, and presentation of data, on the other hand, ***inferential statistics*** is concerned with using sample data to make an inference about a population of data. By ***population*** it means, the collection of all the individuals or objects, investigator is primarily interested in for his statistical study. Descriptive statistics— applied to populations— includes construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles you came across in the previous unit. However, one generally does not have access to data from the entire population and thus must base his information about a population on a sample. For example, researcher might be interested in the exam grades of all students in a State. It is not feasible to measure exam grades of all students in whole of the State, so he or she has to restrict to a smaller sample of students (e.g., 100 students), which is used to represent the larger population of all students in the State. Thus, a ***sample*** from statistical population is a subset of the units or elements of a population that is actually collected in the course of an investigation. Inferential statistics involves using these samples to make generalisations about the populations from which the samples were drawn by employing the methods like point estimation, interval estimation and hypothesis testing.

It is almost always necessary to use methods of descriptive statistics to organise and summarise the information obtained from a sample before methods of inferential statistics can be used to make more thorough analysis of the subject under investigation. In order to get sound hold of the methods of inferential statistics, we discuss below some prerequisites.

7.3.1 Parameters and Statistics

Features of the population under investigation can be summarised by numerical ***parameters*** like mean, standard deviation, etc. you came across in the previous unit. A research problem usually involves investigation of the values of these parameters, which are unknown, and thus are estimated from a sample. The sample estimate of the population parameter is called a ***statistic*** which is used to make inference about unknown parameter.

For example, sample mean ($\bar{X}$) and sample standard deviation *S*, the statistics from the sample data, may be used to say something about the respective population parameters, the population mean (μ) and population standard deviation (σ).

Definition: *A parameter is an unknown numerical summary of the population. A statistic is a known numerical summary of the sample which can be used to make inference about parameters.*

Please note: Values of sample statistics ($\bar{X}$ and *S*) depend on the sample selected, whereas the parameters of the population (μ and σ) remain constant. In this sense, sample statistics are called ***random variables***. Now we discuss below what is meant by a random variable, its types and associated probability distributions.

7.3.2 Random Variables and Probability Distributions

By random it means unpredictable; a ***random variable***, usually denoted by capital letters like, X, Y, etc., is a numerical outcome of a random phenomenon (a process for which the outcome cannot be predicted with certainty, for instance, a process of flipping a coin, rolling a die, etc.). Random variables can be either discrete or continuous. A ***discrete*** random variable takes a countable number of values, that is, such variables do not assume all values between any two given values of the variables. For example, on rolling a die 5 times, the number of times an even digit (2, 4 or 6) appears on the top is a discrete random variable which can take values 0, 1, 2, 3, 4, 5 which are finite values, but the variable does not take the values like 1.2, 2.5, 4.6, or any other number than the whole number. On the other hand, a ***continuous*** random variable can assume all values between any two given values of the variables. Examples of continuous variables are the heights of adult men, body temperatures of rats, and cholesterol levels of adults.

Probability plays an important role in inferential statistics. A random phenomenon includes a very long sequence of repeated observations, where each observation may or may not result in some particular outcome. That is, some values of the random variable may be more likely than others. The ***probability*** of a particular outcome is the proportion of times that outcome would occur in a long run of repeated observations (for instance, probability of getting an even digit on rolling a die would be $\frac{3}{6}$ or $\frac{1}{2}$).

Probability distribution of a random variable X lists the possible outcomes along with their associated probabilities. For instance, probability distribution of a discrete random variable X, where X is the digit that appears on rolling a six-sided symmetric die is given by:

X:	1	2	3	4	5	6
P(X):	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

If we denote a possible outcome of a random variable by x_i where i = 1, 2, 3…k and the associated probability by P($X = x_i$) = P(x_i), then the following holds true:

i) $0 \leq P(x_i) \leq 1$, that is, each probability falls between 0 and 1 including 0 and 1;

ii) $\sum_{i=1}^{k} P(x_i) = 1$, the total probability equals 1.

Continuous random variable X, on the other hand, takes all values in some interval of numbers between *a* and *b*. That is, for all x_1 and x_2 (with $x_1 \leq x_2$, such that they themselves form an interval $[x_1, x_2]$) belonging to interval of *a* and *b*, the probability that random variable X falls in any particular interval $[x_1, x_2]$ is a number between 0 and 1, and the probability of the interval [a, b], containing all possible values, equals 1. That is, following should be satisfied:

$$0 \leq P(x_1 \leq X \leq x_2) \leq 1 \text{ and } P(a \leq X \leq b) = 1$$

Unlike probability distribution of a discrete random variable, probability distribution of a continuous random variable cannot be presented in a tabular form. It has either a functional form or a graphical form. Graphically, probability distribution of a continuous random variable is pictured by a probability density curve (Refer Fig. 7.1). A ***probability density curve*** is a smooth continuous curve having area exactly 1 underneath it. The function that defines the curve is called ***probability density function f(x)***, which satisfies the following properties:

i) $f(x) \geq 0$ for all x in the interval $[a, b]$

ii) For $x_1, x_2 \in [a, b]$, where $x_1 \leq x_2$, $P(x_1 \leq X \leq x_2)$ = Area under the probability density curve between the ordinates at $X = x_1$ and $X = x_2$, which in calculus terms is given by $\int_{x_1}^{x_2} f(x)dx$*

Whereas, $P(X = x_1) = 0$, that is, area under the probability density curve and the vertical line at $X = x_1$ is 0. Thus, for a continuous random variable, probability at a point is always zero.

iii) Total area under density curve is 1, that is $\int_a^b f(x)dx = 1$

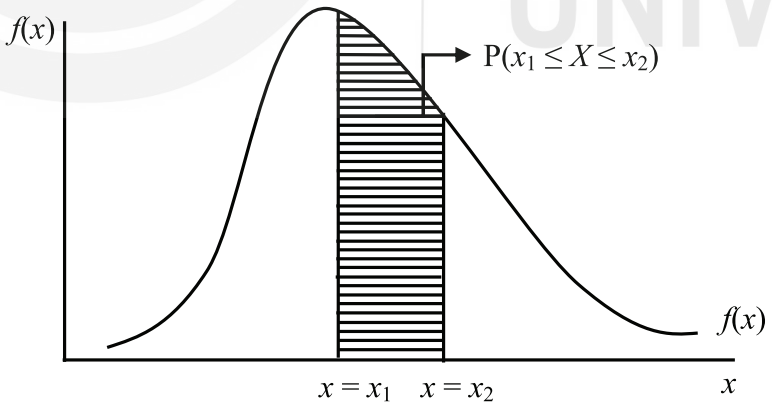


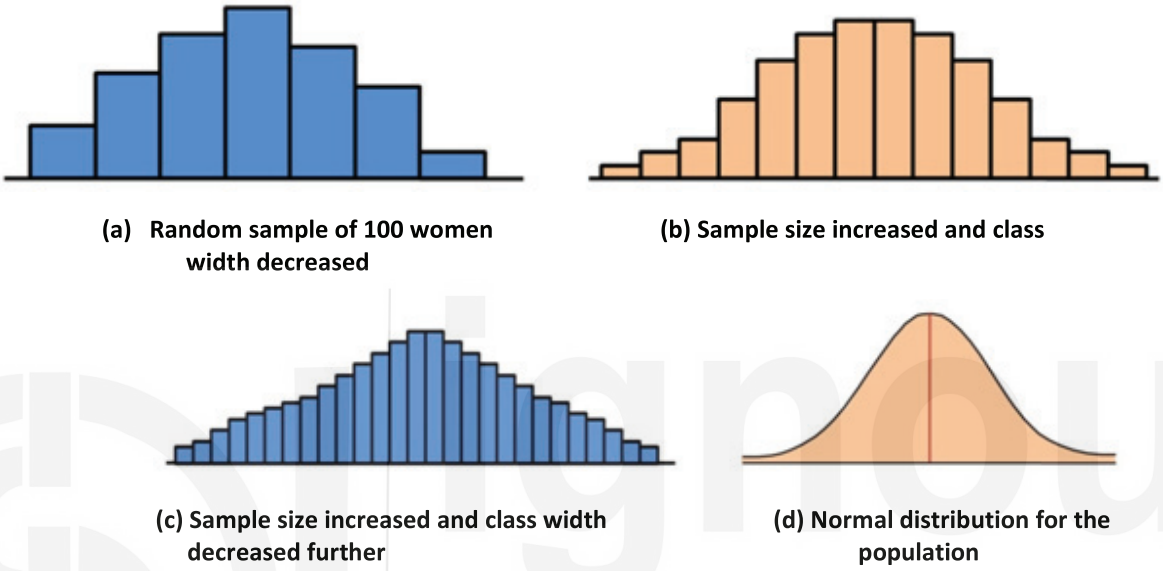
Fig. 7.1: Probability Density Curve

7.3.3 Normal Distribution

One of the most important probability distributions underlying a continuous random variable is *normal probability distribution* or *normal distribution*.

* You came across this tool of definite integrals in your Mathematical Methods in Economics-I course (BECC 102) of Semester 1.

This distribution is significantly used in statistical analysis as it approximates well the distributions of many continuous variables. For example, if a researcher selects a random sample of 100 adult women, measures their heights, and constructs a histogram, the researcher gets a graph similar to the one shown in Fig. 7.2 (a). Now, if the researcher increases the sample size and decreases the width of the classes, the histograms will look like the ones shown in Fig. 7.2 (b) and (c). Finally, if it were possible to measure exactly the heights of all adult females in India and plot them, the histogram would approach what is called a normal distribution, shown in Fig. 7.2 (d).



(a) Random sample of 100 women

(b) Sample size increased and class width decreased

(c) Sample size increased and class width decreased further

(d) Normal distribution for the population

Fig. 7.2

Graphically, a normal distribution is represented by a bell-shaped curve (Refer Fig. 7.3). Similar to a discrete random variable, the values of a continuous random variable referred to as scores or observations are represented on the x-axis. The area under the curve over any interval represents the proportion of scores in that interval. Mean (μ) represents the true or normal value of the score and deviations from this are errors. In intervals around the mean, where the curve is highest, the scores are most crowded (dense). Towards the ends of the curve, the height is lower; the scores become less crowded the further from the mean we go. That is, observations around the mean are more likely to occur than observations further from it.

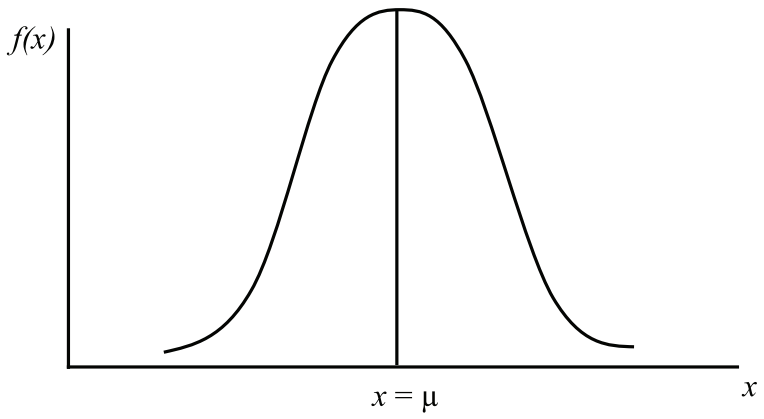


Fig. 7.3: Normal Distribution Curve

A continuous random variable X following normal distribution has two defining parameters—the mean μ and the standard deviation σ. Denoted by $X \sim N(\mu, \sigma)$, the underlying probability density function $f(x)$ of a continuous random variable X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty$$

The shape and position of a normal distribution curve depend on its two parameters (μ and σ). The value of the mean fixes the location of the normal curve, where it is centered, whereas the value of the standard deviation determines the spread; the bigger σ, the more spread out or flatter the curve. The properties of normal distribution, which are its strength, are given below:

1) A normal distribution curve is bell-shaped.
2) The mean, median, and mode are equal and are located at the center of the distribution.
3) A normal distribution curve is uni-modal (i.e., it has only one mode).
4) The curve is symmetric about the mean, which is equivalent to saying that its shape is the same on both sides of a vertical line passing through the center.
5) The curve is continuous; that is, there are no gaps or holes. For each value of *X*, there is a corresponding value of *Y*.
6) The curve never touches the *x* axis. Theoretically, no matter how far in either direction the curve extends, it never meets the *x* axis—but it gets increasingly closer.
7) The total area under a normal distribution curve is equal to 1.00, or 100%. This fact may seem unusual, since the curve never touches the *x* axis, but one can prove it mathematically by using calculus.
8) The area under the part of a normal curve that lies within 1 standard deviation of the mean is approximately 0.68, or 68%; within 2 standard deviations, about 0.95, or 95%; and within 3 standard deviations, about 0.997, or 99.7%. See Fig.7.4, which also shows the area in each region.

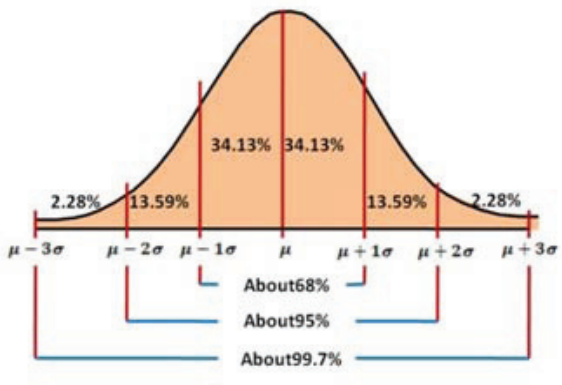


Fig. 7.4

This simply means, area under the normal curve within the range $\mu \pm \sigma$ covers 68.26% of the observations, or in other words, P(μ – $\sigma \leq X \leq$ μ + σ) = 0.68. Similarly, P (μ – 2$\sigma \leq X \leq$ μ + 2σ) = 0.95 and P(μ – 3$\sigma \leq X \leq$ μ + 3σ) = 0.997.

Rationale behind using Normal Distribution

Like we discussed in the beginning, the task performed under inferential analysis is to make inferences from a sample to a parent population. A population is summarised by parameters such as the mean and variance, and a sample selected from the parent population gives the value of sample statistic of the population parameters. We also know that mean of any one sample is likely to differ from the mean of a second sample from the same population. So the sample mean is a random variable, which can take on many different values, with their associated probabilities giving rise to a distribution. This distribution is called the distribution of the sampling mean, whose shape when the sample size is sufficiently large, approximates a normal distribution curve (regardless of the shape of the parent population). This implies, if the parent population is normally distributed then the distribution of the sampling mean is exactly normal. But, if parent population is not normally distributed, distribution of sample mean is approximately normal if the sample size is sufficiently big. This last result comes from a profound and far reaching theorem called the Central Limit Theorem.

Central Limit Theorem

Briefly, according to the central limit theorem, if a random variable is the sum of *n*, independent, identically distributed, non-normal random variables, then its distribution approaches normal as *n* approaches infinity. The following corollary results from the central limit theorem (proof of which is beyond the scope of this course):

The distribution of the sample mean approaches the normal distribution as the sample size n increases, even if the parent distribution from which the samples are drawn is not normal.

The central limit theorem explains why the normal distribution is linked to so many random phenomena.

7.3.4 Standard Normal Distribution

Since each normally distributed variable has its own mean and standard deviation, the shape and location of these curves will vary according to the value of these parameters. In practical applications, then, you would have to have a table of areas under the curve for each variable. To simplify this situation, statisticians use what is called the standard normal distribution.

*The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1.*The standard normal distribution is shown in Fig. 7.5.



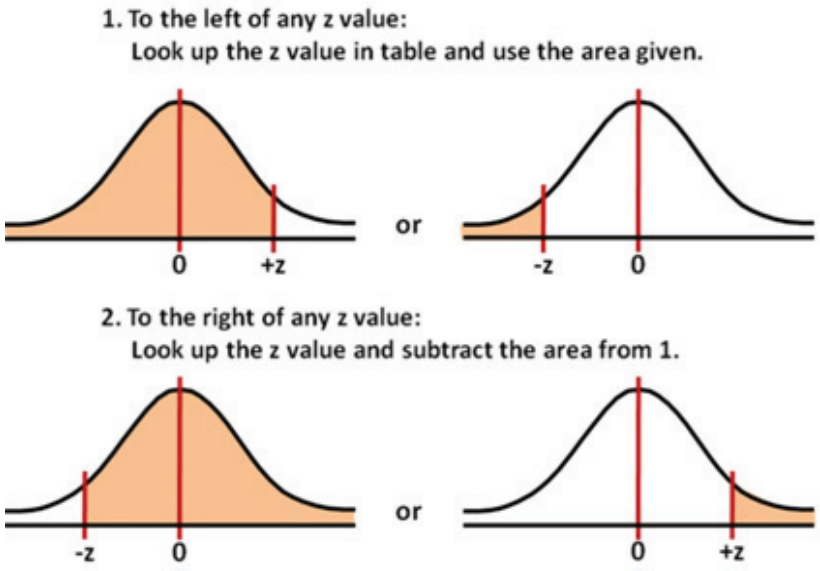
Fig. 7.5

7.3.5 Application of Standard Normal Distribution in Solving Practical Problems

In the figure given above, the values under the curve indicate the proportion of area in each section. All normally distributed variables can be transformed into the standard normally distributed variable by using the formula for the standard score:

$$z = \frac{value - mean}{standard\ deviation} \quad \text{or} \quad z = \frac{X - \mu}{\sigma}$$

As stated earlier, the area under a normal distribution curve is used to solve practical application problems, such as finding the percentage of adult women whose height is between 5 feet 4 inches and 5 feet 7 inches, or finding the probability that a new battery will last longer than 4 years. Hence, it is necessary to understand the procedure for finding the area under the standard normal distribution curve for any z value. Once the X values are transformed by using the above formula, they are called z values. The z value is actually the number of standard deviations that a particular X value is away from the mean. The z-table on Standard Normal Variate gives the area (to four decimal places) under the standard normal curve for any z value from – 3.49 to + 3.49 at various levels of significance. For the solution of problems using the standard normal distribution, a procedure is given in the following Fig. 7.6.



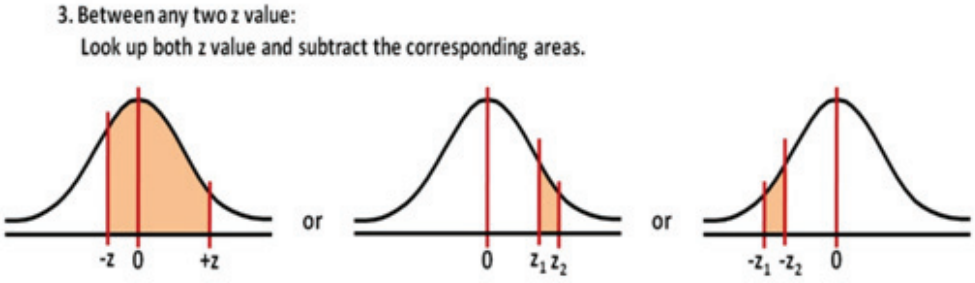


Fig. 7.6

The Figure usually show the area between 0 to some +z. So the area to the right of that +z = .5 – area between 0 to that +z. This will also be the area to the left of —z. Total area to the left of a given +z will be 0.5 + area between 0 to +z.

7.3.6 Hypothesis Testing

Researchers are interested in answering many types of questions, e.g. an educator might wish to see whether a new teaching technique is better than a traditional one. This type of question can be addressed through statistical hypothesis testing, which is a decision-making process for evaluating claims about a population. In hypothesis testing, the researcher must define the population under study, state the particular hypotheses that will be investigated, give the significance level, select a sample from the population, collect the data, perform the calculations required for the statistical test, and reach a conclusion. In inferential statistics, three methods used to test hypotheses are (i) the traditional method, (ii) the p-value method and (iii) the confidence interval method.

Steps in Hypothesis Testing—Traditional Method

Every hypothesis-testing situation begins with the statement of a hypothesis. *A statistical hypothesis is a conjecture about a population parameter. This conjecture may or may not be true.* There are two types of statistical hypotheses for each situation:

i) the null hypothesis; and

ii) the alternative hypothesis.

i) **The null hypothesis**, symbolised by H_0, is a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters.

ii) **The alternative hypothesis**, symbolised by H_1, is a statistical hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters.

As an illustration of how hypotheses should be stated, three different statistical studies will be used as examples.

i) Situation A: A medical researcher is interested in finding out whether a new medication will have any undesirable side effects. The researcher is particularly concerned with the pulse rate of the patients who take the medication. Will the pulse rate increase, decrease, or remain unchanged after a patient takes the medication? Since the researcher knows that the mean pulse rate for the population under study is 82 beats per minute, the hypotheses for this situation are: H_0: $\mu = 82$ and H_1:$\mu \neq 82$. The null hypothesis specifies that the mean will remain unchanged, and the alternative hypothesis states that it will be different. This test is called a two-tailed test, since the possible side effects of the medicine could be to raise or lower the pulse rate.

ii) Situation B: A chemist invents an additive to increase the life of an automobile battery. If the mean lifetime of the automobile battery without the additive is 36 months, then the hypotheses are: H_0: $\mu = 36$ and H_1: $\mu > 36$. In this situation, the chemist is interested only in increasing the lifetime of the batteries, so the alternative hypothesis is that the mean is greater than 36 months. This test is called right-tailed, since the interest is in an increase only.

iii) Situation C: A contractor wishes to lower heating bills by using a special type of insulation in houses. If the average of the monthly heating bills is Rs. 780, the hypotheses about heating costs with the use of insulation are H_0: $\mu = 780$ and H_1: $\mu < 780$. This test is a left-tailed test, since the contractor is interested only in lowering heating costs.

To state hypotheses correctly, researchers must translate the conjecture or claim from words into mathematical symbols. The basic symbols used are as follows:

Equal to	=	Greater than	>
Not equal to	≠	Less than	<

The null and alternative hypotheses are stated together, and the null hypothesis contains the 'equal to' sign, as shown (where k represents a specified number).

Two-tailed test	Right-tailed test	Left-tailed test
$H_0 = k$	$H_0 = k$	$H_0 = k$
$H \neq k$	$H_0 > k$	$H_0 < k$

After stating the hypothesis, the researcher designs the study. The researcher selects the correct statistical test, chooses an appropriate level of significance, and formulates a plan for conducting the study. In situation A, for instance, the researcher will select a sample of patients who will be given the drug. After allowing a suitable time for the drug to be absorbed, the researcher will measure each person's pulse rate. It may be noted that when samples of a specific size are selected from a population, the means of these

samples will vary about the population mean, and the distribution of the sample means will be approximately normal when the sample size is 30 or more. So even if the null hypothesis is true, the mean of the pulse rates of the sample of patients will not, in most cases, be exactly equal to the population mean of 82 beats per minute.

