ગુજરાત રાજ્યના શિક્ષણવિભાગના પત્ર-ક્રમાંક
મશબ/1215/178/છ, તા. 24-11-2016-થી મંજૂર

STATISTICS

(Part 1)

Standard 12

PLEDGE

India is my country.
All Indians are my brothers and sisters.
I love my country and I am proud of its rich and varied heritage.
I shall always strive to be worthy of it.
I shall respect my parents, teachers and all my elders and treat everyone with courtesy.
I pledge my devotion to my country and its people.
My happiness lies in their well-being and prosperity.

રાજ્ય સરકારની વિનામૂલ્યે યોજના હેઠળનું પુસ્તક

Gujarat State Board of School Textbooks
'Vidyayan', Sector 10-A, Gandhinagar-382010

© Gujarat State Board of School Textbooks, Gandhinagar
Copyright of this book is reserved by Gujarat State Board of School Textbooks.
No reproduction of this book in whole or in part, in any form is permitted without
written permission of the Director, Gujarat State Board of School Textbooks.

Subject Adviser
Dr. R. T. Ratani

Authors
Dr. M. N. Patel (Convener)
Prof. Shubha A. Lagvankar
Dr. Chirag J. Trivedi
Dr. Kunjal H. Shah
Dr. Parag B. Shah
Shri Maheshbhai A. Patel
Dr. Yatin A. Parikh

Translation
Dr. M. N. Patel
Prof. Shubha A. Lagvankar
Dr. Chirag J. Trivedi
Dr. Kunjal H. Shah
Dr. Parag B. Shah

Reviewers
Shri Lalita K. Patel
Shri Chirag S. Shah
Shri Haresh N. Doshi
Smt. Deepa K. Lalwani
Shri Girish C. Naik
Shri C. P. Rajpurohit
Smt. Aparna S. Pathak
Shri Manoj G. Naidu

Language Reviewer
Dr. Chetan Mewada

Artist
Media Graphics

Co-ordinator
Dr. Chirag N. Shah
(Subject Co-ordinator : Commerce)

Preparation and Planning
Shri Haren Shah
(Dy. Director : Academic)

Lay-out and Planning
Shri Haresh S. Limbachiya
(Dy. Director : Production)

PREFACE

Gujarat State Board of School Textbooks has prepared new textbooks as per the new curricula developed by the Gujarat State Secondary and Higher Secondary Education Board and which has been sanctioned by the Education Department of the Government of Gujarat. A panel of experts from Universities/Colleges, Teachers Training Colleges and Schools have put lot of efforts in preparing the manuscript of the subject. It is then reviewed by another panel of experts to suggest changes and filter out the mistakes, if any. The suggestions of the reviewers are considered thoroughly and necessary changes are made in the manuscript. Thus, the Textbook Board takes sufficient care in preparing an error-free manuscript. The Board is vigilant even while printing the textbooks.

The Board expresses the pleasure to publish the Textbook of **Statistics (Part 1)** for **Std. 12** which is a translated version of Gujarati. The Textbook Board is thankful to all those who have helped in preparing this textbook. However, we welcome suggestions to enhance the quality of the textbook.

P. Bharathi (IAS)
Director
Executive President
Date : 16-11-2019
Gandhinagar

First Edition : 2017, Reprint : 2018, 2019, 2020

Published by : P. Bharathi, Director, on behalf of Gujarat State Board of SchoolTextbooks, 'Vidhyayn', Sector 10-A, Gandhinagar.

Printed by :

FUNDAMENTAL DUTIES

It shall be the duty of every citizen of India : *

(a) to abide by the Constitution and respect its ideals and institutions, the National Flag and the National Anthem;

(b) to cherish and follow the noble ideals which inspired our national struggle for freedom;

(c) to uphold and protect the sovereignty, unity and integrity of India;

(d) to defend the country and render national service when called upon to do so;

(e) to promote harmony and the spirit of common brotherhood amongst all the people of India transcending religious, linguistic and regional or sectional diversities; to renounce practices derogatory to the dignity of women;

(f) to value and preserve the rich heritage of our composite culture;

(g) to protect and improve the natural environment including forests, lakes, rivers and wild life, and to have compassion for living creatures;

(h) to develop scientific temper, humanism and the spirit of inquiry and reform;

(i) to safeguard public property and to abjure violence;

(j) to strive towards excellence in all spheres of individual and collective activity so that the nation constantly rises to higher levels of endeavour and achievement;

(k) to provide opportunities for education by the parent, the guardian, to his child, or a ward between the age of 6-14 years as the case may be.

*Constitution of India : Section 51-A

CONTENTS

1. Index Number 1

2. Linear Correlation 58

3. Linear Regression 116

4. Time Series 156

- Answers 183

●

1

Index Number

Contents :

1.1 Definition and Meaning of Index Number

1.2 Characteristics of Index Number

1.3 Uses of Index Number

1.4 Base year

1.4.1 Fixed Base method, merits and limitations

1.4.2 Chain Base method, merits and limitations

1.5 Conversion of Fixed Base to Chain Base and Chain Base to Fixed Base

1.6 Computation of Index Number

1.6.1 Laspeyre's Formula

1.6.2 Paasche's Formula

1.6.3 Fisher's Formula

1.7 Cost of Living Index Number

1.7.1 Explanation and Construction

1.7.2 Uses and Limitations

1.1 Definition and Meaning of Index Number

Price of an item, national income, supply, production, employment, unemployment, investment, import-export, cost of living, population of a country, birth rate and death rate vary continuously with time. Generally, the proportion and direction of these variations also keep changing. It is important to study the variations in the price and quantity of an item with respect to change in time. The planning for the future can be suitably done from the knowledge of these changes. The changes taking place in the values of the variable at two different time periods can be measured by the following methods :

(1) Method of absolute measure (difference) and (2) Method of relative measure (ratio)

We will understand this concept by the following illustration :

Suppose the data regarding the average price per kilogram of two items, wheat and rice, for a month in the year 2015 and year 2016 are as follows :

Item	Price per kilogram ₹	
	Year 2015	Year 2016
Wheat	24	30
Rice	40	46

Let us understand the comparision of variations in the prices of wheat and rice using the two methods stated above.

(1) Method of Absolute Measure (difference) : The price of wheat in the year 2015 was ₹ 24 which increased to ₹ 30 in the year 2016. Thus, the price per kilogram increased in the year 2016 by ₹ 6 with respect to the year 2015. Similarly, the price of rice has also increased by ₹ 6. This is obtained by the absolute difference. Thus, it can be said that there is same rise in price in both the items. But this is not true in reality because the prices per unit of these items are not same in the year 2015. Thus, the base for comparative study of prices in 2016 is different. Hence, this method is not appropriate to compare the variations in a variable. We shall now study the method of relative measure which is used in such situations.

(2) Method of Relative Measure (ratio) : A ratio of price of the commodity in the year 2016 is obtained with the price in the year 2015 in this method to find the relative changes of the price of the item in the year 2016.

Thus, ratio of prices of wheat $= \frac{30}{24} = 1.25$

ratio of prices of rice $= \frac{46}{40} = 1.15$

It can be known from these ratios that the relative increase in the prices of wheat and rice in the year 2016 is not same. The price of wheat in the year 2016 is 1.25 times the price in the year 2015, whereas the price of rice is 1.15 times its price in the year 2015. Thus, it can be said that change in the price of wheat is more than the change in the price of rice.