There are two possibilities. Either the null hypothesis is true, and the difference between the sample mean and the population mean is due to chance; or the null hypothesis is false, and the sample came from a population whose mean is not 82 beats per minute but is some other value that is not known. The farther the sample mean is away from the population mean, the more evidence there would be for rejecting the null hypothesis. The probability that the sample came from a population whose mean is 82 decreases as the distance or absolute value of the difference between the means increases. If the mean pulse rate of the sample were, say, 83, the researcher would probably conclude that this difference was due to chance and would not reject the null hypothesis. But if the sample mean were, say, 90, then in all likelihood the researcher would conclude that the medication increased the pulse rate of the users and would reject the null hypothesis. The question is, where does the researcher draw the line? This decision is not made on feelings or intuition; it is made statistically. That is, the difference must be significant and in all likelihood not due to chance. Here is where the concepts of statistical test and level of significance are used.

In this type of statistical test, the mean is computed for the data obtained from the sample and is compared with the population mean. Then a decision is made to reject or not reject the null hypothesis on the basis of the value obtained from the statistical test. If the difference is significant, the null hypothesis is rejected. If it is not, then the null hypothesis is not rejected. In the hypothesis-testing situation, there are four possible outcomes. In reality, the null hypothesis may or may not be true, and a decision is made to reject or not reject it on the basis of the data obtained from a sample. The four possible outcomes are shown in Figure given below. Notice that there are two possibilities for a correct decision and two possibilities for an incorrect decision.

	H_0 true	H_0 False
Reject H_0	**Type I** error α	Correct decision $1-\beta$
Do not reject H_0	Correct decision $1-\alpha$	**Type II** error β

If a null hypothesis is true and it is rejected, then a *type I error* is made. In situation A (as given on page 177), for instance, the medication might not significantly change the pulse rate of all the users in the population; but it might change the rate, by chance, of the subjects in the sample. In this case, the researcher will reject the null hypothesis when it is really true, thus committing a type I error. On the other hand, the medication might not change the pulse rate of the subjects in the sample, but when it is given to the general population, it might cause a significant increase or decrease in the pulse rate of users. The researcher, on the basis of the data obtained from the sample, will not reject the null hypothesis, thus committing a *type II error.*

In situation B, the additive might not significantly increase the lifetimes of automobile batteries in the population, but it might increase the lifetimes of the batteries in the sample. In this case, the null hypothesis would be rejected when it was really true. *This would be a type I error.* On the other hand, the additive might not work on the batteries selected for the sample, but if it were to be used in the general population of batteries, it might significantly increase their lifetimes. The researcher, on the basis of information obtained from the sample, would not reject the null hypothesis, *thus committing a type II error.*

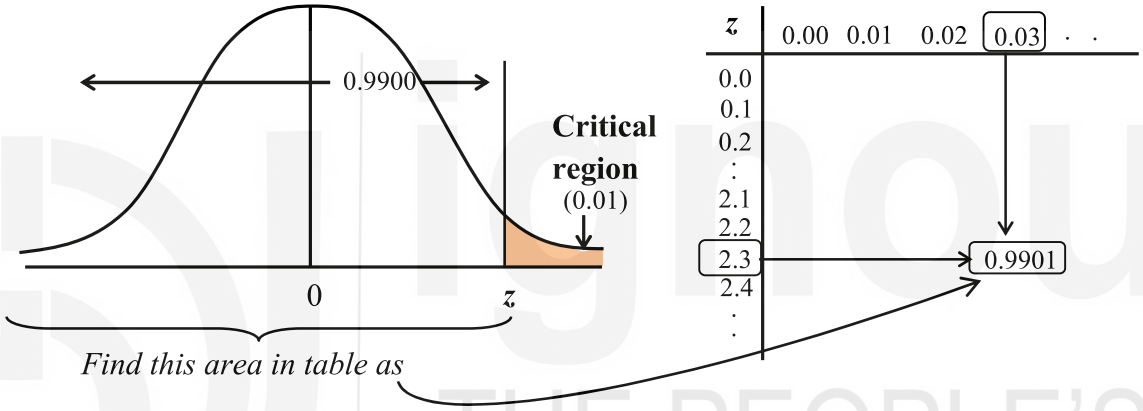
But the decision to reject or not reject the null hypothesis does not prove anything. The decision, then, is made on the basis of probabilities. That is, when there is a large difference between the mean obtained from the sample and the hypothesised mean, the null hypothesis is probably not true. The question is, How large a difference is necessary to reject the null hypothesis? Here is where the level of significance is used. The level of significance is the maximum probability of committing a type I error. This probability is symbolised by α (Greek letter alpha). That is, P (type I error) = α. The probability of a type II error is symbolised by β, (the Greek letter beta). That is, P(type II error) = β. In most hypothesis-testing situations, β cannot be easily computed; however, α and β are related in that decreasing one increases the other.

Statisticians generally agree on using three arbitrary significance levels: the 0.10, 0.05, and 0.01 levels. That is, if the null hypothesis is rejected, the probability of a type-I error will be 10%, 5%, or 1%, depending on which level of significance is used. Another way of putting it: When α = 0.10, there is a 10% chance of rejecting a true null hypothesis; when α= 0.05, there is a 5% chance of rejecting a true null hypothesis; and when α = 0.01, there is a 1% chance of rejecting a true null hypothesis. In a hypothesis-testing situation, the researcher decides what level of significance to use. It can be any level, depending on the seriousness of the type I error. After a significance level is chosen, a critical value is selected from a table for the appropriate test. If a z test is used, for example, the z table is consulted to find the critical value. The critical value determines the critical (rejection) and noncritical (acceptance) regions.

The critical value can be on the right side of the mean or on the left side of the mean for a one-tailed test. Its location depends on the inequality sign of

the alternative hypothesis. For example, in situation B, where the chemist is interested in increasing the average lifetime of automobile batteries, the alternative hypothesis is H_1: $\mu>36$. Since the inequality sign is >, the null hypothesis will be rejected only when the sample mean is significantly greater than 36. Hence, the critical value must be on the right side of the mean. Therefore, this test is called a right-tailed test.

To obtain the critical value, the researcher must choose an alpha level. In situation B, suppose the researcher chose $\alpha= 0.01$. Then the researcher must find a z value (from z-table) such that 1% of the area falls to the right of the z value and 99% falls to the left of the z value, as shown in Fig. 7.7 (a) below. Next, the researcher must find the area value in z-table closest to 0.9900. The critical z value is 2.33, since that value gives the area closest to 0.9900 (that is, 0.9901), as shown in Fig. 7.7 (b) as below:



(a) The critical region **(b) The critical value from z-table E**

Fig. 7.7

The critical and non-critical regions and the critical value are shown in the following Fig. 7.8.

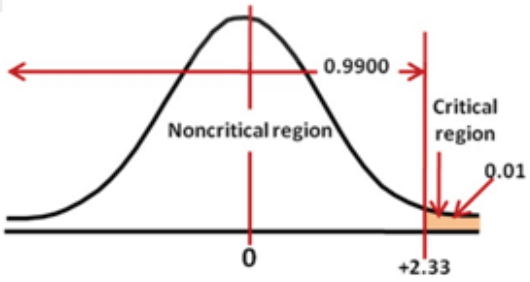


Fig. 7.8

When a researcher conducts a two-tailed test, as in situation A, the null hypothesis can be rejected when there is a significant difference in either direction, above or below the mean. For a two-tailed test, then, the critical region must be split into two equal parts. If $\alpha = 0.01$, then one-half of the area, or 0.005, must be to the right of the mean and one half must be to the left of the mean, as shown in Fig.7.9.

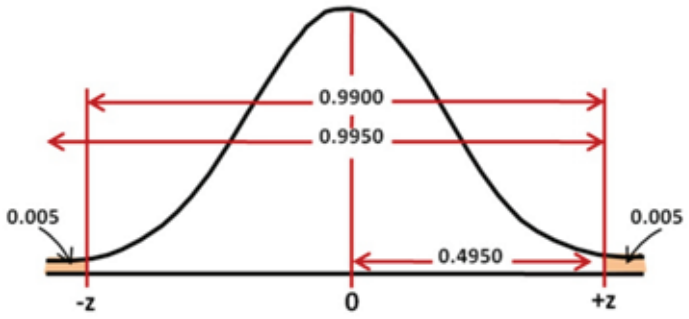


Fig. 7.9

In this case, the z value on the left side is found by looking up the z value corresponding to an area of 0.0050. The z value falls about halfway between −2.57 and −2.58 corresponding to the areas 0.0049 and 0.0051. The average of −2.57 and −2.58 is [(−2.57) + (−2.58)] / 2 = −2.575 so if the z value is needed to 3 decimal places, − 2.575 is used; however, if the z value is rounded to 2 decimal places, −2.58 is used. On the right side, it is necessary to find the z value corresponding to 0.99 + 0.005, or 0.9950. Again, the value falls between 2.57 and 2.58, so −2.575 or 2.58 can be used. The critical values are −2.58 and + 2.58 as shown in the following Fig. 7.10.

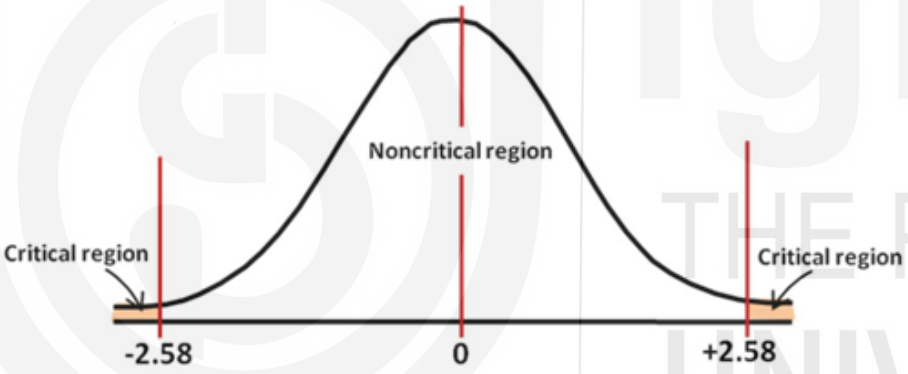


Fig. 7.10

p-value Method of Hypothesis Testing

Generally, we test hypotheses at the common α levels of 0.05 or 0.01 and sometimes at 0.10, but many computer statistical packages give a p-value for hypothesis tests. The p-value (or probability value) is the probability of getting a sample statistic (such as the mean) or a more extreme sample statistic in the direction of the alternative hypothesis when the null hypothesis is true. In other words, the p-value is the actual area under the standard normal distribution curve (or other curve, depending on what statistical test is being used) representing the probability of a particular sample statistic or a more extreme sample statistic occurring if the null hypothesis is true.

For example, suppose that an alternative hypothesis is H_1: μ> 50 and the mean of a sample is $\bar{X}$ = 52. If the computer printed a p-value of 0.0356 for a statistical test, then the probability of getting a sample mean of 52 or

greater is 0.0356 if the true population mean is 50 (for the given sample size and standard deviation). The relationship between the p-value and the α value can be explained in this manner. For p =0.0356, the null hypothesis would be rejected at α = 0.05 but not at α= 0.01. This is clarified in the following Fig. 7.11.

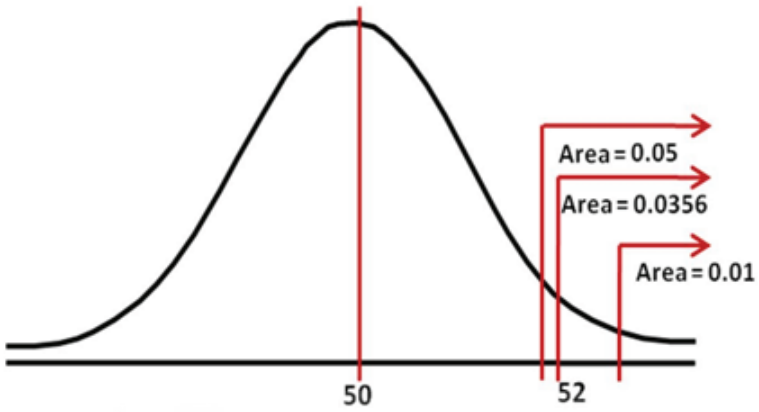


Fig. 7.11

When the hypothesis test is two-tailed, the area in one tail must be doubled. For a two-tailed test, if α is 0.05 and the area in one tail is 0.0356, the p-value will be 2(0.0356) = 0.0712. That is, the null hypothesis should not be rejected at α = 0.05, since 0.0712 is greater than 0.05. In summary, then, if the p-value is less than α, reject the null hypothesis. If the p-value is greater than α, do not reject the null hypothesis. The p-values for the z test can be found by using Standard Normal Distribution Table (z-table). First find the area under the standard normal distribution curve corresponding to the z test value. For a left-tailed test, use the area given in the table; for a right-tailed test, use 1.0000 minus the area given in the table. To get the p-value for a two-tailed test, double the area you found in the tail.

Confidence Intervals and Hypothesis Testing

There is a relationship between confidence intervals and hypothesis testing. When the null hypothesis is rejected in a hypothesis-testing situation, the confidence interval for the mean using the same level of significance will not contain the hypothesised mean. Likewise, when the null hypothesis is not rejected, the confidence interval computed using the same level of significance will contain the hypothesised mean. The concept and construction of confidence intervals is discussed under Section 7.4 of this Unit.

Type II Error and the Power of a Test

As stated above, there are two types of errors: type I and type II. A type I error can occur only when the null hypothesis is rejected, when it is true. By choosing a level of significance, say, 0.05 or 0.01, the researcher can determine the probability of committing a type I error. On the other hand, if the null hypothesis is not rejected, then either it is true or a type II error has been committed. A type II error occurs when the null hypothesis is indeed

false, but is not rejected. The probability of committing a type II error is denoted as β.

The value of β is not easy to compute. It depends on several things, including the value of the size of the sample, the population standard deviation, and the actual difference between the hypothesised value of the parameter being tested and the true parameter. The researcher has control over two of these factors, namely, the selection of α and the size of the sample. The standard deviation of the population is sometimes known or can be estimated. The major problem, then, lies in knowing the actual difference between the hypothesised parameter and the true parameter. If this difference were known, then the value of the parameter would be known; and if the parameter were known, then there would be no need to do any hypothesis testing. Hence, the value of β cannot be computed. But this does not mean that it should be ignored. What the researcher usually does is to try to minimise the size of β or to maximise the size of 1 − β, which is called the power of a test. However, the power of a test can be calculated for different values of true parameter and plotted to examine the power of the test. However, the power of a test can be calculated for different values of true parameter.

The power of a statistical test measures the sensitivity of the test to detect a real difference in parameters if one actually exists. The power of a test is a probability and, like all probabilities, can have values ranging from 0 to 1. The higher the power, the more sensitive the test is to detecting a real difference between parameters if there is a difference. As stated, the power of a test depends on the probability of committing a type II error, and since β is not easily computed, the power of a test cannot be easily computed. However, there are some guidelines that can be used when you are conducting a statistical study concerning the power of a test.

- When you are conducting a statistical study, use the test that has the highest power for the data.
- Another way to increase the power of a test is to select a larger sample size. A larger sample size would make the standard error of the mean smaller and consequently reduce β.

These two methods should be used with adequate caution: Before α can be increased, the researcher must consider the consequences of committing a type I error. If these consequences are more serious than the consequences of committing a type II error, then α should not be increased. Likewise, there are consequences to increasing the sample size. These consequences might include an increase in the amount of money required to do the study and an increase in the time needed to tabulate the data. When these consequences result, increasing the sample size may not be practical. The relationship among α, β, and the revelation that there is no magic formula or statistical test that can guarantee foolproof results when a decision is made about the validity of H_0. Whether the decision is to reject H_0 or not to reject H_0, there is, in either case, a chance of being wrong. The goal, then, is to try to keep the probabilities of type I and type II errors as small as possible.

Check Your Progress 2

1) Define the terms 'Parameter' and 'Statistic'?

.....

.....

.....

.....

2) Explain the characteristics of a normal distribution, and its usage in solving statistical problems?

.....

.....

.....

.....

3) What is z-variate and its applications in statistical problems?

.....

.....

.....

.....

4) What is the method of hypothesis testing?

.....

.....

.....

.....

5) What is type I and type II Error in Hypothesis Testing? Explain their relationship and the tradeoff between them?

.....

.....

.....

.....

6) What is the power of test and its relation with type II error?

.....

.....

.....

.....

7.4 INFERENTIAL STATISTICS: POINT ESTIMATION AND SETTING UP CONFIDENCE INTERVALS FOR POPULATION PARAMETERS

One aspect of inferential statistics is estimation, which is the process of estimating the value of a parameter from information obtained from a sample. *The statistical procedures for estimating the population mean, proportion, variance, and standard deviation are discussed in this section.*

7.4.1 Point Estimation for Population Mean

Suppose a college teacher wishes to estimate the average marks of students in Statistics attending his classes this semester. The teacher could select a random sample of 100 students and find the average marks of these students, say, 65 out of 100. From the sample mean, the teacher could infer that the average mark of all the students is 65. This type of estimate is called a point estimate. The Sample measures (i.e., statistics) are used to estimate population measures (i.e. parameters). These statistics are called estimators. For an estimator to be a good and reliable estimator, it needs to satisfy the following properties:

- *The estimator should be an unbiased estimator*. That is, the expected value or the mean of the estimates obtained from samples of a given size is equal to the parameter being estimated.
- *The estimator should be consistent*. For a consistent estimator, as sample size increases, the value of the estimator approaches the value of the parameter estimated.
- *The estimator should be a relatively efficient estimator*. That is, of all the statistics that can be used to estimate a parameter, the relatively efficient estimator has the smallest variance.

You might ask why other measures of central tendency, such as the median and mode, are not used to estimate the population mean. The reason is that the means of samples vary less than other statistics (such as medians and modes) when many samples are selected from the same population. Therefore, the sample mean is the best estimate of the population mean.

7.4.2 Confidence Intervals for the Mean When Population Standard Deviation (σ) is Known

You may ask the question, how good is a point estimate? The answer is that there is no way of knowing how close a particular point estimate is to the population mean. For this reason, statisticians prefer another type of estimate, called an interval estimate.

An interval estimate of a parameter is an interval or a range of values used to estimate the parameter. This estimate may or may not contain the value of the parameter being estimated.

In an interval estimate, the parameter is specified as being between two values. For example, an interval estimate for the average marks of all

students might be 60 <μ < 65, i.e. ± 5 marks. Either the interval contains the parameter or it does not. A degree of confidence (usually a per cent) can be assigned before an interval estimate is made. For instance, you may wish to be 95% confident that the interval contains the true population mean. Another question then arises. Why 95%? Why not 99 or 99.5%? If you desire to be more confident, such as 99 or 99.5% confident, then you must make the interval larger. To be more confident that the interval contains the true population mean, you must make the interval wider.

The confidence level of an interval estimate of a parameter is the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated. A confidence interval is a specific interval estimate of a parameter determined by using data obtained from a sample and by using the specific confidence level of the estimate. Intervals constructed in this way are called confidence intervals. Three common confidence intervals are used: the 90, the 95, and the 99% confidence intervals.

A brief intuitive explanation based on the central limit theorem states that when the sample size is large, approximately 95% of the sample means taken from a population and same sample size will fall within ±1.96 standard errors of the population mean, that is,

$$\mu \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Now, if a specific sample mean is selected, say, $\bar{X}$, there is a 95% probability that the interval $\mu \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ contains $\bar{X}$. Likewise, there is a 95% probability that the interval specified by

$$\bar{X} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

will contain μ, as will be shown later. Stated another way,

$$P\left\{\bar{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right\} = 0.95$$

Hence, you can be 95% confident that the population mean is contained within that interval when the values of the variable are normally distributed in the population. The value used for the 95% confidence interval is 1.96. For a 99% confidence interval, the value 2.58 is used instead of 1.96 in the formula. These values are obtained from a Table which is based on the standard normal distribution. In the general formula for confidence intervals, the symbol $Z_{\alpha/2}$ is used in statistics. The Greek letter α(alpha) represents the total area in both tails of the standard normal distribution curve, and α/2represents the area in each one of the tails. The relationship between α and the confidence level is that the stated confidence level is the percentage equivalent to the decimal value of 1– α, and vice versa. When the 95% confidence interval is to be found, α = 0.05, since 1 – 0.05 = 0.95, or

95%.When α = 0.01, then 1 − α = 1 − 0.01 = 0.99, and the 99% confidence interval is being calculated.

$$\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 90% confidence interval, $z_{\alpha/2}$ = 1.65; for a 95% confidence interval, $z_{\alpha/2}$= 1.96; and for a 99% confidence interval, $z_{\alpha/2}$= 2.58.

The term $z_{\alpha/2}(\sigma/\sqrt{n})$ is called the maximum error of the estimate (also called the margin of error). The maximum error of the estimate is the maximum likely difference between the point estimate of a parameter and the actual value of the parameter. For a specific value, say, α = 0.05, 95% of the samples means will fall within this error value on either side of the population mean as shown below in the Fig.7.12.

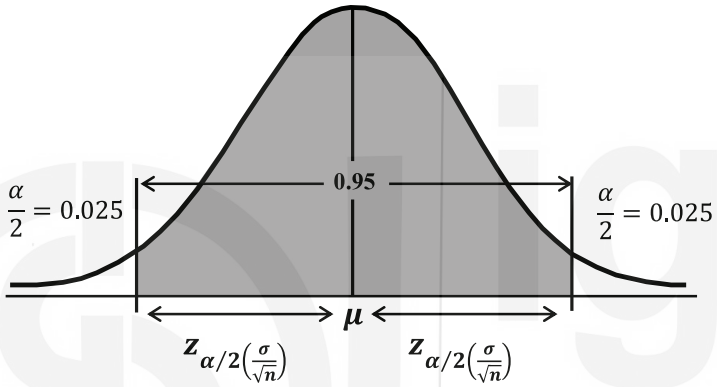


Fig. 7.12: Distribution of $\overline{X'}s$

Since other confidence intervals (besides 90, 95, and 99%) are sometimes used in statistics, an explanation of how to find the values for $Z_{\alpha/2}$is necessary. As stated previously, the Greek letter α represents the total of the areas in both tails of the normal distribution. The value for α is found by subtracting the decimal equivalent for the desired confidence level from 1. For example, if you wanted to find the 98% confidence interval, you would change 98% to 0.98 and find α = 1 − 0.98, or 0.02. Then α /2 is obtained by dividing α by 2. So α /2 is 0.02/2, or 0.01. Finally, $Z_{0.01}$ is the *z* value that will give an area of 0.01 in the right tail of the standard normal distribution curve. This is shown in the following Fig. 7.13.