The ratios of prices of wheat and rice given here indicate the changes in prices at two different time periods. It is also called as relative change or price relative. Generally, the ratios are expressed as percentages to facilitate comparison. Hence,

Percentage change in the price of wheat $= 1.25 \times 100 = 125$ and

Percentage change in the price of rice $= 1.15 \times 100 = 115$

This is the relative percentage measure for the changes. Such a relative measure is called index number.

Thus, the percentage change in the value of a variable associated with any item for the given (current) period compared to its value in a fixed (base) period is called an index number.

Now, we shall obtain a relative measure for the collective change in the prices of these two mutually related items. The absolute method is not useful to find a measure for the overall change because many times the units expressing the prices of these two items may be different and it is not possible to combine the changes in these prices. The method of relative measure is used in such a situation. Since the relative measure is free from the unit of measurement, it is possible to combine the changes in the prices of the two items and it is convenient to find a mathematical measure for these changes. Now, we shall take a relative measure for the overall change from the changes in prices of grains, wheat and rice. We shall denote the price for the year 2016 as p_1 and the price of the year 2015 by p_0. p_0 is called base year price and p_1 is called the current year price. The ratio $\frac{p_1}{p_0}$ is called the price relative of that item.

We shall present this in a tabular form :

Item	**Price of base year 2015 (₹)** p_0	**Price of current year 2016 (₹)** p_1	**Price relative or Relative change** $= \frac{p_1}{p_0}$	**Percentage of Price relative** $= \frac{p_1}{p_0} \times 100$
Wheat	24	30	$\frac{30}{24} = 1.25$	125
Rice	40	46	$\frac{46}{40} = 1.15$	115
Total			**2.40**	**240**

The index obtained by multiplying the average of price relatives of the current year for these two items by 100 is called the price index number of the items for the current year. It is denoted by *I*. Thus,

Price index number of wheat and rice for the current year

$$= \frac{\text{Price relative of wheat} + \text{Price relative of rice}}{\text{No. of items}} \times 100$$

$$= \frac{1.25 + 1.15}{2} \times 100$$

$$= 120$$

Hence, price index number of wheat and rice for the current year $I = 120$. The price index number of wheat and rice $I = 120$ indicates that there is an overall rise of 20 percent in the prices of the two items in the year 2016 as compared to the year 2015. Index number is a relative measure based on ratio. Similarly, measure for the overall change can be obtained using the relative method by combining changes in the values of more than two variables. We can define the general index number for a group as follows.

"The average of the percentage change in the value of a variable associated with one or more items for the given (current) period compared to its value in the fixed (base) period is called a general index number for the group."

General index number for the group $I = \frac{\Sigma\left[\frac{p_{1i}}{p_{0i}}\right]}{n} \times 100$

Where, general index number I = Index number of current period with respect to comparison period

p_{1i} = Value of variable i for current period $\left(i = 1, 2, 3, ..., n\right)$

p_{0i} = Value of variable i for comparison period $\left(i = 1, 2, 3, ..., n\right)$

n = Number of values of the variable

The simple mean is used in the definition of general index number for the group of n items. But the weighted mean or geometric mean can also be used in the definition of general index number, which will be discussed later in this chapter.

In practice, several mutually related items are to be included and the data regarding their prices should be obtained to find a price index number. For example, wheat, rice, pulses, oil, ghee, jaggery, spices, vegetables are included in the category of food items. Thus, price index number for food is an index associated with the relative change or price relative for the prices of several related items.

Now, if we take a group of n such mutually related items then an index is found using relative change in price of each item in that group. An average measure obtained from them is called the price index numbers for the group. It can be written as the following formula :

General price index number $I = \frac{\Sigma\left[\frac{p_{1i}}{p_{0i}}\right]}{n} \times 100$

Where, p_{1i} = price of item i in current period $\left(i = 1, 2, 3, ..., n\right)$

p_{0i} = price of item i in base period $\left(i = 1, 2, 3, ..., n\right)$

n = number of items

Further, if we take a group of n such mutually related items then an index is found using relative change in the quantity of each item in that group. An average measure obtained from them is called the quantity index number for the group.

Note : The index number for production, import, export, unemployment, industrial output, etc. can be obtained by the above formula.

1.2 Characteristics of Index Number

Some of the characteristics of index number deduced from its definition are as follows :

(1) Index number is free from the unit as it is a relative measure.

(2) The changes in the values of the variable having different units can be compared using index number. Hence, index number is a comparative measure.

(3) Index number is a relative measure showing percentage change.

(4) Index number is a special average. It has all the characteristics of an average.

(5) The situation at two different periods can be compared by ratio with the standard (base) period using an index number.

1.3 Uses of Index Number

A general notion about the index number is, an index number is only used to find a measure for changes in the value of a variable or the price level. But now its use is not limited to the study of change in the price level. The index number is used in various fields in the current revolutionary age. Index number is a useful statistical tool to study the challenges in the given economic, political, social and industrial activities. Index number provides important guidance for planning the economic development of a country as it gives a comparative study of economic and industrial scenario of the country. Some of the uses of index number are as follows :

(1) Index Number for Trade : This index number provides useful guidance to study the general situation of economic activities of business and trade in the country.

(2) Wholesale Price Index Number : This index number measures the changes in the general price level in the country. This index number is useful to the government, producers and businessmen to take policy decisions such as knowing the demand and supply of items in the economy, estimating the future values and planning the future. The Reserve Bank of India uses this index number to take necessary steps to control inflation by studying the changes in price levels. Using the wholesale price index number, the rate of inflation is found as follows.

$$\text{Rate of inflation} = \frac{\left(\begin{array}{c}\text{Wholesale index}\\ \text{number of current year}\end{array}\right) - \left(\begin{array}{c}\text{Wholesale index number}\\ \text{of previous year}\end{array}\right)}{\text{Wholesale index number of previous year}} \times 100$$

(3) Cost of Living Index Number : This index number is useful to study the changes in the cost of living of people of different sections. This index number helps to determine purchasing power of money, salary to employees, dearness allowance, bonus, to calculate real wage and to devise tax policies by the government.

(4) Index Number of Human Development : This index number is useful to determine the state of human resource, standard of living, life expectancy and level of education and it gives information about human resource development.

(5) Index Number of National Income : This index number is useful to evaluate economic condition of the country and to determine targets for the five year plans by the government. The suggestions for increasing the national income, production and per capita income of the country can also be given using this index number by studying the changes in the national income of the country.

(6) Index Number of Industrial Production : This index number is very useful to study the changes in the production in industrial and craft fields. This index number is helpful to increase the rate of development of the country, planning industrial and trading activities.

(7) Index Number of Agricultural Production : This index number is useful to study the changes in the prices of agricultural production. The government plans agricultural policies using this index number. Moreover, this index number is helpful in forming policy to give proper support price to the farmers for their production.

(8) Index Number of Import-Export : This index number is useful to determine import-export policy, exchange rate, foreign exchange requirement and the rate of excise on goods and to provide necessary suggestions.

(9) Index Number for Employment : This index number provides the picture of employment, unemployment prevailing in the country. This shows the problems of unemployment which facilitates human resource planning.

(10) Index Number of Capital Investment : The changes in prices of shares and stocks, debentures, government securities and flow of capital investment can be studied by this index number. It also helps to estimate the trend of prices of shares and stocks.

(11) Index Number of Raw Material : This index number provides necessary guidance to traders, businessmen, economists, etc. for the policies of production-sales.

As barometer is used to predict weather, air pressure, cyclone and rain, the index number is a necessary tool for the measurement and comparative study of changes in the economic, business and social activities of the country. Hence, an index number is called the barometer of the economy of a country.