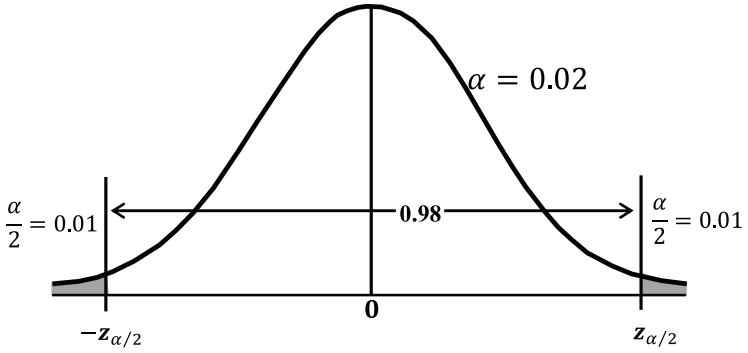


Fig. 7.13

When the original variable is normally distributed and σ is known, the standard normal distribution can be used to find confidence intervals regardless of the size of the sample.

7.4.3 Confidence Intervals for the Mean when Population Standard Deviation (σ) is Not Known

However, most of the time, the value σ is not known, so it must be estimated by using S, namely, the standard deviation of the sample. The S is used, especially when the sample size is small, critical values greater than the values for $Z_{\alpha/2}$ are used in confidence intervals in order to keep the interval at a given level, such as the 95%. These values are taken from the Student t distribution, most often called the t distribution. The t distribution differs from the standard normal distribution in the following ways:

- The variance is greater than 1.
- The t distribution is actually a family of curves based on the concept of degrees of freedom (see in the box below), which is related to sample size.
- As the sample size increases, the t distribution approaches the standard normal distribution.

Concept of Degree of Freedom

The degrees of freedom are the number of values that are free to vary after a sample statistic has been computed, and they tell the researcher which specific curve to use when a distribution consists of a family of curves with varying degrees of freedom. For example, if the mean of 5 values is 10, then 4 of the 5 values are free to vary. But once 4 values are selected, the fifth value must be a specific number to get a sum of 50, since 10×5 = 50. Hence, the degrees of freedom are 5 − 1 = 4, and this value tells us which t curve to use.

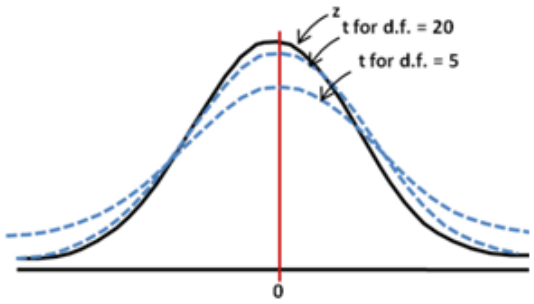


Fig. 7.14

The symbol d.f. will be used for degrees of freedom. The degrees of freedom for a confidence interval for the mean are found by subtracting 1 from the sample size. That is, d.f.= n − 1. As degree of freedom increases, t-distribution approaches the normal distribution.

To use this t-distribution, the samples must be simple random samples, and the population from which the samples were taken must be normally or approximately normally distributed, or the sample size must be 30 or more.

The formula for finding a confidence interval about the mean by using the t distribution is given below:

$$\bar{X} - t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

The degrees of freedom are n – 1.

The values for $t_{\alpha/2}$ are found from the table of t-distribution, the top row of the Table labeled Confidence Intervals, is used to get these values.

Example: Find the $t_{\alpha/2}$ value for a 95% confidence interval when the sample size is 22. The d.f. = 22 − 1, or 21. Find 21 in the left column and 95% in the row labeled Confidence Intervals. The intersection where the two meet, gives the value for $t_{\alpha/2}$ which is 2.080 as shown below:

The t-distribution Table

	Confidence Intervals	50%	80%	90%		95%	99%
d.f.	One tail α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails α	0.50	0.20	0.10	0.05	0.02	0.01
1							
2							
3							
⋮							
21						2.080	2.831
⋮		0.674	1.282[a]	1.645[b]	1.960	2.326[c]	2.576[d]
$(z)^{\infty}$							

7.4.4 Confidence Intervals for Proportions

Proportions can be obtained from samples or populations. Symbols used in proportion notation are:

P = population proportion

$\hat{p}$ (read "p hat") = sample proportion

For a sample proportion,

$\hat{p} = \frac{X}{n}$ and $\hat{q} = \frac{n - X}{n}$ or $\hat{q} = 1 - \hat{p}$

where X = number of sample units that possess the characteristics of interest and n = sample size.

For example, in a study, 200 people were asked if they were satisfied with their job or profession; 162 said that they were. In this case, n = 200, X = 162, and $\hat{p}$ = X/n = 162/200 = 0.81. It can be said that for this sample, 0.81, or

81%, of those surveyed were satisfied with their job or profession. The sample proportion is $\hat{p} = 0.81$.

When $\hat{p}$ and $\hat{q}$ are given in decimals or fractions, $\hat{p} + \hat{q} = 1$. When $\hat{p}$ and $\hat{q}$ are given in percentages, $\hat{p} + \hat{q} = 100\%$. It follows, then, that $\hat{q} = 1 - \hat{p}$, or $\hat{p} = 1 - \hat{q}$, when $\hat{p}$ and $\hat{q}$ are in decimal or fraction form. For the sample survey on job satisfaction, can also be found by using $\hat{q} = 1 - \hat{p}$, or $1 - 0.81 = 0.19$. Similar reasoning applies to population proportions; that is, $p = 1 - q$, $q = 1 - p$, and $p + q = 1$, when p and q are expressed in decimal or fraction form. When p and q are expressed as percentages, $p + q = 100\%$, $p - 100\% = q$, and $q = 100\% - p$.

As with means, point and interval estimates for a population proportion can be made by using the sample proportion. For a point estimate of p (the population proportion),$\hat{p}$(the sample proportion) is used. On the basis of the three properties, a good estimator is unbiased, consistent, and relatively efficient. But as with means, one is not able to decide how good the point estimate of p is. Therefore, statisticians also use an interval estimate for a proportion, and they can assign a probability that the interval will contain the population proportion. The confidence interval for a particular p is based on the sampling distribution of $\hat{p}$. When the sample size n is large then, the sampling distribution of $\hat{p}$ is approximately normal with a mean of p and a standard deviation of $\sqrt{\frac{pq}{n}}$, where q = 1 – p.

Confidence intervals about proportions must meet the criteria that $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$. Formula for a specific confidence interval for proportion is given below:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

When $n\hat{p}$ and $n\hat{q}$ are each greater than or equal to 5.

7.4.5 Confidence Intervals for Variances and Standard Deviations

In statistics, the variance and standard deviation of a variable are as important as the mean. For example, when products that fit together (such as pipes) are manufactured, it is important to keep the variations of the diameters of the products as small as possible; otherwise, they will not fit together properly and will have to be scrapped. In the manufacture of medicines, the variance and standard deviation of the medication in the pills play an important role in making sure patients receive the proper dosage. For these reasons, confidence intervals for variances and standard deviations are necessary.

To calculate these confidence intervals, a new statistical distribution is needed. It is called the chi-square distribution. The chi-square variable is similar to the t variable in that its distribution is a family of curves based on the number of degrees of freedom. The symbol for chi-square is χ^2 (Greek

letter chi, pronounced “ki”). Several of the curves of the distributions are shown below in along with the corresponding degrees of freedom.

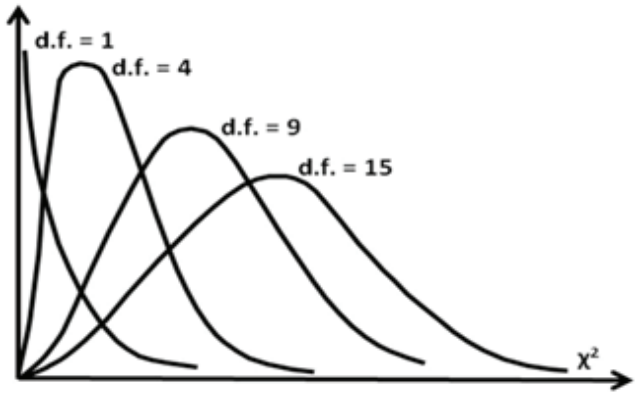


Fig. 7.15

The chi-square distribution is obtained from the values of $(n - 1) S^2/\sigma^2$ when random samples are selected from a normally distributed population whose variance is σ^2. A chi-square variable cannot be negative, and the distributions are skewed to the right. At about 100 degrees of freedom, the chi-square distribution becomes somewhat symmetric. The area under each chi-square distribution is equal to 1.00, or 100%. A standard table gives the values for the chi-square distribution corresponding to various degrees of freedom. These values are used in the denominators of the formulas for confidence intervals. Two different values are used in the formula because the distribution is not symmetric. One value is found on the left side of the table, and the other is on the right. For example, to find the table values corresponding to the 95% confidence interval, you must first change 95% to a decimal and subtract it from 1 (1 − 0.95 = 0.05). Then divide the answer by 2 (α/2 = 0.05/2 = 0.025). This is the column on the right side of the table, used to get the values for χ^2 right. To get the value for χ^2 left, subtract the value of α/2 from 1 (1−0.05/2 = 0.975). Fig.7.16 below illustrates the point:

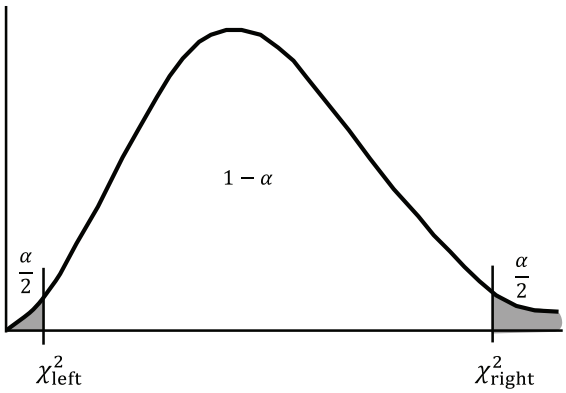


Fig. 7.16

Finally, find the appropriate row corresponding to the degrees of freedom n−1. A similar procedure is used to find the values for a 90 or 99% confidence interval.

Example: Find the values for χ^2_{right} and χ^2_{left} for a 90% confidence interval when n =25.

To find χ^2_{right}, subtract 1– 0.90 =0.10 and divide by 2 to get 0.05.To find χ^2_{left} subtract 1–0.05 to get 0.95. Hence, use the 0.95 and 0.05 columns and the row corresponding to 24 d.f. as shown below:

The Chi-Square Distribution

Degrees of Freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1										
2										
⋮										
24				13.848			36.415			

χ^2_{left} χ^2_{right}

The answers are

$$\chi^2_{right} = 36.415$$

$$\chi^2_{left} = 13.848$$

The estimates for σ^2 and σ are S^2 and S, respectively. To find confidence intervals for variances and standard deviations, it is assumed that the variable is normally distributed. The formulas for the confidence intervals are shown here.

Formula for the Confidence Interval for a Standard Deviation

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\text{right}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{left}}}$$

Formula for the Confidence Interval for a Variance

$$\frac{(n-1)s^2}{\chi^2_{right}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{left}}$$

d.f. = n ─ 1

It should be clear by now that an important aspect of inferential statistics is estimation. There are two types of estimates of a parameter: point estimates and interval estimates, which are adequately discussed under this section.

Check Your Progress 3

1) What is the difference between point and interval estimations? What are the properties of a good estimator?

……

……

……

2) Explain the z distribution and define the confidence intervals for mean and proportions under the assumptions of this distribution.

……

……

……

3) When do you use t distribution? Work out the confidence intervals for mean and proportions using t distribution.

……

……

……

……

4) What is a chi square distribution? How do you use it to work out confidence interval for standard deviation?

……

……

……

……

7.5 LET US SUM UP

This Unit explains how the boxplot and five-number summaries, (which are parts of exploratory data analysis) are used to examine data to see what they reveal. The significance of the central tendency measures and dispersions is explained in case of normal frequency distribution which is widely used in solving the real practical problems.

An important aspect of data analysis is estimation. There are two types of estimates of a parameter: point estimates and interval estimates. A point estimate is a specific value. Estimations of parameters of populations are accomplished by selecting a random sample from that population and choosing and computing a statistic that is the best estimator of the parameter. A good estimator must be unbiased, consistent, and relatively efficient. The problem with point estimates is that the accuracy of the estimate cannot be determined. For this reason, statisticians prefer to use the interval estimate. By computing an interval about the sample value, statisticians can be 95 or 99% (or some other percentage) confident that their estimate contains the true parameter. The confidence level is determined by the researcher. The higher the confidence level, the wider the interval of the estimate must be. Various probability distributions are used to set the confidence intervals for mean, proportions and standard

deviation and variance depending upon the availability of population standard deviation, size of the sample, degree of freedom available with researcher.

7.6 REFERENCES

1) Blaikie, N. (2003). *Analyzing quantitative data: From description to explanation.* Sage.
2) Bohrnstedt, G. W., &Knoke, D. (1994). *Statistics for social data analysis.*
3) Bryman, A., & Cramer, D. (1994). *Quantitative data analysis for social scientists (rev. Taylor & Frances/Routledge.*
4) Cramer, D. (2003). *Advanced quantitative data analysis.* McGraw-Hill International.
5) Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative data.* Prentice Hall.
6) Suen, H. K., &Ary, D. (2014). *Analyzing quantitative behavioral observation data.* Psychology Press.

7.7 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Sub-Section 7.2.1
2) See Sub-Section 7.2.2

Check Your Progress 2

1) See Sub-Section 7.3.1
2) See Sub-Section 7.3.3 and 7.3.4
3) See Sub-Section 7.3.5
4) See Sub-Section 7.3.6
5) See Sub-Section 7.3.6
6) See Sub-Section 7.3.6 under the head “Type II error and the power of a test”

Check Your Progress 3

1) See Sub-Section 7.4.1 and 7.4.2
2) See Sub-Section 7.4.2
3) See sub-section 7.4.3 and 7.4.4
4) See sub section 7.4.5

UNIT 8 BIVARIATE DATA ANALYSIS

Structure

8.1 Introduction

8.2 Scatter Plots and Correlation

8.3 Analysing Extent of Relationship in Bivariate Data Analysis – Correlation

- 8.3.1 Concept of Correlation
- 8.3.2 Correlation Coefficient
- 8.3.3 Test of Significance for the Correlation Coefficient
- 8.3.4 Correlation and Causation

8.4 Analysing Cause and Effect Relationship in Bivariate Data – Regression

- 8.4.1 Line of Best Fit
- 8.4.2 Regression Lines Equation
- 8.4.3 Regression Coefficients
- 8.4.4 Predictability of Regression Equations
- 8.4.5 Coefficient of Determination

8.5 Reliability of Regression Model – Standard Error of Estimate

- 8.5.1 Standard Error of Estimate: Concept and Estimation
- 8.5.2 Prediction Interval

8.6 Testing the Difference between two Means, two Proportions, and two Variances

- 8.6.1 Testing the Difference between Two Means: Using the z-test and t-test
- 8.6.2 Testing the Difference between Proportions Using z-test
- 8.6.3 Testing the Difference between Two Variances: F-Test

8.7 Analysis of Variances

8.8 Let Us Sum Up

8.9 References

8.10 Answers or Hints to Check Your Progress Exercises

8.0 OBJECTIVES

After going through this unit on bivariate analysis, you should be able to:

a) answer the following questions:

- are two variables related?
- if so, what is the strength of the relationship between the two variables?
- what type of relationship exists between them?

- what kind of predictions can be made about the relationship of these two variables?

b) explain

- the difference between correlation and causation;
- the cause and effect relationship in Bivariate Data using Regression Analysis; and
- predictability of cause and effect relation in population given the sample drawn from the population under some assumptions.

8.1 INTRODUCTION

In Unit 2 and 7, we discussed exploratory data analysis as well as the methods of estimation of univariate population parameters based on the corresponding statistics calculated from the samples drawn from the parent population. In this Unit, we will discuss the bivariate data description, *i.e.* descriptive methods used to explore the joint distribution of the pairs of values of two variables. The statistical analysis related to the study of the relationship between two variables is known as Bi-Variate Analysis. The joint distribution of a pair of variables is the way in which the pairs of possible values of these variables are distributed among the units in the group of interest. When we measure or observe pairs of values for a pair of variables, we want to know how the two variables behave together (the joint distribution of the two variables). Another area of inferential statistics involves determining whether a relationship exists between two or more numerical or quantitative variables. For example, a business person may want to know whether the volume of sales for a given month is related to the amount of advertising the firm does that month. Educators are interested in determining whether the number of hours a student studies is related to the student's score on a particular exam. Medical researchers are interested in questions such as, is caffeine related to heart damage? Or is there a relationship between a person's age and his or her blood pressure? These are only a few of the many questions that can be answered by using the techniques for bivariate data analysis.

To answer the first question listed under Section 8.0, statisticians use a numerical measure to determine whether two or more variables are related and to determine the strength of the relationship between or among the variables. This measure is called a correlation coefficient. For example, there are many variables that contribute to heart disease, among them lack of exercise, smoking, heredity, age, stress, and diet. Of these variables, some are more important than others; therefore, a physician who wants to help a patient must know which factors are most important.

To answer the second question listed under Section 8.0, in a simple relationship, there are two variables—an independent variable, also called

an explanatory variable or a predictor variable, and a dependent variable, also called a response variable. A simple relationship analysis is called simple regression, and there is one independent variable that is used to predict the dependent variable. For example, a manager may wish to see whether the number of years the sales people have been working for the company has anything to do with the amount of sales they make. This type of study involves a simple relationship, since there are only two variables—years of experience and amount of sales. Simple relationships can also be positive or negative. A positive relationship exists when both variables increase and decrease at the same time. In a negative relationship, as one variable increases, the other variable decreases, and vice versa.

Finally, under Section 8.0, the fourth question asks what type of predictions can be made. Predictions are made in all areas and daily. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price predictions, and sports predictions. Some predictions are more accurate than others, due to the strength of the relationship. That is, the stronger the relationship is between the variables, the more accurate the prediction is.

8.2 SCATTER PLOTS AND CORRELATION

In simple bivariate studies, the researcher collects data on two numerical or quantitative variables on the same subjects to see whether a relationship exists between the variables. For example, if a researcher wishes to see the relationship between the following variables and collects the data from the relevant sizes of their samples:

i) Number of car rentals and the revenue of the car rental companies

ii) Number of absences and grade in the exams

iii) Number of hours of exercises in each week and the amount of milk consumed by the people

As mentioned earlier, after identifying the dependent and independent variables under these studies, these variables can be plotted on a graph called a scatter plot. A scatter plot is a graph of the ordered pairs (x, y) of numbers consisting of two variables x and y. The variable x is plotted on the horizontal axis, and the variable y is plotted on the vertical axis. After labelling the axis and plotting the data, the graph emerged from the three sample studies are drawn below:

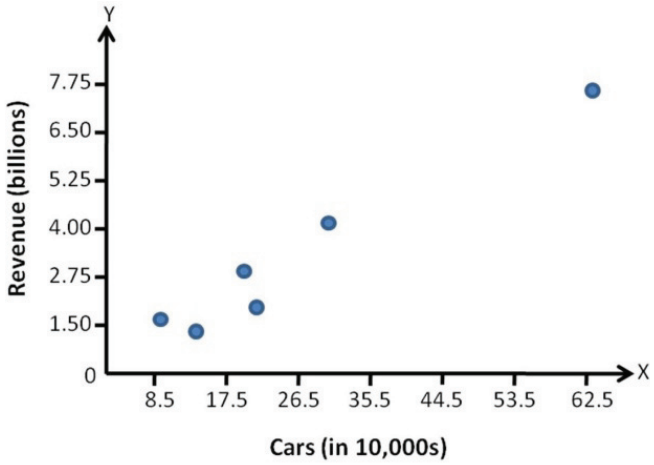


Fig. 8.1: Scatter Plot 1

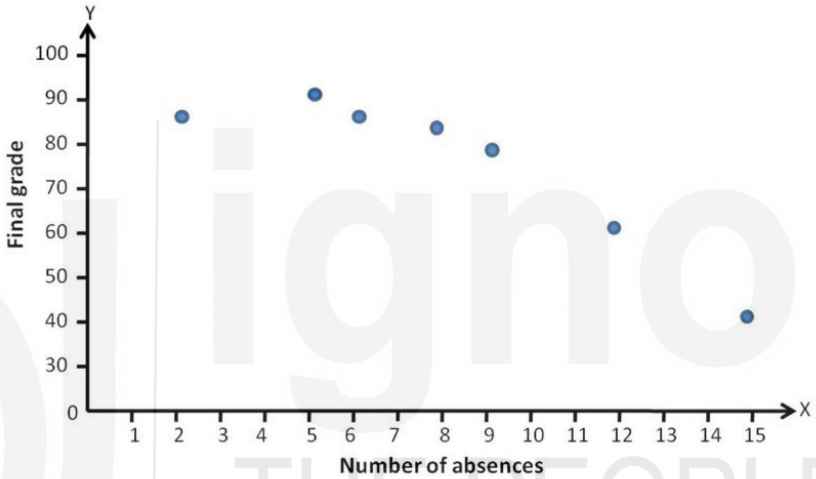


Fig. 8.2: Scatter Plot 2

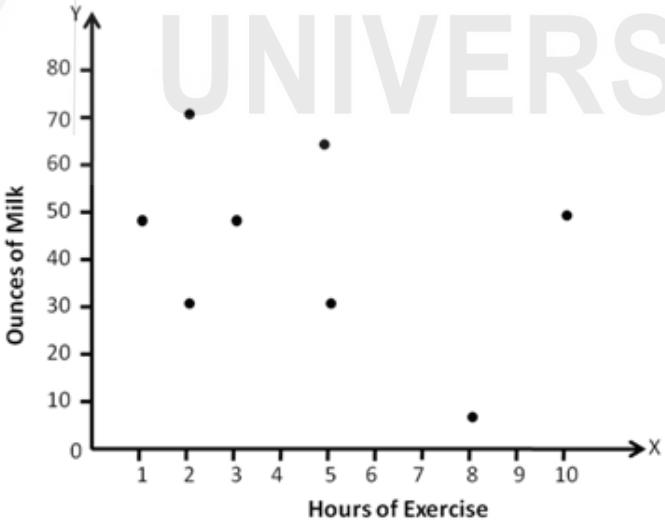


Fig. 8.3: Scatter Plot 3

After the plot is drawn, it should be analysed to determine which type of relationship, if any, exists. For example, the plot shown in scatter plot 1 suggests a positive relationship, since as the number of cars rented increases, revenue tends to increase also. The plot of the data shown in

scatter plot 2 suggests a negative relationship, since as the number of absences increases, the final grade decreases. Finally, the plot of the data shown in scatter plot 3 shows no specific type of relationship, since no pattern is discernible. Note that the data shown in plots 1 and 2 also suggest a linear relationship, since the points seem to fit a straight line, although not perfectly.

Check Your Progress 1

1) What is the scatter plot and what does it indicate?

……

……

……

……

2) What do the three scatter plots drawn under Section 8.2 depict?

……

……

……

……

8.3 ANALYSING EXTENT OF RELATIONSHIP IN BIVARIATE DATA ANALYSIS – CORRELATION

8.3.1 Concept of Correlation

In bivariate analysis, the variables are normally observed to be inter-related. For example, in health sciences we study the relationship between blood pressure and age, consumption level of some nutrients and weight gain, medicine consumption and the associated effects, etc. The strength of relationship may be examined by correlation analysis.

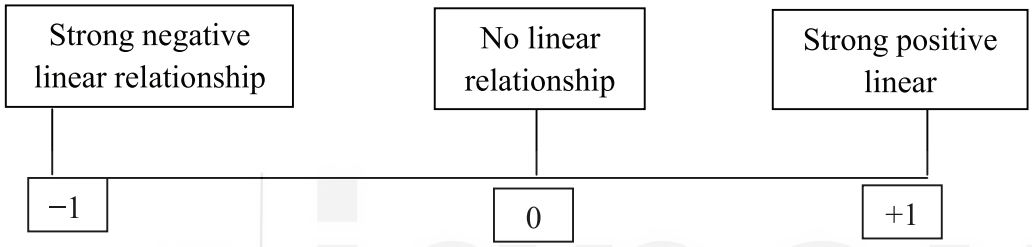
Correlation is statistical analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated. The analysis of correlation is widely used in physical and social sciences, e.g. it is useful for economists to study the relationship between variables like price, quantity and costs-sales-price, income and expenditure, etc. using correlation. Most importantly, it forms the basis for the concept of regression.

8.3.2 Correlation Coefficient

There are several types of correlation coefficients. The one explained in this section is called the **Pearson product moment correlation coefficient**

(PPMC), named after statistician Karl Pearson, who pioneered the research in this area.

*The **correlation coefficient** computed from the sample data measures the strength and direction of a linear relationship between two variables. The symbol for the sample correlation coefficient is r. The symbol for the population correlation coefficient is ρ (Greek letter rho).* The range of the correlation coefficient is from − 1 to + 1. If there is a strong positive linear relationship between the variables, the value of *r* will be close to +1. If there is a strong negative linear relationship between the variables, the value of *r* will be close to − 1. When there is no linear relationship between the variables (or only a weak relationship), the value of *r* will be 0 (or close to 0).



The following graphs show the relationship between the correlation coefficients and their corresponding scatter plots. Notice that as the value of the correlation coefficient increases from 0 to +1 (parts a, b, and c), data values become closer to an increasingly stronger relationship. As the value of the correlation coefficient decreases from 0 to – 1 (parts d, e, and f), the data values also become closer to a straight line. Again this suggests a stronger relationship.

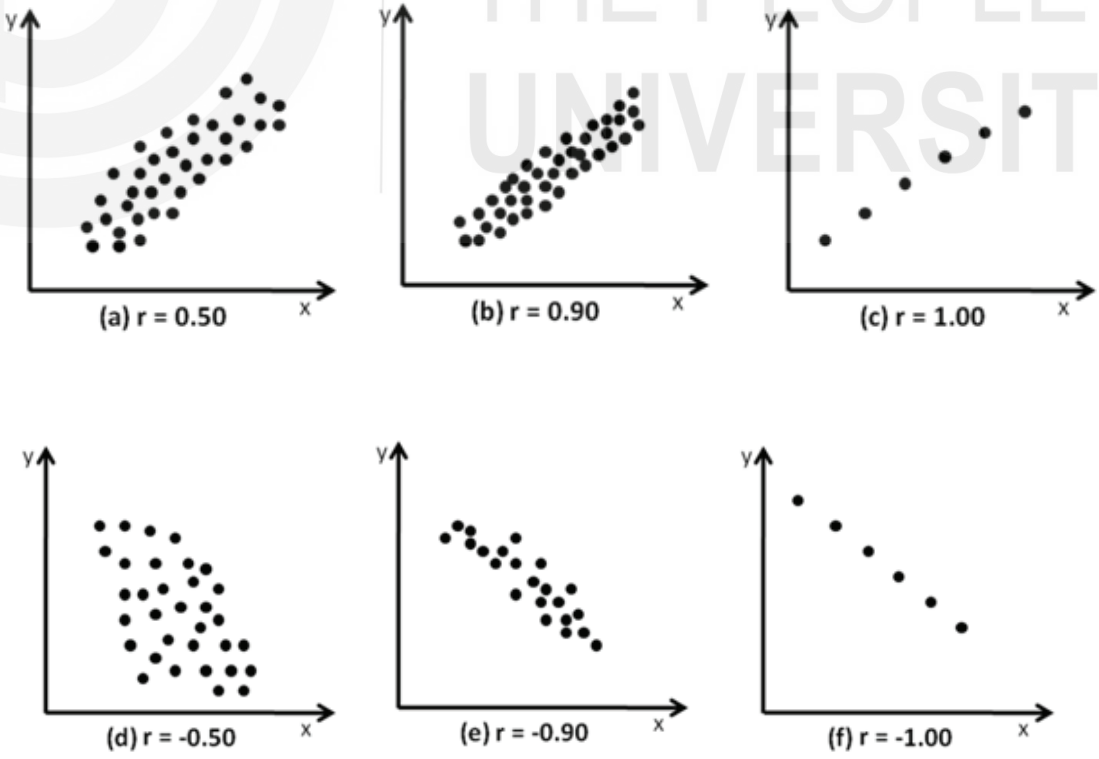


Fig. 8.4

Computation of Correlation Coefficient

The mutually exclusive formulae for computation of Karl Pearson's coefficient of correlation widely used in statistics are given below:

i) $r = \frac{Cov(X,Y)}{\sigma_X.\sigma_Y}$

ii) $r = \frac{\sum xy}{n\,\sigma_X.\sigma_Y}$

iii) $r = \frac{\sum xy}{\sqrt{\sum x^2.\sum y^2}}$

where, *n* is the total number of observations; $x = X - \bar{X}$, $y = Y - \bar{Y}$; σ_X and σ_Y are standard deviation of *X* and *Y* given by $\sigma_X = \sqrt{\frac{\sum(X-\bar{X})^2}{n}}$ or $\sqrt{\sum X^2 - \frac{(\sum X)^2}{n}}$ and $\sigma_Y = \sqrt{\frac{\sum(Y-\bar{Y})^2}{n}}$ or $\sqrt{\sum Y^2 - \frac{(\sum Y)^2}{n}}$ respectively; *Cov* (*X*,*Y*) is the covariance between X and Y, given by $Cov(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n}$ or $\sum XY - \frac{\sum X \sum Y}{n}$.

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of *X* and *Y* series, respectively. The steps involved in this method are:

1) Find the mean of the two series *X* and *Y*.
2) Take derivations of the two series from their respective means $\bar{X}$ and $\bar{Y}$, that is, $x = X - \bar{X}$, $y = Y - \bar{Y}$
3) Square the deviations and get the total of the respective square of deviations of X and Y and denote by $\sum x^2$, $\sum y^2$, respectively.
4) Multiply the deviations of *X* and *Y* and get the total (i.e. $\sum xy$).
5) Substitute the values in the formula

$$r = \frac{\sum xy}{\sqrt{\sum x^2.\sum y^2}}$$

Alternatively, it can also be written as;

$$r = \frac{n\,(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$$

The properties of the above correlation coefficient are:

- Property 1: Correlation coefficient lies between –1 and +1, i.e. $-1 \leq r \leq +1$
- Property 2: 'r' is independent of change of origin and scale.
- Property 3: It is a pure number independent of units of measurement.
- Property 4: Independent variables are uncorrelated.
- Property 5: Correlation coefficient is the geometric mean of two regression coefficients.

- Property 6: The correlation coefficient of *X* and *Y* is symmetric i.e., $r_{XY} = r_{YX}$.

But this coefficient has some limitations too, for instance,

- Correlation coefficient assumes linear relationship regardless of whether the assumption is correct or not.
- Extreme values of variables have unduly high influence on correlation coefficient.
- Existence of correlation does not necessarily indicates cause-effect relation.

8.3.3 Test of Significance for the Correlation Coefficient

As stated above, the range of the correlation coefficient is between − 1 and + 1. When the value of *r* is near +1 or − 1, there is a strong linear relationship. When the value of *r* is near 0, the linear relationship is weak or nonexistent. Since the value of *r* is computed from data obtained from samples, there are two possibilities when *r* is not equal to zero: either the value of *r* is high enough to conclude that there is a significant linear relationship between the variables, or the value of *r* is due to chance.

To make this decision, we use a hypothesis-testing procedure, the steps of which are listed below:

- Step 1: State the hypotheses.
- Step 2: Find the critical values.
- Step 3: Compute the test value.
- Step 4: Make the decision.
- Step 5: Summarise the results.

The population correlation coefficient is computed from taking all possible (X, Y) pairs; it is designated by the Greek letter rho (ρ). Formally defined, the population correlation coefficient ρ is the correlation computed by using all possible pairs of data values (X, Y) taken from a population. The sample correlation coefficient (*r*) can then be used as an estimator of ρ if the following assumptions are valid:

- The variables X and Y are linearly related.
- The variables are random variables.
- The two variables have a bivariate normal distribution.

A bivariate normal distribution means that for the pairs of (X, Y) data values, the corresponding Y values have a bell-shaped distribution for any given X value, and the X values for any given Y value have a bell-shaped distribution.

In hypothesis testing, one of these is true:

H_0: ρ = 0 This null hypothesis means that there is no correlation between the X and Y variables in the population.

$H_1: \rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.

When the null hypothesis is rejected at a specific significance level α, it means that there is a significant difference between the value of ρ and 0. When the null hypothesis is not rejected, it means that the value of ρ is not significantly different from 0 (zero) and whatever difference appears is probably due to chance.

There are several methods to test the significance of the correlation coefficient but here we will discuss only the t-test. Formula for the t-test for the correlation coefficient is given below:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$with\ degrees\ of\ freedom\ equal\ to\ n-2.$$

The two-tailed critical values are used. These values are found in t-distribution table. Also, when you are testing the significance of a correlation coefficient, both variables X and Y must come from normally distributed populations.

Example: Compute the correlation coefficient from the following sample data and test its significance.

Company	Cars (in ten thousands)	Revenue (in $ billions)
A	63.0	7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

Solution:

Computation of *r*

Company	Cars X (in 10,000s)	Income Y (in $ billions)	XY	X^2	Y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84

E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	2.75	72.25	2.25
	$\sum X = 153.8$	$\sum Y = 18.7$	$\sum XY = 682.77$	$\sum X^2 = 5859.26$	$\sum Y^2 = 80.67$

Substitute in the formula and solve for r.

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$$

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The correlation coefficient suggests a strong relationship between the number of cars a rental agency has and its annual income.

Significance of the Correlation Coefficient

Step 1: State the hypotheses

H_0: ρ = 0 and H_1: ρ ≠ 0

Step 2: Find the critical values and mark the acceptance and the rejection region. Since $\alpha = 0.05$ and there are 6 – 2 = 4 degree of freedom, the critical values obtained from t-distribution table are ± 2.776

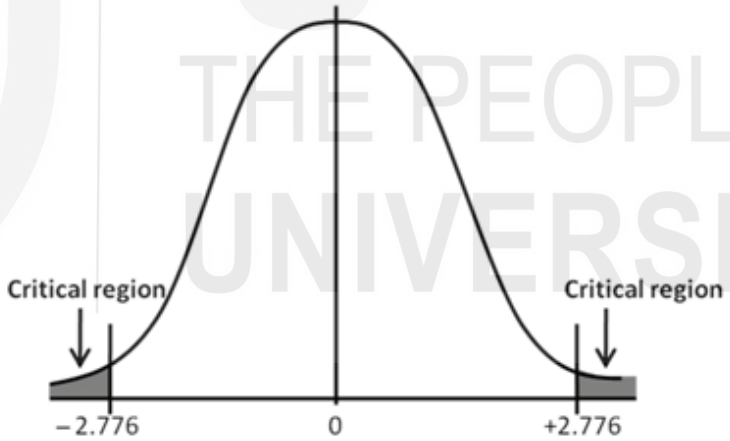


Fig.8.5

Step 3: Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.982\sqrt{\frac{6-2}{1-(0.982)^2}} = 10.4$$

Step 4: Make the decision. Reject the null hypothesis (that is, ρ = 0) as the computed test values falls **i**n the rejection region as shown below:

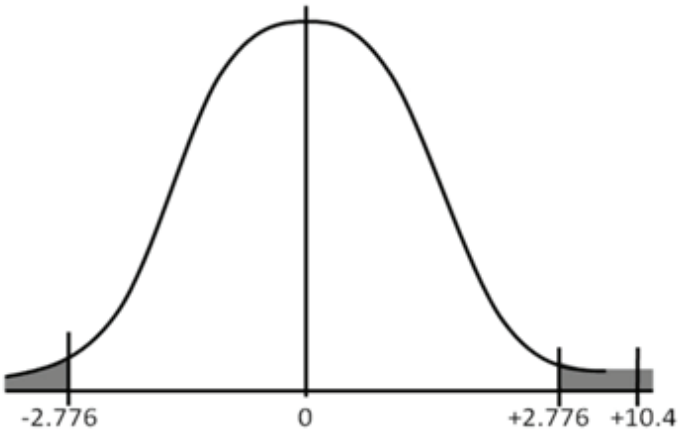


Fig. 8.6

Step 5: Summarise the results. There is a significant relationship between the number of cars a rental agency owns and its annual income.

8.3.4 Correlation and Causation

The nature of the linear relationship between the independent variable X and the dependent variable Y must be understood clearly. When a hypothesis test indicates that a significant linear relationship exists between the variables, researchers must consider all the possibilities of relationship. When the null hypothesis has been rejected for a specific α value, any of the following five possibilities can exist.

1) There is a direct cause-and-effect relationship between the variables. That is, X causes Y. For example, water causes plants to grow, poison causes death, and heat causes ice to melt.

2) There is a reverse cause-and-effect relationship between the variables. That is, Y causes X. For example, suppose a researcher believes excessive coffee consumption causes nervousness, but the researcher fails to consider that the reverse situation may occur. That is, it may be that an extremely nervous person craves coffee to calm his or her nerves.

3) The relationship between the variables may be caused by a third variable. For example, if a statistician correlated the number of deaths due to drowning and the number of cans of soft drink consumed daily during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.

4) There may be a complexity of interrelationships among many variables. For example, a researcher may find a significant relationship between students' high school grades and college grades. But there probably are many other variables involved, such as IQ, hours of study, influence of parents, motivation, age, and instructors.

5) The relationship may be coincidental. For example, a researcher may be able to find a significant relationship between the increase in the number of people who are exercising and the increase in the number

of people who are committing crimes. But common sense dictates that any relationship between these two values must be due to coincidence.

When two variables are highly correlated, item 3 above states that there exists a possibility that the correlation is due to a third variable. If this is the case and the third variable is unknown to the researcher or not accounted for in the study, it is called a lurking variable. An attempt should be made by the researcher to identify such variables and to use methods to control their influence. It is important to restate the fact that even if the correlation between two variables is high, it does not necessarily mean causation. There are other possibilities, such as lurking variables or just a coincidental relationship. Also, you should be cautious when the data for one or both of the variables involve averages rather than individual data. It is not wrong to use averages, but the results cannot be generalised to individuals since averaging tends to smooth out the variability among individual data values. The result could be a higher correlation than actually exists. Thus, when the null hypothesis is rejected, the researcher must consider all possibilities and select the appropriate one as determined by the study. Remember, **correlation does not necessarily imply causation.**

Activity 1: Creating a Scatter plot

Create the Scatter plot for the data of the example in sub-section 8.3.3.

How to do it?

- Insert the data in the spreadsheet (refer to Fig. 8.7)
- To create Scatter plot select the ***Insert*** tab and from the Ribbon Select ***Scatter*** from the *Chart* group and from the list of Scatter select ***Scatter with only Markers.***
- A Blank chart will be displayed, in addition in the tabs you will see Chart Tools - Design, Layout and Format.

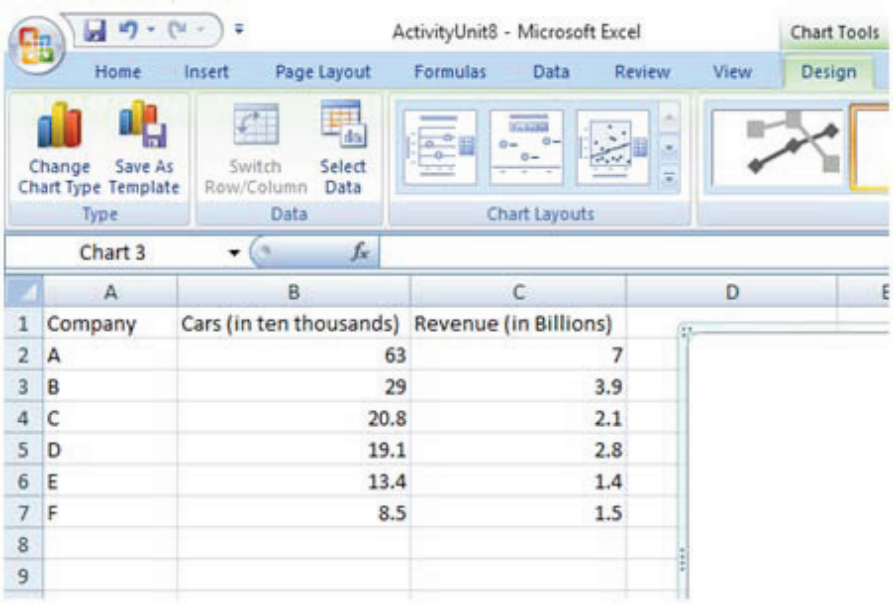


Fig. 8.7: Inserting of Blank Scatter plot

- In the *Design* tab, select *Select Data* option in the *Data* Group. You will see a *Select Data Source* dialog box with Chart Data Range text box highlighted. Take your mouse pointer to the cell having first value of table (in our example cell B2 having value 63, press and hold left mouse button on this cell and while holding left mouse button move the mouse pointer to last value in the table that is cell C7 having value 1.5 in this example. The Chart data range now will display *=Sheet1!$B$2:$C$7* (refer to Fig. 8.8). Please note that the cells B2 to C7 are shown in a dotted box as they have been selected.

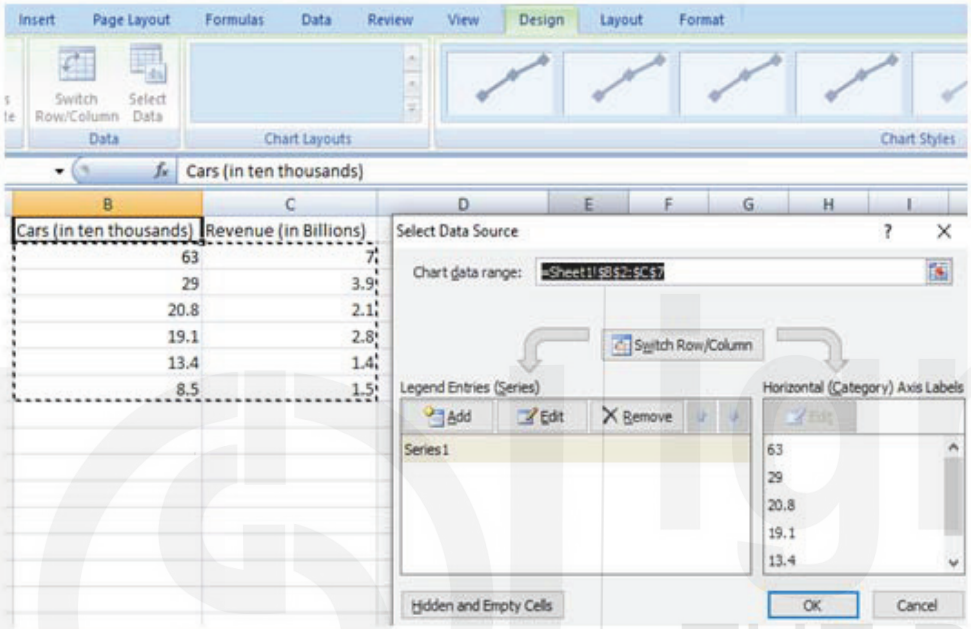
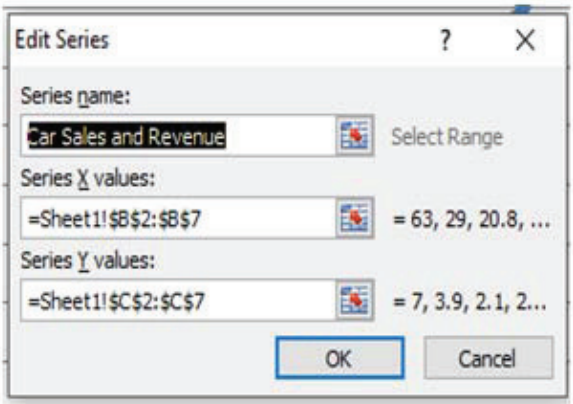


Fig. 8.8: Selection of Data Range B2 to C7

- Please notice it shows *Series1* in *Legend Entries (Series)* Area. To change this press *Edit* Button, you will see *Edit Series* dialog box in this box type "*Car Sales and Revenue*" in the *Series name* text box. Please check the X values and Y values and press *Ok* button. You will be back to Fig. 8.8 except that Series name will change from *Series1* to *Car Sales and Revenue.* Press *OK* button in *Select Data Source* dialog box.



- You will now see a scatter plot with Title and Legend Entry, and axis values. (refer to Fig. 8.9) To add axis titles, select *Layout* tab and then

Select *Axis Title* button in the Ribbon followed by *Primary Horizontal Axis Title* and *Title Below Axis* as shown in Fig. 8.9. Similarly, add vertical axis title using *Primary Vertical Axis Title* and selecting *Rotated Titles.* Please note this will just create an area for axis titles. Actual titles are to be added by you.