1.4 Base Year

In the construction of index numbers, the value of a variable for the current period is compared with the value of the variable with a fixed period (usually from the past). This fixed period or year is called the base year. The fixed year from the past can be the preceding year or any year before that. The period or year for which the value is to be compared with the base period or year is called the current period or year. For example, if the price of an item in the year 2016 is to be compared with the price of the same item in the year 2015, the year 2015 is called base year and the year 2016 is called the current year.

The year selected as base year should be standard or normal. It should be free from natural calamities like floods, draught, earthquake, abnormal man-made events like war, revolt, riot, strike, agitation, political events, economic disturbance or any unusual events. It is also necessary that the base year should not be from a distant past. If the base year selected is an unusual year and the values of the variable are unusually high or low then the value of index number could be misleading and it will not reveal the realistic picture of the current situation. Thus, the base year should be carefully selected while constructing index number.

The base year can be selected in two ways : (1) Fixed Base Method (2) Chain Base Method

1.4.1 Fixed Base Method

In this method, a stable period or year with usual events or situation is selected as a normal year or base year. But sometimes it becomes difficult to select a normal or base year. In this case, an average value of certain years is taken as the value of the variable for the base year. Index number is obtained by comparing value of the variable in the current year with the value of variable for the base year. The base year should be changed periodically so that it does not become a year of the distant past. The index number by fixed base method is obtained from the following formula :

$$\text{Index number } I = \frac{\text{Value of the variable in current year (period)}}{\text{Value of the variable in base year (period)}} \times 100$$

$$= \frac{p_1}{p_0} \times 100$$

Where, p_1 = Value of the variable in current year (period)

p_0 = Value of the variable in base year (period)

Illustration 1 : The data about wholesale prices of wheat in a region are as follows. Taking the year 2005 as the base year, prepare the index numbers for the price of the item for the remaining years. State the percentage increase in the price of wheat in the year 2013 from these index numbers.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013
Price per Quintal (₹)	1650	1690	1730	1750	1810	1850	1870	1900	1950

We will find the fixed base index number as the year 2005 is to be taken as the base year. The index number for the price of wheat in the year 2005 will be taken as 100.

Year	**Price of wheat per Quintal (₹)**	**Index number = $\frac{p_1}{p_0} \times 100$**
2005	1650	$\frac{1650}{1650} \times 100 = 100$
2006	1690	$\frac{1690}{1650} \times 100 = 102.42$
2007	1730	$\frac{1730}{1650} \times 100 = 104.85$
2008	1750	$\frac{1750}{1650} \times 100 = 106.06$
2009	1810	$\frac{1810}{1650} \times 100 = 109.70$
2010	1850	$\frac{1850}{1650} \times 100 = 112.12$
2011	1870	$\frac{1870}{1650} \times 100 = 113.33$
2012	1900	$\frac{1900}{1650} \times 100 = 115.15$
2013	1950	$\frac{1950}{1650} \times 100 = 118.18$

It can be said that the increase in the price of wheat in the year 2013 is $(118.18 - 100) = 18.18$ % with respect to the year 2005.

Illustration 2 : The prices per unit (₹) of six food items in the year 2014 and 2015 are given in the following table. Taking 2014 as the base year, compute the general index number for the price of food items and state the overall rise in prices of these food items.

Item	Unit	Price per unit (₹) of the item	
		Year 2014	Year 2015
Bread	Packet	25	28
Eggs	Dozen	30	35
Ghee	Tin	375	380
Milk	Litre	36	40
Cheese	Kilogram	440	500
Butter	Kilogram	265	300

A general index number for the price of these items for the current year 2015 is to be obtained with the base year 2014. We will find price relatives $\frac{p_1}{p_0}$ by taking base year price as p_0 and current year price as p_1. The calculation is shown in the following table :

Item	Price of item (₹)		Price relative = $\frac{p_1}{p_0}$	
	p_0	p_1		
Bread	25	28	$\frac{28}{25}$	= 1.1200
Eggs	30	35	$\frac{35}{30}$	= 1.1666
Ghee	375	380	$\frac{380}{375}$	= 1.0133
Milk	36	40	$\frac{40}{36}$	= 1.1111
Cheese	440	500	$\frac{500}{440}$	= 1.1364
Butter	265	300	$\frac{300}{265}$	= 1.1321
Total				**= 6.6795**

General index number of six food items I $= \dfrac{\Sigma\left[\frac{p_1}{p_0}\right]}{n} \times 100$

$= \dfrac{6.6795}{6} \times 100$

$= 111.33$

$\therefore$ General price index number of six food items is $I = 111.33$.

It can be seen from the value of the index number I that there is an overall rise in prices of food items by $(111.33 - 100) = 11.33$ % in the year 2015 as compared to the year 2014.

Illustration 3 : The data about sugar production of a sugar manufacturing company from the year 2008 to 2015 are as follows. Prepare index number by fixed base method from these data by taking average production of the years 2009, 2010 and 2011 as the production of the base year.

Year	2008	2009	2010	2011	2012	2013	2014	2015
Production (thousand tons)	186	196	202	214	229	216	226	230

The average production of the years 2009, 2010 and 2011 $= \dfrac{196 + 202 + 214}{3} = \dfrac{612}{3} = 204$

Year	**Production (thousand tons)**	**Index number by fixed base method** $= \dfrac{p_1}{p_0} \times 100$
2008	186	$\frac{186}{204} \times 100 = 91.18$
2009	196	$\frac{196}{204} \times 100 = 96.08$
2010	202	$\frac{202}{204} \times 100 = 99.02$
2011	214	$\frac{214}{204} \times 100 = 104.90$
2012	220	$\frac{220}{204} \times 100 = 107.84$
2013	216	$\frac{216}{204} \times 100 = 105.88$
2014	226	$\frac{226}{204} \times 100 = 110.78$
2015	230	$\frac{230}{204} \times 100 = 112.75$

Merits and limitations of fixed base method

Merits : (1) Uniformity is maintained in calculation and comparison of the relative changes in the values of the variable as the base year is constant in this method.

(2) This method is useful to compare the long term changes in the values of the variable.

(3) This method is easy to understand and compute.

Limitations : (1) The taste, habits and fashion of consumers change with time and hence there is a change in the items used by the consumer. The items with reduced usage which were used in the past can not be removed in this method.

(2) It is not always possible to have a standard year with normal conditions as the base year. Therefore, selection of the base year is difficult.

The reliability of the index number reduces if the base year is not selected appropriately.

(3) This method is not suitable to compare the short term changes in the value of the variable.

(4) The quality of selected items keeps changing. It is not possible to make necessary change in their weights in this method.

(5) If the base year is a year of very remote past, the comparison can not be considered to be appropriate.

1.4.2 Chain Base Method

A fixed year or period is not taken as a base year or period in this method. For every current year, its preceding year is taken as a base year. For example, the year 2015 is taken as a base year for the index number of the year 2016. The base year keeps changing in this method. Since the base year is repeatedly changed, this method is called chain base method. The current situation is compared with the recent past situation in this method. The index number by this method is found using the following formula :

$$\text{Index number} = \frac{\text{Value of the variable for current year (period)}}{\text{Value of the variable for preceding year (period)}} \times 100$$

$$\therefore \quad I = \frac{p_1}{p_0} \times 100$$

Illustration 4 : The data about bi-monthly closing prices of shares of a company in the year 2014 are given. Compute the chain base index numbers from these data.