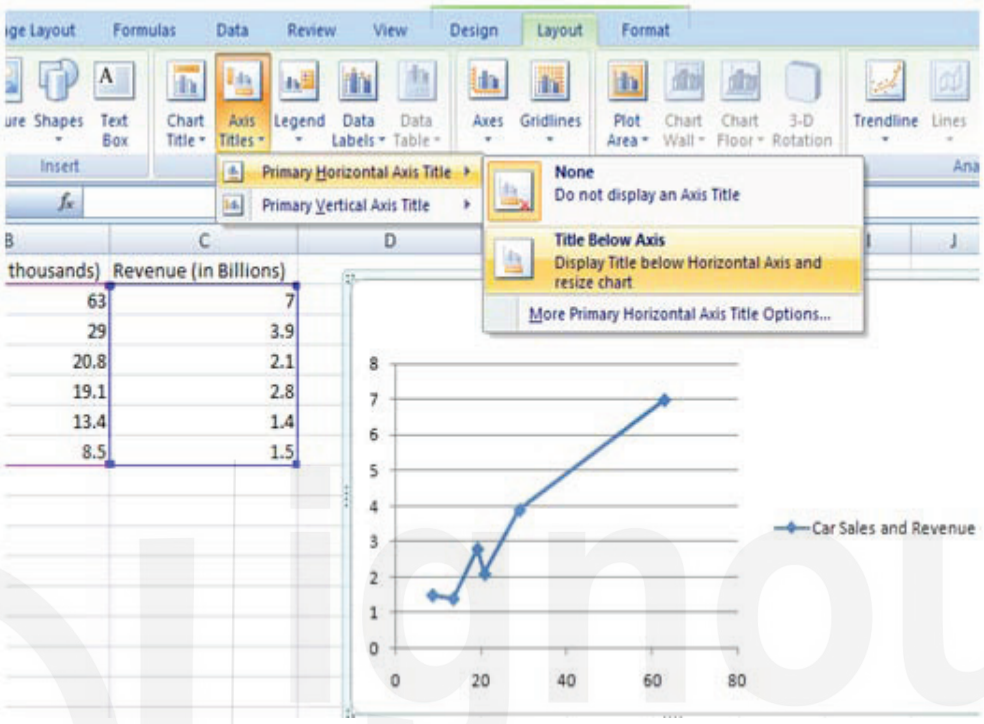
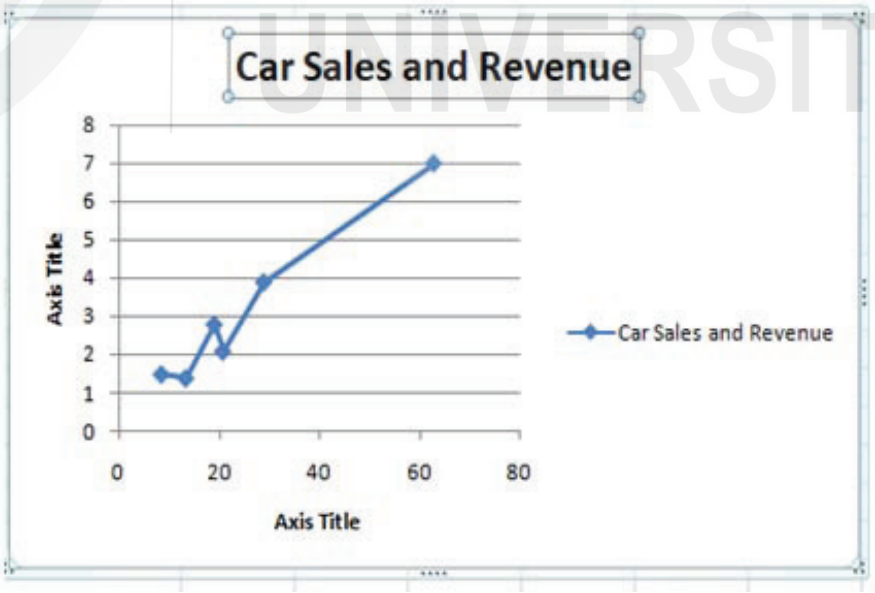


Fig. 8.9: Process of adding Axis Titles

- After selecting both the Titles, your scatter plot will change to the following:



- You need to put actual titles and remove the legend as it is only one series. Select the legend and delete it. Select the Axis Title and in the formula cell type =B1, if cell B1 contains the Axis title (refer to Fig. 8.10) which shows the final chart.

- Please note that in the Fig. 8.10, in the chart Vertical Axis is selected and the related value shown in the formula bar is *=Sheet1!$C$1* this is the address of the cell (C1) that contains the Title of the Vertical Axis. You can try changing the cell C1 in your sheet and you will find the change in the Title in the graph.

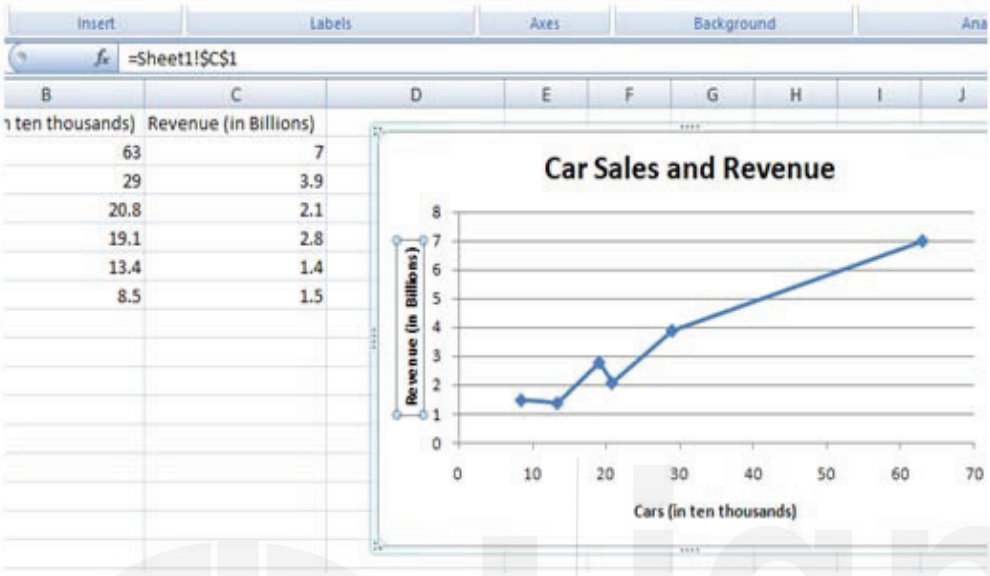


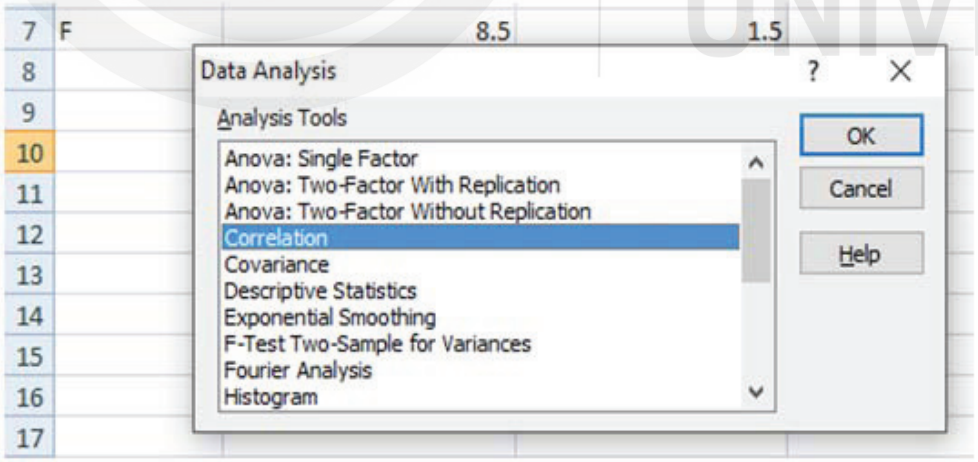
Fig. 8.10: Final Chart

Activity 2: Find the Correlation coefficient for the Data of Activity 1

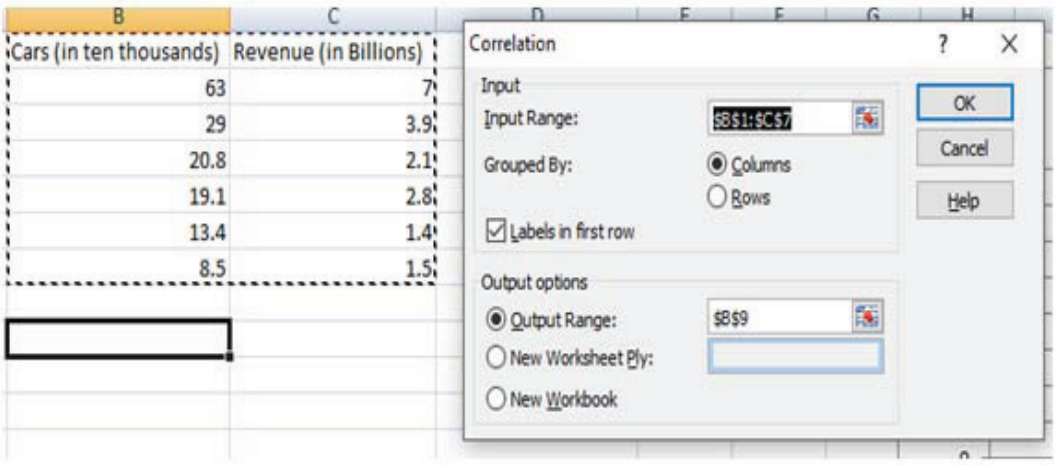
Find the correlation coefficient for the data shown in Fig. 8.10.

How to do it?

- From the *Data* Tab Select *Analysis* icon (right most) option on the ribbon. It will open a *Data Analysis* dialog box. Select *Correlation* from this dialog box as shown below and Press *OK* button.



- In the *Correlation* dialog box select Input Range as *$B$1:$C$7* and check the check box Labels in first row, which is the case for our data as first row contains headings. Keep the Grouped By option to Columns as columns contain data of one variable. Select output range as *$B$9* (following figure shows this dialog box) and press the OK button.



- The correlation coefficient matrix will be displayed as a table (Refer to Fig. 8.11)

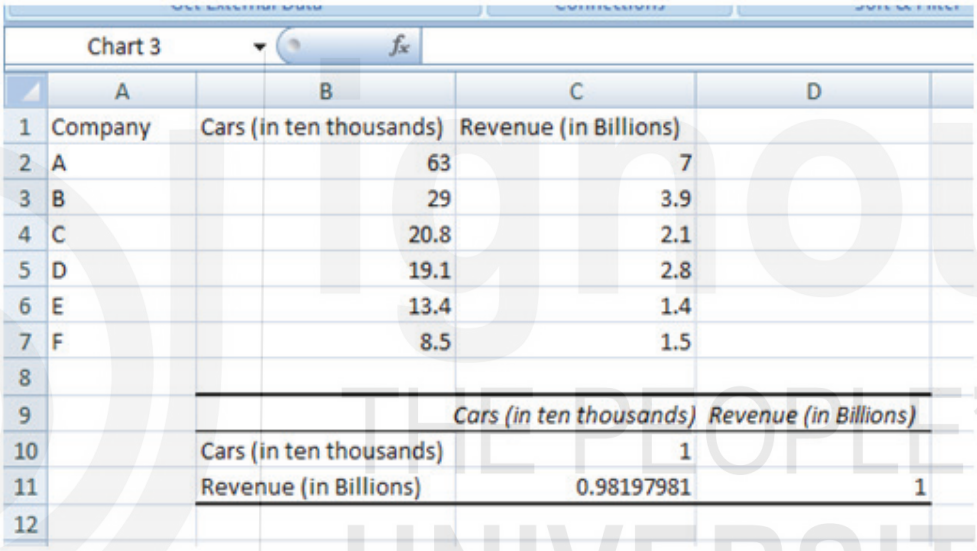


Chart 3

	A	B	C	D
1	Company	Cars (in ten thousands)	Revenue (in Billions)	
2	A	63	7	
3	B	29	3.9	
4	C	20.8	2.1	
5	D	19.1	2.8	
6	E	13.4	1.4	
7	F	8.5	1.5	
8				
9			*Cars (in ten thousands)*	*Revenue (in Billions)*
10		Cars (in ten thousands)	1	
11		Revenue (in Billions)	0.98197981	1
12				

Fig. 8.11: Correlation coefficients

Please observe that correlation between Revenue and Cars is 0.98197981. Compare it with the value calculated in sub-section 8.3.3.

Check Your Progress 2

1) Explain the concept of correlation in bivariate data analysis?

.....

.....

.....

2) Define the term 'correlation coefficient' and its relationship with the co-variation between the two variables?

.....

.....

.....

3) Discuss the computation procedure for correlation coefficient and also highlight the properties and limitations of correlation coefficient?

……

……

……

……

4) Using t-test, explain the procedure for testing the significance of correlation coefficient?

……

……

……

……

8.4 ANALYSING CAUSE AND EFFECT RELATIONSHIP IN BIVARIATE DATA – REGRESSION

In studying relationships between two variables, as discussed in sub-section 8.3.2, we first construct a scatter plot of the collected data of the two variables. The purpose of the scatter plot is to determine the nature of the relationship. The possibilities include a positive linear relationship, a negative linear relationship, or no discernible relationship. After the scatter plot is drawn, the next steps are to compute the value of the correlation coefficient and to test the significance of the relationship. If the value of the correlation coefficient is significant, the next step is to determine the value of one variable given the value of another. The variable predicted on the basis of other variables is called the "dependent" or the 'explained' variable and the other the 'independent' or the 'predicting' variable. For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of advertising expenditure for attaining a given amount of sales. Regression is the measure of the average relationship between two or more variables in terms of the original units of the data. It is important to note that determining the regression line when *r* is not significant and then making predictions using the regression line are meaningless. The purpose of the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

8.4.1 Line of Best fit

The following figure shows a scatter plot for the data of two variables. It shows that several line scan be drawn on the graph near the points.

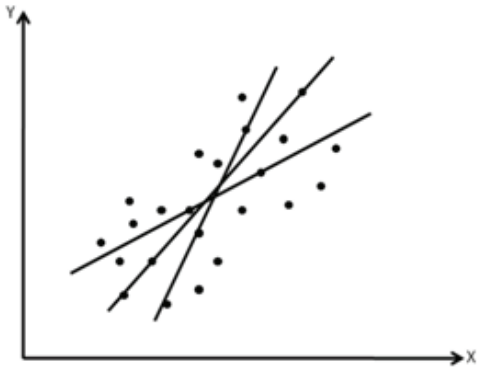


Fig. 8.12

Given a scatter plot, you must be able to draw the line of best fit. Best fit means that the sum of the squares of the vertical distances from each point to the line called deviation or error denoted by d_1, d_2…d_7 in Fig. 8.13 is at a minimum.

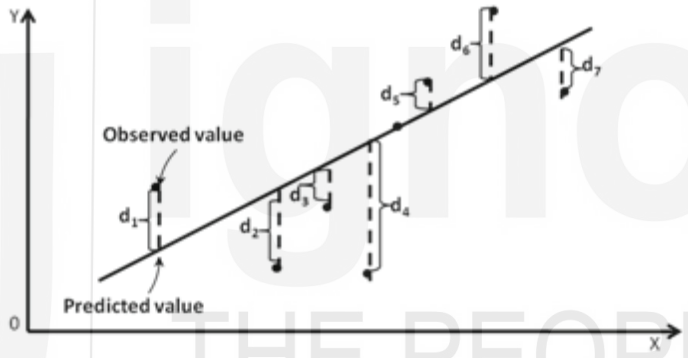


Fig. 8.13

The reason you need a line of best fit is that the values of Y will be predicted from the values of X; hence, the closer the points are to the line, the better the fit and the prediction will be. When *r* is positive, the line slopes upward and to the right. When *r* is negative, the line slopes downward from left to right. A line fitted by the method of least squares is known as the line of best fit, as it satisfy the following rules:

a) The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.

$$\sum(Y - Y_e) = 0$$

Where Y is the observed value and Y_e the value obtained from the regression analysis.

b) The sum of the squares of these deviations is less than the sum of squares of deviations from any other line, i.e.

$$\sum(Y - Y_e)^2 < \sum (Y - A_i)^2$$

Where A_i = corresponding values of points on any other straight line.

8.4.2 Regression Lines Equation

When we consider X as independent and Y as dependent variable, we have the linear regression model of Y on X, on the other hand with X as dependent and Y as independent variable, we have the linear regression model of X on Y. The corresponding two regression equations are given below:

Equation for X on Y:

$$X = a' + b'Y$$

Equation for Y on X:

$$Y = a + bX,$$

Where X, Y are variables, and a', b', a, b are constants whose values are to be determined.

For the equation, X = a' + b'Y, the normal equations are

$$\sum X = na' + b' \sum Y \text{ and}$$

$$\sum XY = a' \sum Y + b' \sum Y^2$$

For the equation, Y= a + bX, the normal equations are

$$\sum Y = na + b \sum X \text{ and}$$

$$\sum XY = a \sum X + b \sum X^2$$

From these normal equations the values of a', b', a and b can be determined.

8.4.3 Regression Coefficients

Solving the above sets of normal equations simultaneously, the values of regression coefficients can be estimated using the following formulae:

$$b = b_{YX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\sum xy}{\sum x^2}$$

$$b' = b_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2} = \frac{\sum xy}{\sum y^2}$$

Where x = $X - \bar{X}$, y = $Y - \bar{Y}$

The slight algebraic treatment (i.e. multiplying and dividing the formula of b_{yx} by σ_y and of b_{xy} by σ_x, and re-arranging the values) on these formulae also provides the following alternative formulae for regression coefficients:

$$b = b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b' = b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

Properties of Regression Co-efficient:

1) The regression line will always pass through the point whose X coordinate is the mean of the X values and whose Y coordinate is the mean of the Y values, i.e. $(\bar{X}, \bar{Y})$.

2) Both regression coefficients must have the same sign, i.e. either both will be positive or negative.

3) Correlation coefficient is the geometric mean of the regression coefficients, i.e.

$$r = \pm\sqrt{b_{XY} b_{YX}}$$

4) The sign of the correlation coefficient and the sign of the slope of the regression line will always be the same. That is, if *r* is positive, then both b_{XY} and b_{YX} will be positive; if *r* is negative, then both slope coefficients will be negative. The reason is that the numerators of the formulas are the same and determine the signs of *r* and regression coefficients, and the denominators being squares, are always positive.

5) If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.

6) Regression coefficients are independent of origin but not of scale.

7) Arithmetic mean of b_{XY} and b_{YX} is equal to or greater than the coefficient of correlation. Symbolically

$$\frac{b_{XY} + b_{YX}}{2} \geq r$$

8) If *r* = 0, the variables are uncorrelated, the lines of regression become perpendicular to each other.

9) If r = ± 1, the two lines of regression coincide each other

10) The angle between the regression lines indicates the degree of dependence between the variables.

11) The lines of regression intersect at the mean values of the variables X and Y, i.e., $(\bar{X}, \bar{Y})$.

8.4.4 Predictability of Regression Equations

Using the values of the regression coefficients as derived above, regression equations can be re-written in more practical forms as follows:

Regression equation for Y on X:

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

Regression equation for X on Y:

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

These two regression equations represent entirely two different lines. In other words, regression of Y on X is a function of X, which can be written as

Y = F(X) and regression of X on Y is a function of Y, which can be written as X = F(Y). The variables X and Y are not inter changeable. It is mainly due to the fact that in, regression of Y on X, Y is the dependent variable, X is the independent variable. That is to say: For the given values of X we can find the estimates of Y. Similarly, the estimates X for the values of Y can be obtained only from regression equation of X on Y.

The magnitude of the change in one variable, when the other variable changes exactly by one unit, is called a marginal change. The value of slope of the regression line equation represents the marginal change. When r is not significantly different from 0, the best predictor of Y is the mean of the data values of Y. For valid predictions, the value of the correlation coefficient must be significant. Also, the following assumptions must be met:

1) For any specific value of the independent variable X, the value of the dependent variable Y must be normally distributed about the regression line, as shown in the following Fig. 8.14(a)

2) The standard deviation of each of the dependent variables must be the same for each value of the independent variable, as shown in Fig. 8.14(b)

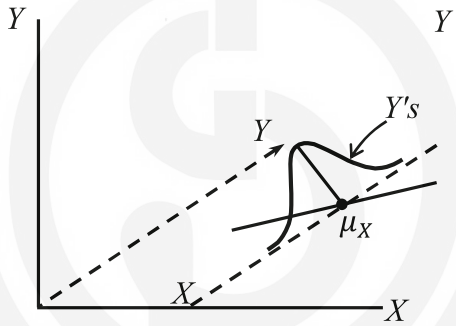


Fig. 8.14(a)

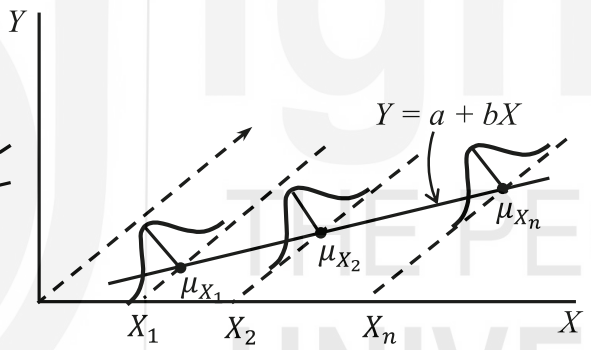


Fig. 8.14(b)

Identifying the Influencing Observation or Outlier

A scatter plot should be checked for outliers. An outlier is a point that seems out of place when compared with the other points. Some of these points can affect the equation of the regression line. When this happens, the points are called influential points or influential observations. When a point on the scatter plot appears to be an outlier, it should be checked to see if it is an influential point. An influential point tends to "pull" the regression line toward the point itself. To check for an influential point, the regression line should be graphed with the point (suspected to be the influential point) included in the data set. Then a second regression line should be graphed that excludes that point from the data set. If the position of the second line is changed considerably, the point is said to be an influential point. Points that are outliers in the horizontal direction are more likely to be influential points than that are outliers in the vertical direction.

We should use our judgement as to whether to include influential observations in the final analysis of the data. If the researcher feels that the

observation is not necessary, then it should be excluded so that it does not influence the results of the study. However, if it is felt that it is necessary, then we should obtain additional data values whose values are near the value of the influential point and then include them in the equations.