Month	January	March	May	July	September	November
Price (₹)	22	21.20	22	23	24.70	26.00

The price for the month before January is not given here. Hence, we will take the index number for January, 2014 as 100. The calculation of index numbers for remaining months using the chain base method are shown in the following table.

Month	Price of share (₹)	Chain base index number $= \frac{\text{Value of the variable in current month}}{\text{Value of the variable in preceding month}} \times 100$
January	22.00	= 100
March	21.20	$\frac{21.20}{22.00} \times 100$ = 96.36
May	22.00	$\frac{22.00}{21.20} \times 100$ = 103.77
July	23.00	$\frac{23.00}{22.00} \times 100$ = 104.55
September	24.70	$\frac{24.70}{23.00} \times 100$ = 107.39
November	26.00	$\frac{26.00}{24.70} \times 100$ = 105.26

Merits and limitations of chain base method

Merits : (1) The problem of selecting the base year does not arise in this method because at any given time the preceding year (period) is taken as the base year (period).

(2) As the comparison is with the preceding year, new items can be included according to the taste and choice of the consumers. It is possible to remove the items not in use.

(3) This method is useful in the fields of economics, trade and commerce as the value of the variable in the current period is compared with the period in the recent past.

Limitations : (1) This method is suitable only for short term comparison of the value of the variable in the current period as the preceding year is taken as the base year. The method is not very convenient for long term comparison.

(2) If there is an error in the calculation of index number by this method then the effect of that error continues in the interpretation of the index number of the succeeding year.

(3) There is no uniformity in the computation of index numbers obtained by this method.

(4) If the information for a year is not available then the index number for the next year can not be obtained.

Illustration 5 : The data about the purchase of groundnut by an edible oil mill from the year 2008 to 2015 are as follows. Prepare the index numbers by fixed base method with the year 2008 as the base year, with chain base and by taking the average quantity purchased in the year 2010 and 2011 as the purchase for the base year.

Year	2008	2009	2010	2011	2012	2013	2014	2015
Purchase of groundnut (ton)	230	250	230	250	270	280	300	300

Year	Quantity Purchase of groundnut (ton)	Index number with base year 2008 $= \frac{\text{Value of variable in current year}}{\text{Value of variable in base year}} \times 100$	Chain base Index number $= \frac{\text{Value of variable in current year}}{\text{Value of variable in preceding year}} \times 100$	Index number by taking average of quantity in year 2010 and 2011 $= \frac{230 + 250}{2} = 240$ as base year quantity
2008	230	$= 100$	$= 100$	$\frac{230}{240} \times 100 = 95.83$
2009	250	$\frac{250}{230} \times 100 = 108.70$	$\frac{250}{230} \times 100 = 108.70$	$\frac{250}{240} \times 100 = 104.17$
2010	230	$\frac{230}{230} \times 100 = 100$	$\frac{230}{250} \times 100 = 92.00$	$\frac{230}{240} \times 100 = 95.83$
2011	250	$\frac{250}{230} \times 100 = 108.70$	$\frac{250}{230} \times 100 = 108.70$	$\frac{250}{240} \times 100 = 104.17$
2012	270	$\frac{270}{230} \times 100 = 117.39$	$\frac{270}{250} \times 100 = 108$	$\frac{270}{240} \times 100 = 112.5$
2013	280	$\frac{280}{230} \times 100 = 121.74$	$\frac{280}{270} \times 100 = 103.70$	$\frac{280}{240} \times 100 = 116.67$
2014	300	$\frac{300}{230} \times 100 = 130.43$	$\frac{300}{280} \times 100 = 107.14$	$\frac{300}{240} \times 100 = 125$
2015	300	$\frac{300}{230} \times 100 = 130.43$	$\frac{300}{300} \times 100 = 100$	$\frac{300}{240} \times 100 = 125$

Illustration 6 : The data about sale of three grain flour wheat, bajri and chana at a flour mill from the year 2011 to 2015 are as follows. Compute the general index number using simple average with (i) fixed base method (taking base year 2011) and (ii) Chain base method.

Year → / Grain flour	Sale (lakh ₹)				
	2011	2012	2013	2014	2015
Wheat flour	40	46	50	56	64
Bajri flour	20	30	36	42	54
Chana flour	50	64	80	96	112

(i) Fixed base method

$$\text{Fixed base index number } I = \frac{\text{Value of variable in current year (period)}}{\text{Value of variable in base year (period)}} \times 100$$

Year / Grain flour	2011	2012	2013	2014	2015
Wheat flour	100	$\frac{46}{40} \times 100 = 115$	$\frac{50}{40} \times 100 = 125$	$\frac{56}{40} \times 100 = 140$	$\frac{64}{40} \times 100 = 160$
Bajri flour	100	$\frac{30}{20} \times 100 = 150$	$\frac{36}{20} \times 100 = 180$	$\frac{42}{20} \times 100 = 210$	$\frac{54}{20} \times 100 = 270$
Chana flour	100	$\frac{64}{50} \times 100 = 128$	$\frac{80}{50} \times 100 = 160$	$\frac{96}{50} \times 100 = 192$	$\frac{112}{50} \times 100 = 224$
Total	300	393	465	542	654
General index number of sale $= \frac{\text{Total}}{3}$	$\frac{300}{3}$ $= 100$	$\frac{393}{3}$ $= 131$	$\frac{465}{3}$ $= 155$	$\frac{542}{3}$ $= 180.67$	$\frac{654}{3}$ $= 218$

(ii) General index number by chain base method :

$$\text{Chain base index number } I = \frac{\text{Value of variable in current year (period)}}{\text{Value of variable in preceding year (period)}} \times 100$$

Year / Grain flour	2011	2012	2013	2014	2015
Wheat flour	100	$\frac{46}{40} \times 100 = 115$	$\frac{50}{46} \times 100 = 108.70$	$\frac{56}{50} \times 100 = 112$	$\frac{64}{56} \times 100 = 114.29$
Bajri flour	100	$\frac{30}{20} \times 100 = 150$	$\frac{36}{30} \times 100 = 120$	$\frac{42}{36} \times 100 = 116.67$	$\frac{54}{42} \times 100 = 128.57$
Chana flour	100	$\frac{64}{50} \times 100 = 128$	$\frac{80}{64} \times 100 = 125$	$\frac{96}{80} \times 100 = 120$	$\frac{112}{96} \times 100 = 116.67$
Total	300	393	353.7	348.67	359.53
Aggregate index number $= \frac{\text{Total}}{3}$	$\frac{300}{3}$ $= 100$	$\frac{393}{3}$ $= 131$	$\frac{353.7}{3}$ $= 117.90$	$\frac{348.67}{3}$ $= 116.22$	$\frac{359.53}{3}$ $= 119.84$

Illustration 7 : The following data are available about the crimes in a city. Find the general index number by fixed base method considering the year 2010 as base year.