8.4.5 Coefficient of Determination

In previous sections it is stated that if the correlation coefficient is significant, the equation of the regression line can be determined. Also, for various values of the independent variable X, the corresponding values of the dependent variable Y can be predicted. The predictability of a regression equation could also be judged by its coefficient of determination.

It will be worthwhile to define the different types of variation associated with the regression model. Consider the following hypothetical regression model:

X	1	2	3	4	5
Y	10	8	12	16	20

The equation of the regression line is Y_e= 4.8 + 2.8X, and r = 0.919. The sample Y values are 10, 8, 12, 16, and 20. The predicted values, designated by Y_e, for each X can be found by substituting each X value into the regression equation and finding Y_e. For example, when X = 1,

$$Y_e = 4.8 + 2.8X = 4.8 + (2.8)(1) = 7.6$$

Now, for each X, there is an observed Y value and a predicted Y_e value. It may be recalled from the previous section that the closer the observed values are to the predicted values, the better the fit is and the closer r is to +1 or −1.

The total variation $\sum(Y - \bar{Y})^2$ is the sum of the squares of the vertical distances each point is from the mean. The total variation can be divided into two parts: that which is attributed to the relationship of X and Y and that which is due to chance. The variation obtained from the relationship (i.e., from the predicted Y_e values) is $\sum(Y_e - \bar{Y})^2$ and is called the explained variation. Most of the variations can be explained by the relationship. The closer the value r is to +1 or −1, the better the points fit the line and the closer $\sum(Y_e - \bar{Y})^2$ is to $\sum(Y - \bar{Y})^2$. In fact, if all points fall on the regression line, $\sum(Y_e - \bar{Y})^2$ will equal $\sum(Y - \bar{Y})^2$, since Y_e is equal to Y in each case.

On the other hand, the variation due to chance, found by $\sum(Y - Y_e)^2$, is called the unexplained variation. This variation cannot be attributed to the relationship. When the unexplained variation is small, the value of r is close to +1 or − 1. If all points fall on the regression line, the unexplained variation will be 0. Hence, the total variation is equal to the sum of the explained variation and the unexplained variation. That is,

$$\sum(Y - \bar{Y})^2 = \sum(Y_e - \bar{Y})^2 + \sum(Y - Y_e)^2$$

These values are shown in Fig. 8.15.

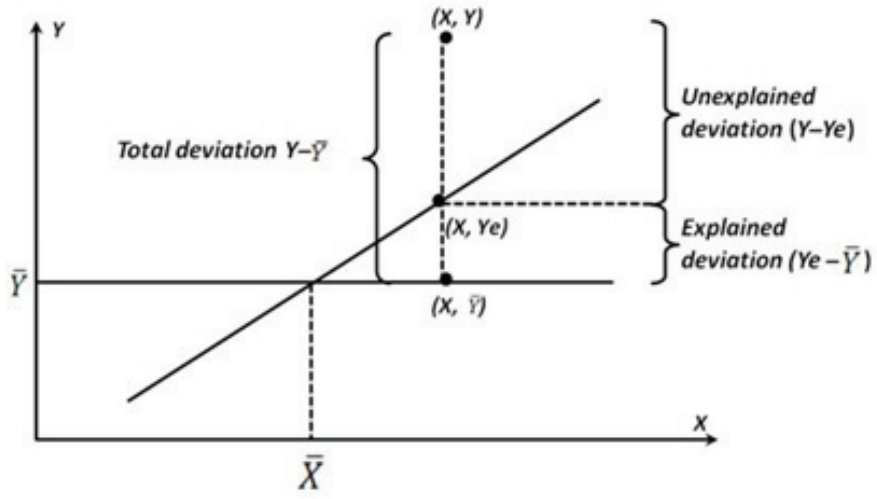


Fig. 8.15

The *coefficient of determination* is thus the ratio of the explained variation to the total variation and is denoted by r^2. That is,

$$r^2 = \frac{explained\ variation}{total\ variation} = \frac{\sum(Y_e - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

The coefficient of determination is a measure of the variation of the dependent variable that is explained by the regression line and the independent variable. It is usually easier to find the coefficient of determination by squaring r and converting it to a percentage. Therefore, if r = 0.90, then r^2 = 0.81, which is equivalent to 81%. This result means that 81% of the variation in the dependent variable is accounted for by the variations in the independent variable. The rest of the variation, 0.19, or 19%, is unexplained. This value is called the *coefficient of non-determination* and is found by subtracting the coefficient of determination from 1. As the value of r approaches 0, r^2 decreases more rapidly. For example, if r = 0.6, then r^2 = 0.36, which means that only 36% of the variation in the dependent variable can be attributed to the variation in the independent variable.

Check Your Progress 3

1) What do you mean by 'line of best fit' in regression analysis?

……

……

……

……

2) In a two variable case, define the regression equations, their regression coefficients and their properties?

……

……

……

……

3) What do you understand by the term 'predictability of regression equations' and briefly explain the assumptions for the valid predictions?

.....

.....

.....

.....

4) Define the term 'coefficient of determination' and discuss its significance in regression analysis?

.....

.....

.....

.....

.....

8.5 RELIABILITY OF REGRESSION MODEL – STANDARD ERROR OF ESTIMATE

When a y_e value is predicted for a specific x value, the prediction is a point prediction. However, a prediction interval about the value can be constructed, just as a confidence interval was constructed for an estimate of the population mean in the Unit 7. The prediction interval uses a statistic called the standard error of the estimate.

8.5.1 Standard Error of Estimate: Concept and Estimation

The standard error of the estimate, denoted by s_{est}, is the standard deviation of the observed Y values about the predicted values. The formula for the standard error of the estimate is

$$s_{est} = \sqrt{\frac{\sum(Y - Y_e)^2}{n-2}}$$

The standard error of the estimate is similar to the standard deviation, but the mean is not used. As can be seen from the formula, the standard error of the estimate is the square root of the unexplained variation—that is, the variation due to the difference of the observed values and the expected values—divided by n—2. So the closer the observed values are to the predicted values, the smaller the standard error of the estimate will be.

Example:

A researcher collects the following data and determines that there is a significant relationship between the age of a copy machine and its monthly maintenance cost. The regression equation is Y_e = 55.57 + 8.13X. Find the standard error of the estimate.

Machine	Age (X) (in years)	Monthly Cost (Y)
A	1	$62
B	2	78
C	3	70
D	4	90
E	5	93
F	6	103

Solution:

Step 1: Make a table as below:

X	Y	Y_e	$(Y - Y_e)$	$(Y - Y_e)^2$
1	62			
2	78			
3	70			
4	90			
5	93			
6	100			

Step 2: Using the regression line equation $Y_e = 55.57 + 8.13X$, compute the predicted values Y_e for each X and place the results in the column labelled Y_e.

Step 3: For each Y, subtract Y_e and place the answer in the column labelled $(Y - Y_e)$

Step 4: Square the numbers found in step 3 and place the squares in the column labelled $(Y - Y_e)^2$

Step 5: Find the sum of the numbers in the last column, the complete table is given:

X	Y	Y_e	$(Y - Y_e)$	$(Y - Y_e)^2$
1	62	63.70	−1.70	2.89
2	78	71.83	6.17	38.0689
3	70	79.96	−9.96	99.2016
4	90	88.09	1.91	3.6481
5	93	88.09	4.91	24.1081
6	103	104.35	−1.35	1.8225
				169.7392

Step 6: Substitute in the formula and find s_{est}.

$$s_{est} = \sqrt{\frac{\sum(Y - Y_e)^2}{n-2}}$$

$$= \sqrt{\frac{169.7392}{6-2}} = 6.51$$

In this case, the standard deviation of observed values about the predicted values is 6.51.

8.5.2 Prediction Interval

The basic objective of the regression analysis is to make the prediction about the Population Regression Function based on the Sample Regression Function. Therefore, the researcher would attempt to make a point estimate of the population regression coefficients; however, it is not always possible to establish the reliability of the point estimator. Therefore, it is always considered worthwhile to define the intervals with certain confidence within which the estimated values will fall so that estimates are trusted by the decision makers.

The standard error of the estimate can be used for constructing a prediction interval (similar to a confidence interval) about Y_e value. When a specific value X is substituted into the regression equation, Y_e that you get is a point estimate for Y. For example, if the regression line equation for the age of a machine and the monthly maintenance cost is Y_e = 55.57 + 8.13X, then the predicted maintenance cost for a 3-year-old machine would be Y_e= 55.57 + 8.13(3), or $79.96. Since this is a point estimate, you have no idea how accurate it is. But you can construct a prediction interval about the estimate. By selecting a significance level (α), you can achieve a (1−α) × 100% confidence that the interval contains the actual mean of the Y values that correspond to the given value of X.

The reason is that there are possible sources of prediction errors in finding the regression line equation. One source occurs when finding the standard error of the estimate s_{est}. Two others are errors made in estimating the slope and the Y_e intercept, since the equation of the regression line will change somewhat if different random samples are used when calculating the equation.

Formula for the Prediction Interval about a value Y_e , using the t-distribution discussed in the Unit 7, will be as below:

$$Y_e - t_{\alpha/2}\, s_{est}\sqrt{1 + \frac{1}{n} + \frac{n(X - \bar{X})^2}{n\sum X^2 - (\sum X)^2}} < Y$$

$$< Y_e + t_{\alpha/2}\, s_{est}\sqrt{1 + \frac{1}{n} + \frac{n(X - \bar{X})^2}{n\sum X^2 - (\sum X)^2}}$$

with n − 2 degree of freedom.

Example: For the problem taken in the previous section for calculating the Standard Error of Estimate, find the 95% prediction interval for the monthly maintenance cost of a machine that is 3 years old. (Given $t_{\alpha/2}$ = 2.776)

Solution:

Step 1: Find $\sum X, \sum X^2, and\ \bar{X}$

$$\sum X = 21,\ \sum X^2 = 91,\ and\ \bar{X} = \frac{21}{6} = 3.5$$

Step 2: Find Y_e for X = 3,

$Y_e = 55.57 + 8.13(3)$, or 79.96

Step 3: Finds s_{est} as done in the problem in previous section.

Step 4: Substitute in the formula and solve:

$t_{\alpha/2} = 2.776$, d.f. = 6 – 2 = 4 for 95%

$$Y_e - t_{\alpha/2}\, s_{est}\sqrt{1+\frac{1}{n}+\frac{n(X-\bar{X})^2}{n\sum X^2-(\sum X)^2}} < Y < Y_e + t_{\alpha/2}\, s_{est}\sqrt{1+\frac{1}{n}+\frac{n(X-\bar{X})^2}{n\sum X^2-(\sum X)^2}}$$

$$79.96 - (2.776)(6.51)\sqrt{1+\frac{1}{6}+\frac{6(3-3.5)^2}{6(91)-(21)^2}} < Y < 79.96 + (2.776)(6.51)\sqrt{1+\frac{1}{6}+\frac{6(3-3.5)^2}{6(91)-(21)^2}}$$

$$79.96 - (2.776)(6.51)(1.087) < Y < 79.96 + (2.776)(6.51)(1.087)$$

$$79.96 - 19.64 < Y < 79.96 + 19.64$$

$$60.32 < Y < 99.6$$

Hence, you can be 95% confident that the interval 60.53< Y <99.39 contains the actual value of Y.

Activity 3: Find the Regression line for the data of Activity 2

Find the coefficients for the regression line for the data of the regression example of this section (Age X (years) versus Monthly Cost Y)

How to do it?

- From the *Data* Tab Select *Analysis* icon option on the ribbon. It will open a *Data Analysis* dialog box, select *Regression* in this dialog box and Press *OK* button. The Output shown by Regression is shown as the following:

Summary Output

Regression Statistics	
Multiple R	0.9255044
R Square	0.8565585
Adjusted R Square	0.8206981
Standard Error	6.5141985
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1013.594	1013.594	23.885929	0.0081
Residual	4	169.7391	42.43478		
Total	5	1183.333			

	Coefficients	*Standard Error*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	55.565217	6.149984	9.035018	0.0008313	38.49	72.64	38.4901237	72.64031
Age x (Years)	8.1304348	1.663576	4.887323	0.0081177	3.5116	12.749	3.51160674	12.74926

You can analyse and compare the coefficients and as per this section. You can ignore other statistics for the present case.

Check Your Progress 4

1) What do you mean by 'Standard error of estimate' in regression analysis and how it is estimated?

……

……

……

2) Construct the prediction interval for standard error of estimate using t-distribution?

……

……

……

8.6 TESTING THE DIFFERENCE BETWEEN TWO MEANS, TWO PROPORTIONS, AND TWO VARIANCES

There are, however, many instances when researchers wish to compare two population means, using experimental and control groups. For example, the average lifetimes of two different brands of bus tires might be compared to

see whether there is any difference in tread wear. Two different brands of fertilizer might be tested to see whether one is better than the other for growing plants. Or two brands of cough syrup might be tested to see whether one brand is more effective than the other. The basic concepts of hypothesis testing explained in Unit 7 are used here with the z-test for comparing two means and proportions and F-test for comparing two variances.

8.6.1 Testing the Difference between Two Means: Using the z-test and t-test

Suppose a researcher wishes to determine whether there is a difference in the average age of students who enroll in a nursing programme at a community college and those who enroll in a nursing programme at a university. In this case, the researcher is not interested in the average age of all beginning nursing students; instead, he is interested in comparing the means of the two groups. His research question is, does the mean age of nursing students who enroll at a community college differ from the mean age of nursing students who enroll at a university? Here, the hypotheses are:

$H_0: u_1 = u_2$

$H_1: u_1 \neq u_2$

Where

u_1 = mean age of all beginning nursing students at the community college

u_2 = mean age of all beginning nursing students at the university

Another way of stating the hypotheses for this situation is

$H_0: u_1 - u_2 = 0$

$H_1: u_1 - u_2 \neq 0$

If there is no difference in population means, subtracting them will give a difference of zero. If they are different, subtracting will give a number other than zero. Both methods of stating hypotheses are correct. The theory behind testing the difference between two means is based on selecting pairs of samples and comparing the means of the pairs. The population means need not be known. All possible pairs of samples are taken from populations.

The variance of the difference $\bar{X}_1 - \bar{X}_2$ is equal to the sum of the individual variances of $\bar{X}_1$ and $\bar{X}_2$. That is,

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}$$

Where $\sigma^2_{\bar{X}_1} = \frac{\sigma_1^2}{n_1}$ and $\sigma^2_{\bar{X}_2} = \frac{\sigma_2^2}{n_2}$

So the standard deviation of $\bar{X}_1 - \bar{X}_2$ is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Formula for the z-test for comparing two means from independent populations is based on the general format:

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Where $\bar{X}_1 - \bar{X}_2$ is the observed difference, and the expected difference $\mu_1 - \mu_2$ is zero when the null hypothesis is $\mu_1 = \mu_2$, since that is equivalent to $\mu_1 - \mu_2 = 0$. Finally, the standard error of the difference is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

In the comparison of two sample means, the difference may be due to chance, in which case the null hypothesis will not be rejected, and the researcher can assume that the means of the populations are basically the same. The difference in this case is not significant. On the other hand, if the difference is significant, the null hypothesis is rejected and the researcher can conclude that the population means are different.

Confidence intervals for the difference between two means can also be found. When you are hypothesising a difference of zero, if the confidence interval contains zero, the null hypothesis is not rejected. If the confidence interval does not contain zero, the null hypothesis is rejected. Confidence intervals for the difference between two means can be found by using this formula:

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The z-test was used to test the difference between two means when the population standard deviations were known and the variables were normally or approximately normally distributed, or when both sample sizes were greater than or equal to 30. When σ_1 and σ_2 are not known but sample sizes n_1 and n_2 are large (greater than 30) then their estimates s_1^2 and s_2^2 can be used and z-test given in the following can be used:

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

8.6.2 Testing the Difference between Proportions Using z-test

The z-test with some modifications can be used to test the equality of two proportions. For example, a researcher might ask, is the proportion of men who exercise regularly less than the proportion of women who exercise regularly? Is there a difference in the percentage of students who own a personal computer and the percentage of non-students who own one?

For two proportions, $\hat{p}_1 = \frac{X_1}{n_1}$ is used to estimate p_1 and $\hat{p}_2 = \frac{X_2}{n_2}$ is used to estimate p_2. The standard error of the difference is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Where $\sigma_{p_1}^2$ and $\sigma_{p_2}^2$ are the variances of the proportions, $q_1 = 1 - p_1$, $q_2 = 1 - p_2$, and n_1 and n_2 are the respective sample sizes.

There are two requirements for use of the z-test: (1) The samples must be independent of each other, and (2) n_1p_1 and n_1q_1 must be 5 or more, and n_2p_2 and n_2q_2 must be 5 or more.

The formula for the confidence interval for the difference between two proportions will be as below:

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

8.6.3 Testing the Difference between Two Variances: F-Test

In addition to comparing two means, statisticians are interested in comparing two variances or standard deviations. For example, is the variation in the temperatures for a certain month for two cities different? In another situation, a researcher may be interested in comparing the variance of the cholesterol of men with the variance of the cholesterol of women. For the comparison of two variances or standard deviations, an F-test is used.

If two independent samples are selected from two normally distributed populations in which the variances are equal ($\sigma_1^2 = \sigma_2^2$) and if the estimates of the variances s_1^2 and s_2^2 are compared as $\frac{s_1^2}{s_2^2}$, the sampling distribution of the variances is called the **F distribution**.

The characteristics of F-distribution are indicated below:

- The values of F cannot be negative, because variances are always positive or zero.
- The distribution is positively skewed.
- The mean value of F is approximately equal to 1.
- The F-distribution is a family of curves based on the degrees of freedom of the variance of the numerator and the degrees of freedom of the variance of the denominator.

The formula for F-statistic is given below:

$$F = \frac{s_1^2}{s_2^2}$$

where the larger of the two variances is placed in the numerator regardless of the subscripts. The F-test has two terms for the degrees of freedom: that

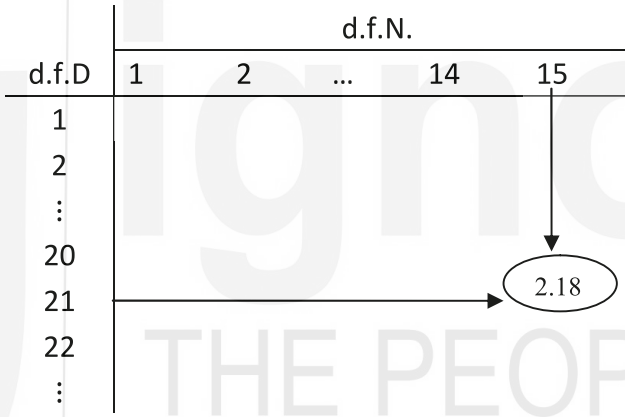
of the numerator, $n_1 - 1$, and that of the denominator, $n_2 - 1$, where n_1 is the sample size from which the larger variance was obtained.

F-distribution table gives the critical F values for α = 0.005, 0.01, 0.025, 0.05, and 0.10 (each α value involves a separate table for F-distribution). These are one-tailed values; if a two-tailed test is being conducted, then the α/2 value must be used. For example, if a two-tailed test with α = 0.05 is being conducted, then the 0.05/2 = 0.025 Table of F-distribution should be used.

Example: Find the critical value for a right-tailed F-test when α = 0.05, the degrees of freedom for the numerator (abbreviated d.f.N.) are 15, and the degrees of freedom for the denominator (d.f.D.) are 21.

Solution: Since this test is right-tailed with α = 0.05, use the 0.05 table. The d.f.N. is listed across the top, and the d.f.D. is listed in the left column. The critical value is found where the row and column intersect in the table. In this case, it is 2.18, as shown below:

$$\boldsymbol{\alpha = 0.05}$$



As noted above, when the F-test is used, the larger variance is always placed in the numerator of the formula. When you are conducting a two-tailed test, α is split; and even though there are two values, only the right tail is used. The reason is that the F statistic value is always greater than or equal to 1.

Check Your Progress 5

1) Giving examples, explain why the difference between the two sample statistics drawn from the same population or the different population is required to be tested?

   ……

   ……

   ……

2) Why difference between the two means is tested by using both z-test and t-test?

   ……

   ……

3) Describe the characteristics of F-distribution and using it construct the formula for testing the difference between the two variances drawn from different population?

……

……

……

8.7 ANALYSIS OF VARIANCES

The F-test, used to compare two variances discussed in the previous section can also be used to compare three or more means. This technique is called analysis of variance, or ANOVA.

It is used to test claims involving three or more means. For example, suppose a researcher wishes to see whether the means of the time it takes three groups of students to solve a computer problem using Fortran, Basic, and Pascal are different. The researcher will use the ANOVA technique for this test. The z- and t-tests should not be used when three or more means are compared due to the reasons given below:

- First, when you are comparing two means at a time, the rest of the means under study are ignored. With the F-test, all the means are compared simultaneously.
- Second, when you are comparing two means at a time and making all pair wise comparisons, the probability of rejecting the null hypothesis when it is true is increased, since the more t-tests that are conducted, the greater is the likelihood of getting significant differences by chance alone.
- Third, the more means there are to compare, the more t-tests are needed. For example, for the comparison of 3 means two at a time, 3 t tests are required. For the comparison of 5 means two at a time, 10 tests are required, and for the comparison of 10 means two at a time, 45 tests are required.

Assumptions for the F-Test for comparing three or more means include:

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent of one another.
- The variances of the populations must be equal.

It should be noted that even though you are comparing three or more means in this use of the F-test, variances are used in the test instead of means. With the F-test, two different estimates of the population variance are made. The first estimate is called the between-group variance, and it involves finding the variance of the means. The second estimate, the within-group variance, is made by computing the variance using all the data and is not affected by differences in the means. If there is no difference in the means, the between-group variance estimate will be approximately equal to

the within-group variance estimate, and the F-test value will be approximately equal to 1. The null hypothesis will not be rejected. However, when the means differ significantly, the between-group variance will be much larger than the within-group variance; the F-test value will be significantly greater than 1; and the null hypothesis will be rejected. Since variances are compared, this procedure is called analysis of variance (ANOVA). For a test of the difference among three or more means, the following hypotheses should be used:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

H_1: At least one mean is different from the others.

The degrees of freedom for this F test are d.f.N. = k–1, where k is the number of groups, and d.f.D. = N – k, where N is the sum of the sample sizes of the groups N = n_1+ n_2+.....+ n_k. The sample sizes need not be equal. The F-test to compare means is always right-tailed. The following example illustrates the computational procedure of ANOVA.

Example: A researcher wishes to try three different techniques to lower the blood pressure of individuals diagnosed with high blood pressure. The subjects are randomly assigned to three groups; the first group takes medication, the second group exercises, and the third group follows a special diet. After four weeks, the reduction in each person's blood pressure is recorded. At α = 0.05, the test claim that there is no difference among the means. The data are shown.

Medication	Exercise	Diet
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4
$\bar{X}_1 = 11.8$	$\bar{X}_2 = 3.8$	$\bar{X}_3 = 7.6$
$s_1^2 = 5.7$	$s_2^2 = 10.2$	$s_3^2 = 10.3$

Solution:

Step 1: State the hypotheses and identify the claim.

$H_0: \mu_1 = \mu_2 = \mu_3$ (claim)

H_1: At least one mean is different from the others

Step 2: Find the critical value. Since k = 3 and N = 15,

d.f.N = k – 1 = 3 – 1 = 2

d.f.D = N – k = 15 – 3 = 12

The critical value is 3.89, obtained from Table H in Appendix C with $\alpha = 0.05$.

Step 3: Compute the test value, using the procedure outlined here:

a) Find the mean and variance of each sample (these values are shown below the data).

b) Find the grand mean. The grand mean, denoted by $\bar{X}_{GM}$, is the mean of all values in the samples.

$$\bar{X}_{GM} = \frac{\sum X}{N} = \frac{10+12+9+\cdots+4}{15} = \frac{116}{15} = 7.73$$

When samples are equal in size, find $\bar{X}_{GM}$ by summing the $\overline{X's}$ and dividing by k, where k = the number of groups.

c) Find the between-groups varaince, denoted by s_B^2.

$$s_B^2 = \frac{\sum n_i(\bar{X}_i - \bar{X}_{GM})^2}{k-1}$$

$$= \frac{5(11.8-7.73)^2 + 5(3.8-7.73)^2 + 5(7.6-7.73)^2}{3-1}$$

$$= \frac{160.13}{2} = 80.07$$

Note: This formula find the variance among the means by using the sample sizes as weights and considers the differences in the means.

d) Find the within-group variance, denoted by s_w^2

$$s_w^2 = \frac{\sum(n_i-1)s_i^2}{\sum(n_i-1)}$$

$$= \frac{(5-1)(5.7) + (5-1)(10.2) + (5-1)(10.3)}{(5-1)+(5-1)+(5-1)}$$

$$= \frac{104.80}{12} = 8.73$$

Note: This formula finds an overall variance by calculating a weighted average of the individual variances. It does not involve using differences of the means.

e) Find the F-statistic value.

$$F = \frac{s_B^2}{s_w^2} = \frac{80.07}{8.73} = 9.17$$

Step 4: Make the decision. The decision is to reject the null hypothesis, since 9.17 > 3.89.

Step 5: Summarise the results. There are enough evidences to reject the claim and conclude that at least one mean is different from the others.

Activity 4: ANOVA

Perform the ANOVA test for the example given in this section.

How to do it?

- Enter the data of the change in Blood pressure in the example.
- From the *Data* Tab Select *Analysis* icon (right most) option on the ribbon. It will open a *Data Analysis* dialog box, select *Anova: Single Factor* in this dialog box and Press *OK* button. In the resulting dialog box select the range *$A$2:$C$7* (as our data is in this range; refer to Fig.

8.16); tick the Label in the first row box; select the output range to *$E$2* and Press *OK* button.

L21

	A	B	C	D	E	F	G	H	I	J	K
1	**Decrease in Blood Pressure**										
2	Medication	Exercise	Diet		Anova: Single Factor						
3	10	6	5								
4	12	8	9		SUMMARY						
5	9	3	12		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
6	15	0	8		Medication	5	59	11.8	5.7		
7	13	2	4		Exercise	5	19	3.8	10.2		
8					Diet	5	38	7.6	10.3		
9											
10											
11					ANOVA						
12					*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
13					Between Groups	160.1	2	80.067	9.167939	0.003831	3.885294
14					Within Groups	104.8	12	8.7333			
15											
16					Total	264.9	14				
17											

Fig. 8.16: Result on applying ANOVA: Single factor on data

You may analyse the results from the results as calculated in this section.

Check Your Progress 6

1) Define the term ANOVA and when this technique is used?

……

……

……

2) Why F-test is used for ANOVA instead of z-test or t-test?

……

……

……

3) Construct the formula for ANOVA and explain its degree of freedom carefully?

……

……

……

8.8 LET US SUM UP

Bivariate data analysis explains the relationships between the two variables existing in the real world. One way to determine whether a relationship exists is to use the statistical techniques known as correlation and

regression. The strength and direction of a linear relationship are measured by the value of the correlation coefficient. It can assume values between +1 and − 1 (both included). The closer the value of the correlation coefficient to +1 or − 1, the stronger the linear relationship is between the variables. A value of +1 or − 1 indicates a perfect linear relationship.

A positive relationship between two variables means that for small values of the independent variable, the values of the dependent variable will be small, and that for large values of the independent variable, the values of the dependent variable will be large. A negative relationship between two variables means that for small values of the independent variable, the values of the dependent variable will be large, and that for large values of the independent variable, the values of the dependent variable will be small. Relationships can be linear or curvilinear. To determine the shape, you are required to draw a scatter plot of the variables. If the relationship is linear, the data can be approximated by a straight line, called the regression line, or the line of best fit. The closer the value of r is to +1 or −1, the more closely the points will fit the line.

8.9 REFERENCES

1) Sharma, A.K. (2005). Text Book of Correlations and Regression, Discovery Publishing House, New Delhi.
2) O'Brien, D., P. Sharkey Scott, 2012. "Correlation and Regression", in Approaches to Quantitative Research – A Guide for Dissertation.
3) Jeremy Miles, Mark Shevlin (2000). Applying Regression and Correlation: A Guide for Students and Researchers, Published in UK.
4) Thomas Archdeacon (1994). Correlation and Regression Analysis: A Historian's Guide, Paperback, University of Wisconsin Press; USA.
5) R S N Pillai and Bagavathi (2016). Statistics: Theory and Practice, S Chand Publishing; Eight edition.

8.10 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 8.2
2) See Section 8.2

Check Your Progress 2

1) See Sub-Section 8.3.1
2) See Sub-Section 8.3.2
3) See Sub-Section 8.3.2
4) See Sub-Section 8.3.3

Check Your Progress 3

1) See Sub-Section 8.4.1

2) See Sub-Section 8.4.2 and 8.4.3
3) See Sub-Section 8.4.4
4) See Sub-Section 8.4.5

Check Your Progress 4

1) See Sub-Section 8.5.1
2) See Sub-Section 8.5.2

Check Your Progress 5

1) See Section 8.6
2) See Sub-Section 8.6.1
3) See Sub-Section 8.6.3

Check Your Progress 6

1) See Section 8.7
2) See Section 8.7
3) See Section 8.7

ignou
THE PEOPLE'S
UNIVERSITY

UNIT 9 MULTIVARIATE DATA ANALYSIS

Structure

9.0 Objectives

9.1 Introduction: Characteristics and Applications

9.2 What is Multivariate Analysis?

- 9.2.1 Multivariate Analysis in Statistical Terms
- 9.2.2 Some Basic Concepts of Multivariate Analysis

9.3 Classification of Multivariate Techniques

9.4 Types of Multivariate Techniques

- 9.4.1 Principal Components and Common Factor Analysis
- 9.4.2 Multiple Regression
- 9.4.3 Multiple Discriminant Analysis (MDA) and Logistic Regression
- 9.4.4 Canonical Correlation Analysis
- 9.4.5 Multivariate Analysis of Variance (MANOVA)
- 9.4.6 Conjoint Analysis
- 9.4.7 Cluster Analysis
- 9.4.8 Perceptual Mapping
- 9.4.9 Correspondence Analysis
- 9.4.10 Structural Equation Modeling (SEM)

9.5 Guidelines for Multivariate Techniques and Interpretation

- 9.5.1 Establish Practical as well as Statistical Significance
- 9.5.2 Recognise that Sample Size affects all Results
- 9.5.3 Know Your Data
- 9.5.4 Strive for Model Parsimony
- 9.5.5 Look to Your Errors
- 9.5.6 Validate Your Result
- 9.5.7 Using Software for Multivariate Analysis

9.6 A Structured Approach to Multivariate Model Building

9.7 Let Us Sum Up

9.8 References

9.9 Answers or Hints to Check Your Progress Exercises

9.0 OBJECTIVES

After going through this unit on multivariate analysis, you should be able to:

- state the concept of multivariate analysis and its application;
- know the measurement scales, measurement errors and their impact on multivariate analysis;
- explain the specific techniques included in the multivariate analysis;

- determine which multivariate technique is appropriate for a specific research problem;
- discuss the guidelines for application and interpretation of multivariate analysis; and
- discuss the approach to multivariate model building.

9.1 INTRODUCTION: CHARACTERISTICS AND APPLICATIONS

In Unit 2 and Unit 8, you have become familiar with the statistical techniques used in univariate (single variable) and bivariate (two variable) data analysis. As you may be well aware that most of the real time situations in social sciences, especially in economics, cannot be explained with the statistical techniques used in univariate or bivariate data analysis. In such situations most often more than two variables interact and reinforce each other making the explanation of such situations more complicated.

As a result, during the last fifty years, a number of statisticians have contributed to the development of several multivariate techniques. Today, these techniques are being applied in many fields such as economics, sociology, psychology, agriculture, anthropology, biology and medicine. These techniques are used in analysing social, psychological, medical and economic data, especially when the variables concerning research studies of these fields are supposed to be correlated with each other and when rigorous probabilistic models cannot be appropriately used. Applications of multivariate techniques in practice have been accelerated in modern times because of the advent of high speed electronic computers.

Multivariate techniques are largely empirical and deal with the reality; they possess the ability to analyse complex data. Accordingly in most of the applied and behavioural researches, we generally resort to multivariate analysis techniques for realistic results. Besides being a tool for analysing the data, multivariate techniques also help in various types of decision-making. For example, take the case of college entrance examination wherein a number of tests are administered to candidates, and the candidates scoring high total marks based on many subjects are admitted. This system, though apparently fair, may at times be biased in favour of some subjects with the larger standard deviations. Multivariate techniques may be appropriately used in such situations for developing norms as to who should be admitted in college. We may also cite an example from medical field. Many medical examinations such as blood pressure and cholesterol tests are administered to patients. Each of the results of such examinations has significance of its own, but it is also important to consider relationships between different test results or results of the same tests at different occasions in order to draw proper diagnostic conclusions and to determine an appropriate therapy. Multivariate techniques can assist us in such a situation. In view of all this, we can state that “if the researcher is interested in making probability statements on the basis of sampled multiple measurements, then the best strategy of data analysis is to use

some suitable multivariate statistical technique." This Unit presents a simplified overview of multivariate analysis. It stresses that multivariate analysis methods will necessarily influence not only the analytical aspects of the research but also the design and approach to data collection for decision making and problem solving. Although multivariate techniques show many characteristics similar to univariate or bivariate counterparts, several key differences arise in transition to a multivariate analysis. To illustrate this transition, this Unit presents a classification of multivariate techniques. However, the discussion is kept confined to detailed presentation of multiple regression and correlation and analysis of variance only as the present course, being of introductory nature, does not permit us to explore all the areas of multivariate analysis. It then provides general guidelines for the application of these techniques as well as the structural approach to the formulation, estimation and interpretation of results.

9.2 WHAT IS MULTIVARIATE ANALYSIS?

9.2.1 Multivariate Analysis in Statistical Terms

Multivariate analysis refers to all statistical techniques that simultaneously analyse multiple measurements on individuals or objects under investigation. Some researchers use multivariate simply to mean examining relationship between or among more than two variables. Others use the term only for the problems in all the multiple variables assumed to have a multivariate normal distribution. To be considered truly multivariate, however, all the variables must be random and interrelated in such ways that their different effects cannot meaningfully be estimated and interpreted separately.

Some authors state that purpose of multivariate analysis is to measure, explain and predict the degree of relationship among variates (weighted combination of variables). Thus, the multivariate character lies in the multiple variates (multiple combination of variables) and not only in the number of variables or observations. For the purpose of this Unit, multivariate analysis will include both multivariate techniques because we believe that the knowledge of multivariate techniques is an essential first step in understanding the multivariate analysis.

9.2.2 Some Basic Concepts of Multivariate Analysis

Although the roots of multivariate analysis lie in univariate and bivariate statistics, the extension to multivariate domain introduces additional concepts and issues of particular relevance. These concepts range from basic concept of variate to the specific issues dealing with the types of measurement scales used and the statistical issues of significance testing and the confidence levels.

The Variate

The building block of multivariate analysis is the variate, a linear combination of variables with empirically determined weights. The variables

are specified by the researchers whereas the weights are determined by multivariate technique to meet a specific objective. A variate of n weighted variables (X_1 to X_n) can be stated mathematically as:

$$\text{Variate value} = w_1X_1+w_2X_2+w_3X_3+ \ldots\ldots+w_nX_n$$

The result is single value representing a combination of the entire set of variables that best achieve the objective of the specific multivariate analysis. For example, in multiple regression, the variate is determined in a manner that maximises the correlation between the multiple independent variable and the dependent variable. In discriminant analysis, variate is formed so as to create scores for each observation that maximally differentiate between groups of observations. In each instance, the variate captures the multivariate character of the analysis. Thus in our discussion of techniques the variate is the focal point of analysis in many respects.

Measurement Scales

Various types of measurement scales like metric or non-metric, numeric or ordinal scales have already been discussed in Unit 7. The measurement scale is critical in determining which multivariate techniques are more suitable to the data, with consideration made both for dependent and independent variables.

Measurement Error and Multivariate Measurement

In particular, it assumes that any observation is composed of the true value plus some error comprising of random error value and systematic error. Random error is caused by any factors that randomly affect measurement of the variable across the sample. For instance, each person's mood can inflate or deflate their performance on any occasion. Systematic error is caused by any factors that systematically affect measurement of the variable across the sample. Apart from taking the full caution to reduce such errors, sometimes researcher choose to develop the multivariate measurement, also known as summed scales under which researcher repeat the observation number of times and take the average or by combining similar responses obtained through various variables develop the composite measure generally called the indicators.

The impact of measurement error and the poor reliability cannot directly be seen because they are embedded in the observed variables. Poor results are not always due to measurement error but the presence of measurement error is guaranteed to distort the observed relationships and make the multivariate techniques less powerful.

Types of Statistical Errors and Statistical Power

These concepts are already discussed in Unit 7 but certainly relevant for interpreting statistical inferences of multivariate analysis by the researcher to specify the acceptable levels of statistical error those results from using a sample (known as sampling error).

Having addresses the issues in extending multivariate techniques from univariate and bivariate origins, we present in the next section a

classification scheme to assist in the selection of the appropriate techniques by specifying the research objectives and the type of the data.

Check Your Progress 1

1) Define the term 'multivariate analysis'? Also discuss its characteristics and applications?

.....

.....

.....

.....

.....

2) In what way the terms like measurement scales, measurement errors, statistical errors and statistical power affect the multivariate analysis why?

.....

.....

.....

.....

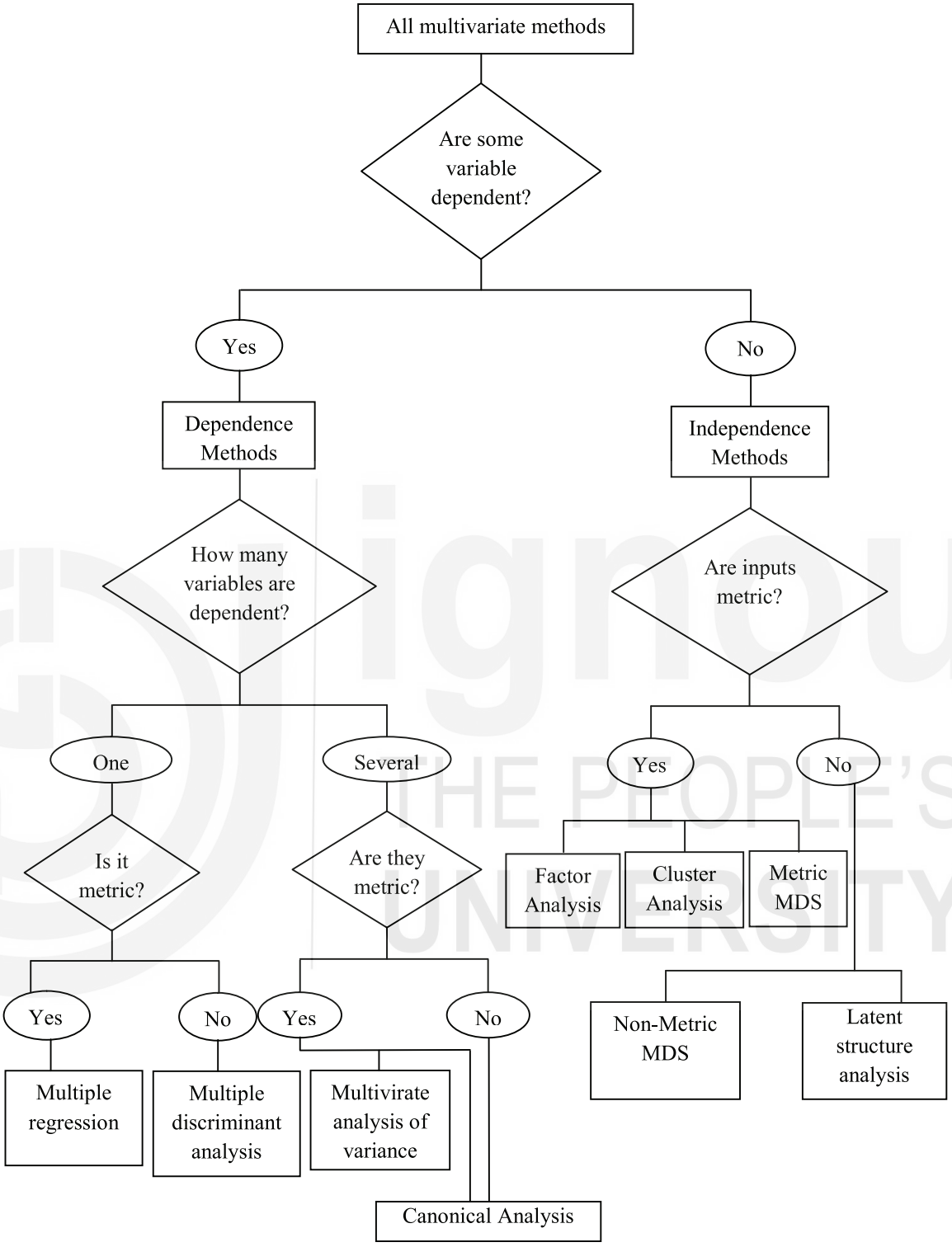
.....

9.3 CLASSIFICATION OF MULTIVARIATE TECHNIQUES

At present, there exist a great variety of multivariate techniques which can be conveniently classified into two broad categories viz., dependence methods and interdependence methods. This sort of classification depends upon the question: Are some of the involved variables dependent upon others? If the answer is 'yes', we have dependence methods; but in case the answer is 'no', we have interdependence methods. Two more questions are relevant for understanding the nature of multivariate techniques. Firstly, in case some variables are dependent, the question is how many variables are dependent? The other question is, whether the data are metric or non-metric? This means whether the data are quantitative, collected on interval or ratio scale, or whether the data are qualitative, collected on nominal or ordinal scale. The technique to be used for a given situation depends upon the answers to all these very questions. A flow chart that clearly exhibits the nature of some important multivariate techniques is shown in Fig. 9.1.

Thus, we have two types of multivariate techniques: one type for data containing both dependent and independent variables, and the other type for data containing several variables without dependency relationship. In the former category are included techniques like multiple regression analysis, multiple discriminant analysis, multivariate analysis of variance and canonical analysis, whereas in the latter category we put techniques like

factor analysis, cluster analysis, multidimensional scaling or MDS (both metric and non-metric) and the latent structure analysis.



Check Your Progress 2

1) Explain the basis of Classification of Multivariate Techniques?

.....

.....

9.4 TYPES OF MULTIVARIATE TECHNIQUES

Multivariate analysis is an ever expanding set of techniques for data analysis that encompasses a wide range of possible research institutions as evidenced by the classification scheme discussed in the previous section. The most established as well as emerging techniques include the following:

1) Principal Components and Common Factor Analysis
2) Multiple Regression and multiple correlation
3) Multiple discriminant analysis and logistic regression
4) Canonical correlation analysis
5) Multivariate analysis of variance
6) Conjoint Analysis
7) Cluster Analysis
8) Perceptual mapping
9) Correspondence analysis
10) Structural equation modelling

Now we introduce each of the multivariate techniques and briefly indicate their applications.

9.4.1 Principal Components and Common Factor Analysis

Factor analysis, including both principal component analysis and factor analysis, is a statistical approach that can be used to analyse interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions (actors). The objective is to find a way of condensing the information contained in the number of original variables into a smaller set of variates (factors) with a minimal loss of information. By providing an empirical estimate of structure of the variables considered, factor analysis becomes an objective basis for creating summed scales.

9.4.2 Multiple Regression

While discussing the bivariate analysis in Unit 8, we explained the concepts of simple linear regression and correlation. In simple linear regression, the regression equation contains one independent variable x and one dependent variable y_e and is written as

$$y_e = a + bx$$

In multiple regression, there are several independent variables and one dependent variable, and the equation is

$$y_e = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots + b_kx_k$$

where $x_1, x_2, \ldots, x_k$ are the independent variables. A multiple regression correlation R can also be computed to determine if a significant relationship exists between the independent variables and the dependent variable. Multiple regression analysis is used when a statistician thinks there are

several independent variables contributing to the variation of the dependent variable. This analysis then can be used to increase the accuracy of predictions for the dependent variable over one independent variable alone. Two examples for multiple regression analysis are when a store manager wants to see whether the amount spent on advertising and the amount of floor space used for a display affect the amount of sales of a product, and when a sociologist wants to see whether the amount of time children spent watching television and playing video games is related to their weight.

The assumptions for multiple regression are similar to those for simple regression:

1) For any specific value of the independent variable, the values of the y variable are normally distributed. (This is called the normality assumption.)

2) The variances (or standard deviations) for the y variables are the same for each value of the independent variable. (This is called the equal-variance assumption.)

3) There is a linear relationship between the dependent variable and the independent variables. (This is called the linearity assumption.)

4) The independent variables are not correlated. (This is called the non-multi-collinearity assumption.)

5) The values for the y variables are independent. (This is called the independence assumption.)