Type of Crime \ Year	2007	2008	2009	2010
Murder	110	128	134	129
Violence and rape	30	45	40	48
Robbery	610	720	770	830
Theft of property	2450	2630	2910	2890

$$\text{Fixed base index number } I = \frac{\text{Value of variable in current year (period)}}{\text{Value of variable in base year (period)}} \times 100$$

Type of Crime \ Year	2007	2008	2009	2010
Murder	$\frac{110}{129} \times 100 = 85.27$	$\frac{128}{129} \times 100 = 99.22$	$\frac{134}{129} \times 100 = 103.88$	$\frac{129}{129} \times 100 = 100$
Violence and rape	$\frac{30}{48} \times 100 = 62.5$	$\frac{45}{48} \times 100 = 93.75$	$\frac{40}{48} \times 100 = 83.33$	$\frac{48}{48} \times 100 = 100$
Robbery	$\frac{610}{830} \times 100 = 73.49$	$\frac{720}{830} \times 100 = 86.75$	$\frac{770}{830} \times 100 = 92.77$	$\frac{830}{830} \times 100 = 100$
Theft of property	$\frac{2450}{2890} \times 100 = 84.78$	$\frac{2630}{2890} \times 100 = 91.00$	$\frac{2910}{2890} \times 100 = 100.69$	$\frac{2890}{2890} \times 100 = 100$
Total	306.04	370.72	380.67	400
General Index number of crime $= \frac{\text{Total}}{4}$	$\frac{306.04}{4} = 76.51$	$\frac{370.72}{4} = 92.68$	$\frac{380.67}{4} = 95.17$	$\frac{400}{4} = 100$

EXERCISE 1.1

1. The data about average daily wage of a group of workers employed in a factory in a city during the year 2008 to 2015 are as follows. Find the index number by (1) Fixed base method (taking base year 2008) (2) Chain base method (3) Fixed base method by taking average of average daily wages of the years 2011 to 2013 as the wage for the base year.

Year	2008	2009	2010	2011	2012	2013	2014	2015
Average daily wage (₹)	275	284	289	293	297	313	328	345

2. From the following data about the retail prices of sugar in a city, find the index numbers of price of sugar by (1) Fixed base method with year 2008 as base year (2) Chain base method (3) taking the average price of sugar for the year 2009 and 2010 as the base year price.

Year	2008	2009	2010	2011	2012	2013	2014	2015
Price of Sugar per kilogram (₹)	28	28.50	29.50	30	31	32	34	36

3. The following data are obtained about the annual average prices of wheat, rice and sugar in the wholesale market of a city. Find the general index number for three items by fixed base method with base year 2011 and by chain base method.

Item \ **Year**	2011	2012	2013	2014	2015
Wheat	18	18.50	18.90	19	19.50
Rice	30	36	38	38	39
Sugar	30	31	32	34	36

4. The prices of five fuel related items in the years 2012 and 2014 are as follows. Calculate the general index number for five fuel items by taking the year 2012 as the base year and state the overall increase in the prices of fuel items.

Item	Electricity	Gas	Match Box	Kerosene	Wood
Unit	Unit	Cylinder	Box	Litre	Kilogram
Price in 2012 (₹)	3	345	1.00	15	12
Price in 2014 (₹)	3.5	370	1.50	20	15

*

1.5 Conversion from Fixed Base to Chain Base and from Chain Base to Fixed Base

Generally, whenever the fixed base or chain base index numbers only are available instead of the original information about the values of the variable, the conversion of base is necessary for the following reasons. If the need arises to find the short term changes in the values of the variable then it becomes difficult to find it from the fixed base index numbers. It is easier to find the short term variations after converting the given fixed base index numbers into chain base index numbers.

Sometimes, it is necessary to compare the value of the variable at a given period to the value of another period in a series of values of the variable. This is not possible if only chain base index numbers are available. The above comparison is possible in this situation if the chain base index numbers are converted to the fixed base index numbers. Thus, it is necessary to convert the chain base index numbers into the fixed base index numbers. Hence, the base conversion is carried out as follows :

Conversion of the fixed base index number to the chain base index numbers : The formula for the conversion of the fixed base index numbers into the chain base index numbers is as follows.

$$\text{Chain base index number} = \frac{\text{Fixed base index number of current year}}{\text{Fixed base index number of preceding year}} \times 100$$

Note : If the base year is not mentioned then we will take the chain base index number for the first year as 100. If the base year is mentioned then the fixed base index number of the first year will be taken as its chain base index number.

Illustration 8 : Convert the following index numbers obtained by fixed base method about the production of craft industry of a state into the chain base index numbers.

Year	2009	2010	2011	2012	2013	2014
Fixed base index numbers	120	132	96	144	138	108

Since the base year is not mentioned here, we will take 100 as the chain base index number for the first year.

$$\text{Chain base index number} = \frac{\text{Fixed base index number of current year}}{\text{Fixed base index number of preceding year}} \times 100$$

Year	**Index number**	**Chain base index number**		
2009	120		=	100
2010	132	$\frac{132}{120} \times 100$	=	110
2011	96	$\frac{96}{132} \times 100$	=	72.73
2012	144	$\frac{144}{96} \times 100$	=	150
2013	138	$\frac{138}{144} \times 100$	=	95.83
2014	108	$\frac{108}{138} \times 100$	=	78.26

Illustration 9 : The wholesale price index numbers for commodities with the base year 2007-08 are as follows. Compute the chain base index numbers.

Year	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16
Wholesale price index number	126	130.8	143.3	156.1	167.6	177.6	181.2	177.2

The base year 2007-08 is mentioned here. Hence, we will take the given fixed base index number for the year 2008-09 as chain base index number. Thus, the chain base index number for the first year is 126.

$$\text{Chain base index number} = \frac{\text{Fixed base index number of the current year}}{\text{Fixed base index number of the preceding year}} \times 100$$

Year	Wholesale price index number of commodities	Chain base index number
2008-09	126	= 126
2009-10	130.8	$\frac{130.8}{126} \times 100$ = 103.81
2010-11	143.3	$\frac{143.3}{130.8} \times 100$ = 109.56
2011-12	156.1	$\frac{156.1}{143.3} \times 100$ = 108.93
2012-13	167.6	$\frac{167.6}{156.1} \times 100$ = 107.37
2013-14	177.6	$\frac{177.6}{167.6} \times 100$ = 105.97
2014-15	181.2	$\frac{181.2}{177.6} \times 100$ = 102.03
2015-16	177.2	$\frac{177.6}{181.2} \times 100$ = 97.79

Conversion of chain base index numbers to fixed base index number : If the year-wise chain base index numbers are given, the fixed base index number can be found accordingly. To obtain the fixed base index numbers, the chain base index number of that year is multiplied by the fixed base index number of the previous year and the product is divided by 100.

$$\text{Thus, Fixed base index number of current year} = \frac{\left(\begin{array}{c}\text{Chain base index number}\\ \text{of the current year}\end{array}\right) \times \left(\begin{array}{c}\text{Fixed base index number of the}\\ \text{preceding year to current year}\end{array}\right)}{100}$$

Let us understand this method with an illustration.

Illustration 10 : The chain base index numbers obtained for food items from the year 2008-09 to 2015-16 are as follows. Compute the fixed base index numbers. (Take 2007-08 as base year)

Year	2008–09	2009–10	2010–11	2011–12	2012–13	2013–14	2014–15	2015–16
Index number of food items	134.8	115.28	115.57	107.29	109.91	112.80	106.24	102.48

The year 2007-08 is to be taken as the base year here. Hence, the fixed base index number for the year 2008-09 will not change.

$$\text{Current year fixed base index number} = \frac{\left(\begin{matrix}\text{Current year chain base}\\ \text{index number}\end{matrix}\right) \times \left(\begin{matrix}\text{Fixed base index number of}\\ \text{preceding year to current year}\end{matrix}\right)}{100}$$

Year	Index number of food items	Fixed base index number
2008-09	134.8	= 134.8
2009-10	115.28	$\frac{115.28 \times 134.8}{100}$ = 155.40
2010-11	115.57	$\frac{115.57 \times 155.40}{100}$ = 179.60
2011-12	107.29	$\frac{107.29 \times 179.60}{100}$ = 192.69
2012-13	109.91	$\frac{109.91 \times 192.69}{100}$ = 211.79
2013-14	112.80	$\frac{112.80 \times 211.79}{100}$ = 238.9
2014-15	106.24	$\frac{106.24 \times 238.9}{100}$ = 253.81
2015-16	102.48	$\frac{102.48 \times 253.81}{100}$ = 260.10

EXERCISE 1.2

1. The chain base index numbers of agricultural production of a state from the year 2008 to 2014 are as follows. Compute the fixed base index numbers. (Take 2007 as base year.)