In multiple regression, as in simple regression, the strength of the relationship between the independent variables and the dependent variable is measured by a correlation coefficient. This multiple correlation coefficient is symbolised by R^2. The value of R^2 can range from 0 to +1; R^2 can never be negative. Closer the R^2 is to +1, the stronger the relationship; the closer to 0, the weaker the relationship. The value of R^2 takes into account all the independent variables and can be computed by using the values of the individual correlation coefficients. The formula for the multiple correlation coefficient when there are two independent variables is given below:

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}.r_{yx_2}.r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

$$R = \frac{Variability\ explained}{Total\ variability} = \frac{\sum(y_e - \overline{y})^2}{\sum(y - \overline{y})^2}$$

As with simple regression, R^2 is the coefficient of multiple determination, and it is the amount of variation explained by the regression model. The expression $1 - R^2$ represents the amount of unexplained variation, called the error or residual variation.

F-test is used to test the significance of R. The hypothesis are:

$H_0 : \rho = 0$ and $H_1 : \rho \neq 0$

Where ρ represents the population correlation coefficient for multiple correlation

The formula for the F-test is

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

where *n* is the number of data groups (x_1, x_2, ……. y)and k is the number of independent variables.

The degrees of freedom are d.f.N. = $n - k$ and d.f.D. = $n - k - 1$.

Adjusted R^2

Since the value of R^2 is dependent on n (the number of data pairs) and k (the number of variables), statisticians also calculate what is called an adjusted R^2, denoted by R^2_{adj}. This is based on the number of degrees of freedom. The formula for R^2_{adj}

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

The adjusted R^2 is smaller than R^2 and takes into account the fact that when n and k are approximately equal, the value of R may be artificially high, due to sampling error rather than a true relationship among the variables. This occurs because the chance variations of all the variables are used in conjunction with each other to derive the regression equation. Even if the individual correlation coefficients for each independent variable and the dependent variable were all zero, the multiple correlation coefficient due to sampling error could be higher than zero. Hence, both R^2 and R^2_{adj} are usually reported in a multiple regression analysis.

9.4.3 Multiple Discriminant Analysis (MDA) and Logistic Regression

Logistic regression and linear discriminant analyses are multivariate statistical methods which can be used for the evaluation of the associations between various covariates and a categorical outcome. Both methodologies have been extensively applied in research, especially in medical and sociological sciences.

MDA is the appropriate multivariate technique if the single dependent variable is dichotomous (e.g. male and female) or multi-chotomous (e.g. high-medium-low) and therefore nonmetric.

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Type of questions that a logistics regression can examine include such as how does the probability of

getting lung cancer (yes vs. no) change for every additional pound of overweight and for every pack of cigarettes smoked per day?

Or do body weight, calorie intake, fat intake, and participant age have an influence on heart attacks (yes vs. no)?

9.4.4 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a method for quantifying the linear dependence between two sets of variables. More precisely, CCA finds a linear combination of variables in one data set, and another linear combination of variables in a second data set, such that the two linear combinations have maximum correlation. The resulting maximised correlation is called a canonical correlation and generalises the concept of correlation to more than two variables.

9.4.5 Multivariate Analysis of Variance (MANOVA)

The purpose of a t-test is to assess the likelihood that the means for two groups are sampled from the same sampling distribution of means. The purpose of an ANOVA is to test whether the means for two or more groups are taken from the same sampling distribution. The multivariate equivalent of the t test is Hotelling's T^2. Hotelling's T^2 tests whether the two vectors of means for the two groups are sampled from the same sampling distribution. The purpose of MANOVA is to test whether the vectors of means for the two or more groups are sampled from the same sampling distribution. Just as Hotelling's T will provide a measure of the likelihood of picking two random vectors of means out of the same hat, MANOVA gives a measure of the overall likelihood of picking two or more random vectors of means out of the same hat.

There are two major situations in which MANOVA is used. The first is when there are several correlated dependent variables, and the researcher desires a single, overall statistical test on this set of variables instead of performing multiple individual tests. The second, and in some cases, the more important purpose is to explore how independent variables influence some patterning of response on the dependent variables.

For example, suppose a researcher is interested in the effect of different types of treatments on several types of anxieties: test anxiety, anxiety in reaction to minor life stresses, and so-called free-floating anxiety. The IV is different treatment with three levels (desensitisation, relaxation training, and a waiting-list control). After random assignment of subjects to treatments and a subsequent period of treatment, subjects are measured for test anxiety, stress anxiety, and free floating anxiety. Scores on all three measures for each subject serve as DVs. MANOVA is used to ask whether a combination of the three anxiety measures varies as a function of treatment.

MANOVA is statistically identical to discriminant analysis. The difference between the techniques is one of emphasis only. MANOVA emphasises the mean differences and statistical significance of differences among groups.

Discriminant analysis emphasises prediction of group membership and the dimensions on which groups differ.

9.4.6 Conjoint Analysis

Conjoint Analysis is a technique developed since the 1970s that allows businesses to work out and quantify the hidden rules people use to make trade-offs between different products and services and to quantify the values they place on different features or component parts of the offer. By understanding precisely how people make decisions and what they value in your products and services, you can work out the sweet spot or optimum level of features and services that balance value to the customer against cost to the company and forecast potential demand or market share in a competitive market situation.

9.4.7 Cluster Analysis

The Cluster Analysis is an explorative analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases, i.e., observations, participants, respondents. Cluster analysis is used to identify groups of cases if the grouping is not previously known. Because it is explorative it does make any distinction between dependent and independent variables.

9.4.8 Perpetual Mapping

A perceptual map is of the visual technique designed to show how the average target market consumer understands the positioning of the competing products in the marketplace. In other words, it is a tool that attempts to map the consumer's perceptions and understandings in a diagram. The word 'perceptual' comes from the word 'perception', which basically refers to the consumers' understanding of the competing products and their associated attributes.

Most common definitions used in analyses are:

"Perceptual maps measure the way products are positioned in the minds of consumers and show these perceptions on a graph whose axes are formed by product attributes".

"A perceptual map represents customer perceptions and preferences spatially by means of a visual display".

You will note that both definitions highlight that the information is visually presented on some form of graph or display.

9.4.9 Correspondence Analysis

Correspondence analysis is a statistical technique that provides a graphical representation of cross tabulations (which are also known as cross tabs, or contingency tables). Cross tabulations arise whenever it is possible to place events into two or more different sets of categories, such as product and

location for purchases in market research or symptom and treatment in medical testing.

9.4.10 Structural Equation Modelling (SEM)

Structural equation modelling (SEM) is a technique that allows separate relationships for each of a set of dependent variables. In its simplest sense, SEM provides the appropriate and the most efficient estimation technique for a series of separate multiple regression equations estimated simultaneously.

This brief analysis of multivariate techniques reveals its diverse character and provides solution to almost all kinds of real life situations. But its usage needs not only thorough understanding of these techniques but the situations where these are to be used also need to be understood clearly. Therefore, the correspondence between the situation and the identification of the techniques to be used in that situation needs to be clearly established. To assist in such situations, next section provides the guidelines for multivariate techniques and their interpretation.

Check Your Progress 3

1) List out the most emerging and tested Multivariate Techniques?

.....

.....

.....

.....

.....

2) Briefly describe the assumptions underlying the multiple regression analysis?

.....

.....

.....

.....

3) What is the difference between ANOVA and MANOVA?

.....

.....

.....

.....

.....

9.5 GUIDELINES FOR MULTIVARIATE TECHNIQUES AND INTERPRETATION

9.5.1 Establish Practical as well as Statistical Significance

Sometimes researcher solely relies on the statistical significance of the results ignoring their practical utility as a result brings out spurious results. For example, a regression analysis is undertaken to predict repurchase intentions, measured as the probability between 0 and 100 that the consumer will shop again that firm. The study is conducted and the results come back significant at the 0.05. Executives rush to embrace the results and modify the firm's strategy accordingly. What goes unnoticed, however is that even though the relationship is significant, the predictive ability was poor – so poor that the estimate of repurchase probability could vary as much as ±20% at the 0.05 significance level. The statistically significant relationship could thus have a range of 40% points. A customer predicted to have a 50% chance of return could really have probabilities from 30% to 70%, representing unacceptable levels upon which to take action. Had researchers and managers probed the practical or managerial significance of the results, they would have concluded that the relationship still needed refinements before it could be relied upon to guide strategy in any substantial sense.

9.5.2 Recognise that Sample Size Affects all Results

The discussion of statistical power in Unit 7 demonstrated the substantial impact sample size plays in achieving statistical significance both in small and large sample sizes.

9.5.3 Know Your Data

Multivariate techniques by their very nature identify complex relationships that are difficult to represent simply. As a result the tendency to accept the results without the typical examination by techniques of univariate and/or bivariate analysis sometimes leads to disastrous results. The multivariate analysis require an even more rigorous examination of the data because the influence of outliers, violations of assumptions, and missing data can be compounded across several variables to create substantial results.

9.5.4 Strive for Model Parsimony

Multivariate techniques are designed to accommodate multiple variables in the analysis. This feature however should not substitute for conceptual model development before multivariate techniques are applied. Although it is always more important to avoid omitting a critical predictable variable, termed specification error, the researcher must also avoid inserting variable indiscriminately letting the multivariate techniques 'sort out' the relevant variables for two fundamental reasons: (i) irrelevant variables usually increase a technique's ability to fit the sample data, but the expense of over fitting the sample data and making the results less generalisable to the

population., (ii) even though irrelevant variables typically do not bias the estimates of the relevant variables, they can mask the true effects due to an increase in the multi-collinearity. Multi-collinearity represents the degree to which any variable's effect can be predicted or announced for by the other variables in analysis. As multi-collinearity rises, the ability to define any variable effect is diminished. The addition of irrelevant variable can only increase the degree of multi-collinearity which makes the interpretation all the more difficult.

9.5.5 Look to Your Errors

The researcher should use the errors in prediction in multivariate analysis not as measure of failure or merely something to eliminate but as starting point for diagnosing the validity of the obtained results and an indication of the remaining unexplained relationships.

9.5.6 Validate Your Result

The ability of multivariate analyses to identify complex interrelationships also means results can be found that are specific to sample and not generalisable to the populations. The researcher is to ensure to validate results by one of several methods:

- Splitting the sample and using one subsample to estimate the model and second subsample to predict the accuracy
- Gathering a separate sample to ensure that the results are appropriate for other samples
- Employing the bootstrapping techniques which validate the multivariate by drawing large number of subsamples, estimating models for each sample, and then determining the values for the parameter estimates from the set of models by calculating the mean of each estimated coefficients across all the subsample models

Last but not the least, it is to be ensured that whenever a multivariate techniques is employed, focus should not be only to estimate the significant model but to ensure that it is representative of the population as a whole. The objective therefore should be not to just find a best fit to the sample data but instead to develop a model that best describes the population as a whole.

9.5.7 Using Software for Multivariate Analysis

It is very easy to use computer software to perform Multivariate analysis. The software will surely give you some results. However, the basic issue is to determine if the outcome is meaningful or not. In order to do so you need to have insight of the multivariate analysis which is beyond the scope of this Unit.

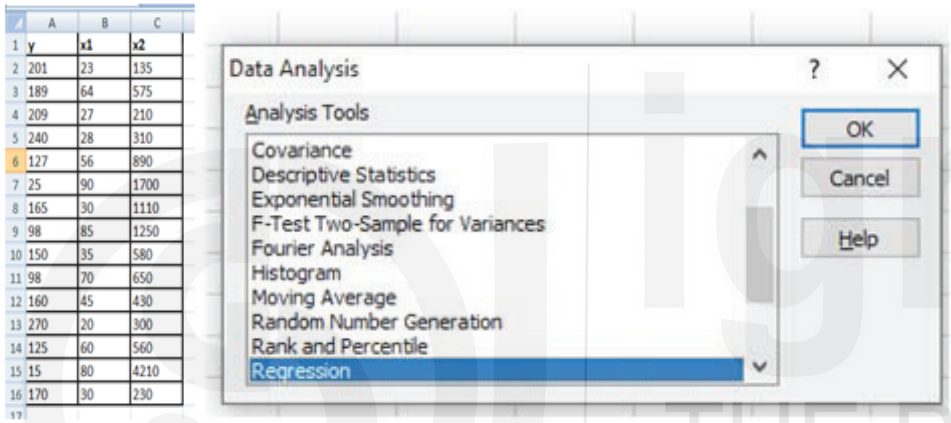
However, the following example will give you a basic idea how multivariable analysis can be performed using computer software.

Example: Find the regression results from the following data keeping y as dependent variable and x1 and x2 as independent variables:

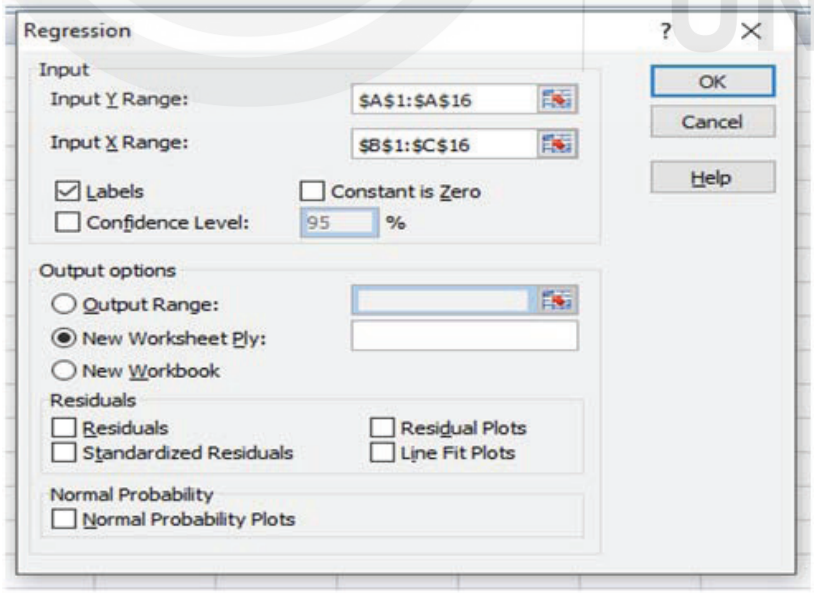
y	201	189	209	240	127	25	165	98	150	98	160	270	125	15	170
x1	23	64	27	28	56	90	30	85	35	70	45	20	60	80	30
x2	135	575	210	310	890	1700	1110	1250	580	650	430	300	560	4210	230

Solution: Please note that the range of x1 and x2 is not similar. When you get such a data in actual situation, it may be a good idea to scale them on comparable values. For this exercise, you can simply apply the multiple regression using MS-Excel.

You may enter the data in the MS-Excel. Following figure shows this data:

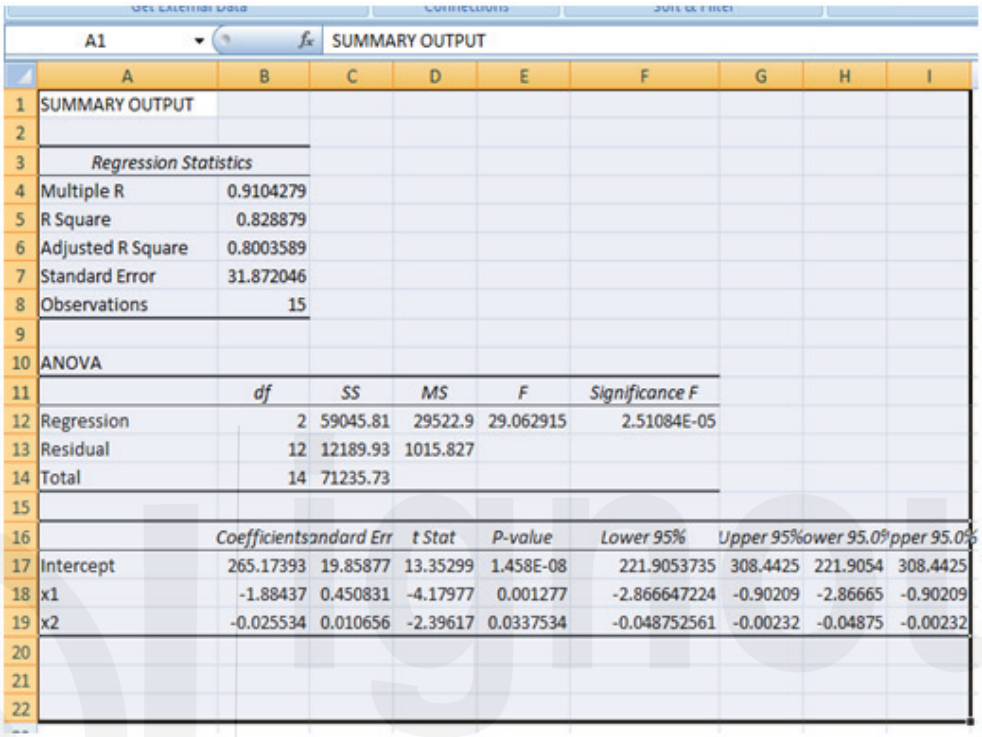


For performing regression anlaysis on the data select *Data* tab and *Data Analysis* option. You will see a Analysis Tools window as show above. Select *Regression* from it. The following window will appear:



Please notice the selection of input Y-range, it is the range of y variable of our data. Please note the Input X-range selects data of column B and C both. This tells excel that you have more than one (two in this case) independent

variables. Please also note that Check box before label has a tick mark. This has been done as the First row in the data contains the label (y, x1, x2 in our case). Also the New Worksheet Ply option has been selected. This will result in creating a new worksheet page for the data of regression analysis. Now, you can press the OK button on this screen. A new worksheet page will display the following:



A1 | SUMMARY OUTPUT

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.9104279							
5	R Square	0.828879							
6	Adjusted R Square	0.8003589							
7	Standard Error	31.872046							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	59045.81	29522.9	29.062915	2.51084E-05			
13	Residual	12	12189.93	1015.827					
14	Total	14	71235.73						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	265.17393	19.85877	13.35299	1.458E-08	221.9053735	308.4425	221.9054	308.4425
18	x1	-1.88437	0.450831	-4.17977	0.001277	-2.866647224	-0.90209	-2.86665	-0.90209
19	x2	-0.025534	0.010656	-2.39617	0.0337534	-0.048752561	-0.00232	-0.04875	-0.00232
20									
21									
22									

The regression data is extensive, but you can just look into the Significance F and p-values. In this case, all of them are below 0.05. Also the correlation coefficient (Multiple R) and R-Square are very close to one (high positive correlation). Please note that the value of Intercept coefficient is very high 265.17393 and coefficients for x1 is –1.88437 but coefficient of x2 is –0.025534. Is it due to difference in magnitude of values?

You may not be able to answer all the questions at present, but this simple illustration may be a good starting point for your future use of software.

9.6 A STRUCTURED APPROACH TO MULTIVARIATE MODEL BUILDING

The six step model building process provides a framework for developing, interpreting, and validating any multivariate analysis. These six steps or stages are:

Step 1: Define the Research Problem, Objective and Multivariate techniques to be used.

Step 2: Develop an Analyses Plan

Step 3: Evaluate the Assumptions underlying the Multivariate Analysis

Step 4: Estimate the Multivariate Model and Assess overall Model Fit

Step 5: Interpret the Variables

Step 6: Validate the Multivariate Model

The intent to present these steps is not to provide a rigid set of procedures but a sequential analysis that emphasise the model building approach. The approach focuses on a well-defined research plan, starting with a conceptual model detailing the relationships to be examined. Once defined in conceptual terms, the empirical issues can be addressed, including the selection of the specific multivariate technique and the implementation issues. After obtaining significant results, focus on interpretation, with special attention directed towards the variate. Finally the diagnostic measures ensure that the model is valid not only for the sample data but that it is generalisable as possible.

Check Your Progress 4

1) List out the important points required to be kept in view while interpreting the results of Multivariate analysis?

……

……

……

……

……

2) What is the difference between the statistical significance and the practical significance of the results?

……

……

……

……

……

3) To avoid any spurious results of a multivariate analysis, what precautions a researcher should take while undertaking such analysis?

……

……

……

……

……

9.7 LET US SUM UP

In real life situation most often more than two variables interact and reinforce each other. The techniques used in univariate or bivariate data analysis are not adequate to handle such types of situation. Multivariate

techniques are helpful to data analysis to handle analysis of multiple variables. Although the basic concepts of multivariate analysis are rooted in univariate and bivariate statistics, yet some additional concepts like measurement scales, assignment of weights, measurement error and multivariate measurement etc are introduced in multivariate domain. The set of techniques used in multivariate analysis include multiple regression, multiple discriminant analysis, multivariate analysis of variance, factor analysis, cluster analysis, correspondence analysis, latent structure analysis – analysis, structural equation modeling etc.

The basic objective underlying multivariate techniques is to represent a collection of massive data in a simplified way in other words, multivariate techniques transform a mass of observations into a smaller number of composite scores in such as a way that they may reflect as much information as possible contained in the raw data obtained concerning a research study. Thus, the main contribution of these techniques is in arranging a large amount of complex information involved in the real data into a simplified visible form. The unit concludes with some software based activities for multiple regression and correlation calculations.

9.10 REFERENCES

1) B.F.J. (1994). *Multivariate Statistical Methods–A Primer*, Chapman and Hall, London

2) Everitt B.S. and Dunn G. (2001). *Applied Multivariate Data Analysis*, Arnold, London

3) Härdle, W., Simar, L. (2007). *Applied Multivariate Statistical Analysis*, Springer.

4) Hair, J. F. Jr. (1995). *Multivariate Data Analysis*, 4th ed. Prentice-Hall.

5) Harris R.J. (1985). *A Primer in Multivariate Statistics*, Academic Press, New York. Manly

6) Johnson, Richard A.; Wichern, Dean W. (2007). *Applied Multivariate Statistical Analysis*, 6th edition,. Prentice Hall

7) Laurence G. Grimm, PhD, and Paul R. Yarnold (Edited), *Reading and Understanding Multivariate Statistics*, American Psychological Association, Washington, USA

8) Sharma, S. (1996). *Applied Multivariate Techniques*, University of South California, John Wiley & Sons, Inc.

9) Tabachnick B., Fidell, L. (2007). *Using Multivariate Statistics*, 5th edition Pearson Education. Inc.

9.11 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 9.1 and Sub-section 9.2.1
2) See Section 9.1 and Sub-section 9.2.1

Check Your Progress 2

1) See Sub-section 9.3.1

Check Your Progress 3

1) See Section 9.4
2) See Sub-section 9.4.3
3) See Sub-section 9.4.5

Check Your Progress 4

1) See Sub-section 9.5.1 to 9.5.6
2) See Sub-section 9.5.1
3) See Section 9.6

ignou THE PEOPLE'S UNIVERSITY

ignou
THE PEOPLE'S
UNIVERSITY