Year	2008	2009	2010	2011	2012	2013	2014
Index number of agricultural production	100	110	95	108	120	106	110

2. Obtain the chain base index number from the fixed base index numbers given below with the year 2007-08 as the base year for the wholesale prices of machines and equipments.

Year	2008 – 09	2009 – 10	2010 – 11	2011 – 12	2012 – 13	2013 – 14	2014 – 15
Index number of machines and equipments	117.4	118	121.3	125.1	128.4	131.6	134.6

3. The fixed base index numbers of food from the month of January to October in the year 2015 for the industrial workers of Ahmedabad are as given below. Compute the chain base index numbers.

Month	January	February	March	April	May	June	July	August	September	October
Index number of food	271	270	268	268	278	283	283	293	293	299

4. The chain base index numbers for sales of a certain type of scooter from the year 2010 to 2015 are as follows. Find fixed base index numbers.

Year	2010	2011	2012	2013	2014	2015
Index number of sale	110	112	109	108	105	111

*

1.6 Specific Formulae for Computing Index Number

We have seen that the index number is useful to study the changes in the values of variable for an item or the values of variables for items in a group. Simple average is used in the construction of an index number and every item is given equal weightage. But the importance of every item is generally not same in practice. For example, the importance given to grains is not same as the importance given to vegetables, pulses or edible oil. Thus, the index number of food will be more realistic and meaningful if each item is assigned weight according to its importance.

The weights of items included are decided in the construction of different types of index numbers. Generally, the weights given to the items for constructing the index number are determined on the basis of their quantity consumed. We shall study some specific formulae for computing index numbers by taking this fact into consideration where different methods of selecting weights are taken for the construction of index number.

Method of weighted average : Suppose I_i is the index number of the ith group among the groups of items (or items) wheat, rice and pulses with the corresponding weight W_i, then the general index number of these groups is obtained using the following formula.

General index number $I = \frac{\Sigma I_i W_i}{\Sigma W_i} = \frac{\Sigma IW}{\Sigma W}$

Note : We shall ignore the suffix 'i' for the simplicity of calculation.

For example, if the index numbers of these groups are 120, 150 and 300 respectively and their corresponding weights are 3, 2 and 1 then the general index number for the group of items is

$$I = \frac{\Sigma IW}{\Sigma W}$$

$$= \frac{120 \times 3 + 150 \times 2 + 300 \times 1}{3 + 2 + 1}$$

$$= \frac{360 + 300 + 300}{6}$$

$$= \frac{960}{6}$$

$$= 160$$

Laspeyre's Formula

This method of finding the index number is given by Laspeyere. It is one of the important methods of finding index number. In this method, base year price is denoted by p_0 and the quantity is denoted by q_0 whereas the prices of items in the current year are denoted by p_1. The expenditure $p_0 q_0$ is assigned as weight to the price relative $\frac{p_1}{p_0}$. The formula of weighted index number thus obtained is called the formula of **Laspeyre's index**, which is denoted by I_L. The Laspeyre's formula is as follows :

Laspeyre's index number $I_L = \frac{\Sigma\left[\frac{p_1}{p_0}\right] \times p_0 q_0}{\Sigma p_0 q_0} \times 100$

$$= \frac{\Sigma \frac{p_1}{p_0} \times p_0 q_0}{\Sigma p_0 q_0} \times 100$$

$$\therefore \quad I_L = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

Paasche's Formula

This method is given by an economist named Paasche. If we denote p_0 as base year price, p_1 as current year price and q_1 as current year quantity then the expenditure $p_0 q_1$ is assigned as weight for the price relative $\frac{p_1}{p_0}$. The formula of weighted index number thus obtained is called the formula of **Paasche's index number**, which is denoted by I_P. The formula of Paasche's index number is as follows :

Paasche's index number $I_P = \frac{\Sigma\left[\frac{p_1}{p_0}\right] \times p_0 q_1}{\Sigma p_0 q_1} \times 100$

$$= \frac{\Sigma \frac{p_1}{p_0} \times p_0 q_1}{\Sigma p_0 q_1} \times 100$$

$$\therefore \quad I_P = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

Fisher's Formula

The base year and current year quantities are taken into condsideration for computing the weight in Laspeyre's and Paasche's method respectively. Prof. Irving Fisher has constructed an index number by considering quantities of both the years. The geometric mean of Laspeyre's and Paasche's index numbers is called **Fisher's index number**, which is denoted by I_F. The formula of Fisher's index number is as follows :

Fisher's index number $I_F = \sqrt{I_L \times I_P}$ or

Fisher's index number $I_F = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$

The Fisher's index number is called ideal index number due to the following reasons :

(1) The quantities of both the years, base year and current year, are taken in the computation for constructing this index number.

(2) This index number satisfies both the important fundamental tests, time reversal and factor reversal tests, of index numbers.

(3) The geometric mean is used to calculate this index number which is the best average for the construction of index number.

(4) This index number is free from bias as it balances the demerits of Laspeyre's and Paasche's index number.

Thus, Fisher's index number is an ideal index number.

Illustration 11 : Find the index number for the year 2016 with base year 2011 by weighted average method from the following data of price and weights of five different items.

Item	Weight	Price (₹)	
		Year 2011	Year 2016
A	40	160	200
B	25	400	600
C	5	50	70
D	20	10	18
E	10	2	3

The weights of different items are given here. We shall compute the general index number from the price relatives of the year 2016 based on the prices of the year 2011.

Item	Weight	Price (₹)		$I = \frac{p_1}{p_0} \times 100$	IW
	W	p_0	p_1		
A	40	160	200	$\frac{200}{160} \times 100 = 125$	5000
B	25	400	600	$\frac{600}{400} \times 100 = 150$	3750
C	5	50	70	$\frac{70}{50} \times 100 = 140$	700
D	20	10	18	$\frac{18}{10} \times 100 = 180$	3600
E	10	2	3	$\frac{3}{2} \times 100 = 150$	1500
Total	**100**				**14,550**

Index number of year 2016 $I = \frac{\Sigma IW}{\Sigma W}$

$= \frac{14550}{100}$

$= 145.50$

Thus, we say that there is an increase of (145.50 – 100) = 45.5 % in prices in the year 2016 as compared to the year 2011.

Illustration 12 : Find Laspeyre's, Paasche's and Fisher's index numbers for the year 2016 with base year 2015 from the data about price and consumption of food items given below.

Item	Unit	Year 2016		Year 2015	
		Price (₹)	Quantity	Price (₹)	Quantity
Rice	Kilogram	40	1.5 Kilogram	39	1 Kilogram
Milk	Litre	44	10 Litre	40	12 Litre
Bread	Kilogram	50	1.5 Kilogram	45	2 Kilogram
Banana	Dozen	36	1.5 Dozen	30	2 Dozen

We will take price p_0 and quantity q_0 for the base year, price p_1 and quantity q_1 for the current year.

Item	Unit	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
Rice	Kilogram	39	1	40	1.5	40	39	60	58.5
Milk	Litre	40	12	44	10	528	480	440	400
Bread	Kilogram	45	2	50	1.5	100	90	75	67.5
Banana	Dozen	30	2	36	1.5	72	60	54	45
Total						**740**	**669**	**629**	**571**

Laspeyre's index number $I_L = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$

$= \frac{740}{669} \times 100$

$= 110.6128$

$\simeq 110.61$

Thus, there is a rise of $(110.61 - 100) = 10.61$ % in prices of the year 2016 as compared to the base year 2015.

Paasche's index number $I_P = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100$

$= \frac{629}{571} \times 100$

$= 110.1576$

$\simeq 110.16$

Thus, there is a rise of $(110.16 - 100) = 10.16$ % in prices of the year 2016 as compared to the base year 2015.

Fisher's index number $I_F = \sqrt{I_L \times I_P}$

$= \sqrt{110.61 \times 110.16}$

$= 110.3847$

$\simeq 110.38$

Thus, there is a rise of (110.38 – 100) = 10.38 % in prices of the year 2016 as compared to the base year 2015.

Illustration 13 : Compute Laspeyre's, Paasche's and Fisher's index numbers for the year 2016 from the data given below by taking 2015 as the base year.

Item	Unit	Price (₹)		Quantity (Consumption)	
		Year 2015	Year 2016	Year 2015	Year 2016
A	20 Kilogram	300	440	5 Kilogram	8 Kilogram
B	Quintal	500	700	10 Kilogram	15 Kilogram
C	Kilogram	60	75	1200 Gram	2000 Gram
D	Meter	14.25	15	15 Meter	25 Meter
E	Litre	32	36	18 Litre	30 Litre
F	Dozen	30	36	8 Pieces	10 Pieces

The base year is 2015 and the current year is 2016. Hence, we will take price p_0 and quantity q_0 for the year 2015, price p_1 and quantity q_1 for the year 2016.

The price of item *A* is per 20 kg here whereas the unit for quantity is kg. The price of item *B* is per quintal but the unit for the quantity is kg. The price of item *C* is per kg whereas the unit for quantity is gram. The price for item *F* is per dozen whereas the unit for quantity is piece. The calculation of the price per item of these four items will be as follows :

The price of item *A* in the year 2015 is ₹ 300 per 20 kg. Hence, its price = $\frac{300}{20}$ = ₹ 15 per kg.

Similarly, the price of item *A* in 2016 is $\frac{440}{20}$ = ₹ 22 per kg.

It is convenient to express the price of item *B* per kg than quintal. Hence, the price for the year 2015 = $\frac{500}{100}$ = ₹ 5 per kg and the price for the year 2016 = $\frac{700}{100}$ = ₹ 7 per kg.

The price of item *C* is per kg. Hence, it is convenient to express its quantity in kg.

Thus, the quantity in the year 2015 = $\frac{1200}{1000}$ = 1.2 kg and the quantity for the year 2016 = $\frac{2000}{1000}$ = 2 kg.

The price of item *F* is per dozen which is convenient to express in per piece. Hence, the price of the year 2015 = $\frac{30}{12}$ = ₹ 2.5 per piece and the price for the year 2016 = $\frac{36}{12}$ = ₹ 3 per piece. Now, the index number will be calculated as follows :

Item	Unit	Year 2015		Year 2016		p_1q_0	p_0q_0	p_1q_1	p_0q_1
		p_0	q_0	p_1	q_1				
A	kg	15	5	22	8	110	75	176	120
B	kg	5	10	7	15	70	50	105	75
C	kg	60	1.2	75	2	90	72	150	120
D	Meter	14.25	15	15	25	225	213.75	375	356.25
E	Litre	32	18	36	30	648	576	1080	960
F	Piece	2.5	8	3	10	24	20	30	25
Total						**1167**	**1006.75**	**1916**	**1656.25**

Laspeyre's index number $I_L = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

$= \frac{1167}{1006.75} \times 100$

$= 115.9175$

$\simeq 115.92$

Thus, we can say that there is a rise of (115.92 – 100) = 15.92 % in the prices in the year 2016 as compared to the year 2015.

Paasche's index number $I_P = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$

$= \frac{1916}{1656.25} \times 100$

$= 115.6830$

$\simeq 115.68$

Thus, it can be said that there is (115.68 – 100) = 15.68 % rise in the prices in the year 2016 as compared to the year 2015.

Fisher's index number $I_F = \sqrt{I_L \times I_P}$

$= \sqrt{115.92 \times 115.68}$

$= 115.7999$

$\simeq 115.8$

Thus, it can be said that there is (115.8 – 100) = 15.8 % rise in the prices in the year 2016 as compared to the year 2015.

Illustration 14 : Find the ideal index number for the year 2015 from the following data.

Item	Base year 2014		Current year 2015	
	Price (₹)	Quantity	Price (₹)	Quantity
A	16	10	20	11
B	20	9	24	9
C	32	16	40	17

Fisher's index number is considered as an ideal index number. So, we will find Fisher's index number here. We will take price p_0 and quantity q_0 for base year, price p_1 and quantity q_1 for the current year.

Item	p_0	q_0	p_1	q_1	p_1q_0	p_0q_0	p_1q_1	p_0q_1
A	16	10	20	11	200	160	220	176
B	20	9	24	9	216	180	216	180
C	32	16	40	17	640	512	680	544
Total					**1056**	**852**	**1116**	**900**

Fisher's index number $I_F = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$

$= \sqrt{\frac{1056}{852} \times \frac{1116}{900}} \times 100$

$= \sqrt{1{\cdot}5369} \times 100$

$= 1.2397 \times 100$

$I_F = 123.97$

Thus, it can be said that there is (123.97 – 100) = 23.97 % rise in the prices in the year 2015 as compared to the year 2014.

Illustration 15 : Find Fisher's index number for the year 2015 by taking the year 2014 as the base year from the data given below about consumption and total expenditure of five different items.

Item	Base Year 2014		Current Year 2015	
	Consumption	Total expenditure	Consumption	Total expenditure
A	50 kg	2500	60 kg	4200
B	120 kg	600	140 kg	700
C	30 litre	330	20 litre	200
D	20 kg	360	15 kg	300
E	5 kg	40	5 kg	50

The consumption and total expenditure for the items are given here.

Total expenditure of item = (Price of item per unit) × (Consumption of item)

$\therefore$ Price of item per unit = $\frac{\text{Total expenditure of item}}{\text{Consumption of item}}$

We will obtain the price per unit of each item using the above formula.

Item	Base year 2014		Current year 2015		p_1q_0	p_0q_0	p_1q_1	p_0q_1
	Quantity	$p_0 = \frac{\text{Expenditure}}{q_0}$	Quantity	$p_1 = \frac{\text{Expenditure}}{q_1}$				
	q_0	p_0	q_1	p_1				
A	50	$\frac{2500}{50} = 50$	60	$\frac{4200}{60} = 70$	3500	2500	4200	3000
B	120	$\frac{600}{120} = 5$	140	$\frac{700}{140} = 5$	600	600	700	700
C	30	$\frac{330}{30} = 11$	20	$\frac{200}{20} = 10$	300	330	200	220
D	20	$\frac{360}{20} = 18$	15	$\frac{300}{15} = 20$	400	360	300	270
E	5	$\frac{40}{5} = 8$	5	$\frac{50}{5} = 10$	50	40	50	40
Total					**4850**	**3830**	**5450**	**4230**

Fisher's index number $I_F = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$

$= \sqrt{\frac{4850}{3830} \times \frac{5450}{4230}} \times 100$

$= \sqrt{1{\cdot}6315} \times 100$

$= 1.2773 \times 100$

$I_F \simeq 127.73$

Thus, it can be said that there is (127.73 – 100) = 27.73 % rise in the prices in the year 2015 as compared to the year 2014.

Illustration 16 : The health department has implemented a certain policy for the industrial units in the year 2003 to control the possibility of cancer due to the chemical process which is hazardous to the health of workers employed in the industrial units of a certain industrial area who are residing in the same area. To evaluate this policy, a survey was conducted about deaths due to cancer of persons in the different age groups. The following data are obtained for the years 2003 and 2008. Find the index number of deaths due to cancer using weighted average method by taking the population of this industrial area in the year 2003 as weight and interpret it.

Age-group (years)	Population in year 2003 (thousand)	Deaths in year 2003	Deaths in year 2008
< 5	10	200	65
5-15	8	145	100
15-40	48	610	480
40-60	38	350	225
> 60	14	550	465

We shall obtain the general index number by finding the relative percentages of cancer deaths for the year 2008 and taking the population in different age-groups in 2003 as weights.

Age-group (years)	Population in year 2003 (thousand) W	Deaths in year 2003 p_0	Deaths in year 2008 p_1	$I = \frac{p_1}{p_0} \times 100$	IW
< 5	10	200	65	$\frac{65}{200} \times 100 = 32.5$	325
5-15	8	145	100	$\frac{100}{145} \times 100 = 68.97$	551.76
15-40	48	610	480	$\frac{480}{610} \times 100 = 78.69$	3777.12
40-60	38	350	225	$\frac{225}{350} \times 100 = 64.29$	2443.02
> 60	14	550	465	$\frac{465}{550} \times 100 = 84.55$	1183.7
Total	**118**				**8280.6**

Index number for year 2008 $I = \frac{\Sigma IW}{\Sigma W}$

$= \frac{8280.6}{118}$

$= 70.1745$

$\simeq 70.17$

Thus, it can be said that there is a decrease of (100 – 70.17) = 29.83 % in the deaths due to cancer in the year 2008 as compared to the year 2003.

EXERCISE 1.3

1. The information about six different items used in the production of an electronics item is as follows. Find the index number and interpret it.

Items	*A*	*B*	*C*	*D*	*E*	*F*
Weight	5	10	10	30	20	25
Percentage price relative	290	315	280	300	315	320

2. The information about six different items used in the furniture items is as follows. Find the index number for the year 2015 with the base year 2014 and interpret it.

Item	*A*	*B*	*C*	*D*	*E*	*F*
Weight	17	15	22	16	12	18
Price in year 2014 (₹)	30	20	50	32	40	16
Price in year 2015 (₹)	24	24	70	40	48	24

3. Find the Laspeyre's, Paasche's and Fisher's index numbers for the year 2015 with the base year 2014 using the following information.

Item		Wheat	Rice	Pulses	Oil	Cloth	Kerosene
	Unit	kg	kg	kg	kg	Meter	Litre
Year 2014	**Quantity**	20	10	10	6	15	18
	Price (₹)	15	20	26.50	24.80	21.25	21
Year 2015	**Quantity**	30	15	15	8	25	30
	Price (₹)	18	31.25	29.50	30	25	28.80

4. Find the Laspeyre's, Paasche's and Fisher's index numbers for the year 2015 with the base year 2014 using the following information.

Item	**Unit**	**Price (₹)**		**Quantity (Consumption)**	
		Year 2014	**Year 2015**	**Year 2014**	**Year 2015**
A	20 kg	80	120	5 kg	7 kg
B	kg	20	24	2400 gm	4000 gm
C	Quintal	2000	2800	10 kg	15 kg
D	Dozen	48	72	30 pieces	35 pieces

5. Find the ideal index number from the following data for the year 2015.

Item	Unit	Base year 2014		Base year 2015	
		Price (₹)	Quantity	Price (₹)	Quantity
A	20 kg	120	10 kg	280	15 kg
B	5 Dozen	120	3 Dozen	140	48 pieces
C	kg	4	5000 gm	8	4 kg
D	5 Litre	52	15 Litre	58	20 Litre

6. Find the Paasche's and Fisher's index numbers for the year 2015 with the base year 2014 using the data given below.

Item		*A*	*B*	*C*	*D*	*E*
Year 2014	Price (₹)	100	100	150	180	250
	Total expenditure	400	500	600	1080	1000
Year 2015	Price (₹)	120	120	160	200	300
	Total expenditure	720	600	800	1000	1200

*

1.7 Cost of Living Index Number

The cost of living index number is constructed to measure and study the changes in the cost of living of people from different sections of the society due to the fluctuations in prices. Thus, "The number showing the percentage of relative changes in the cost of living of the people of a certain section of the society in the current year (period) as compared to the base year (period) is called the **cost of living index number**."

The cost of living index number is prepared separately for the people of different sections of the society and regions.

For example, a family spends ₹ 15,000 per month for their living in the year 2012 and the same family spends ₹ 18,000 per month for their living in the year 2014 for the same lifestyle. Their cost of living index number can be obtained as follows :

$$\text{Cost of living index number} = \frac{\text{Current year (period) monthly expenditure}}{\text{Base year (period) monthly expenditure}} \times 100$$

$$= \frac{18000}{15000} \times 100$$

$$= \frac{600}{5}$$

$$= 120$$

Thus, it can be said that there is a rise of (120 – 100) = 20 % in the monthly expense in the year 2014 as compared to the year 2012.

1.7.1 Construction of Cost of Living Index Number

The following points should be considered while constructing the cost of living index number :

(1) Purpose : The purpose of every index number should be explained before constructing it. We should ascertain the class of people in the society for whom the cost of living index number is to be constructed. The requirements of the people of worker class and rich class are different. For example, the rise in price of grains does not affect much to the cost of living of the people from rich class whereas it affects a lot to the cost of living of the people from worker class. Thus, it is necessary to clarify the purpose of the construction of the cost of living index number.

(2) Family Budget Inquiry : A sample of some families is randomly selected from the families of that class of people for whom the cost of living index is to be prepared. The budget of the families selected in the sample is studied. The information is obtained about the list of different items consumed by them, their consumption, list of retail prices, the expenses incurred on them and place of purchase, etc. This type of inquiry is called sample family budget inquiry.

The data obtained from the inquiry of families included in the sample are generally divided into five sections : (a) food (b) clothing (c) house rent (d) fuel and electricity and (e) miscellaneous.

The importance of different items in the expenditure for living can be known from the sample inquiry of family budget. Hence, the importance of each item selected in the construction of the index number in its group and the importance of each group in the total expenditure can be determined.

(3) Availability of Prices of Items : The retail prices of items are collected from the areas of residence of the people from the class of families for whom the cost of living index number is to be obtained. As far as possible, these prices should be collected from the standard or government approved shops. The average price of items should be taken into consideration when the prices obtained from various shops at different times are different.

(4) Base Year : A normal year is selected as a base year. The price relatives are found as follows for each item by taking the retail prices of the normal year as the base year prices :

$$\text{Price relative} \quad I = \frac{p_1}{p_0} \times 100$$

Where, p_1 = retail price of item in current year

p_0 = retail price of item in base year

(5) Average : It is necessary to find a general price relative from the price relatives of different items. A proper average should be used for this purpose. Theoretically, the geometric mean is the ideal average for the construction of index number. But due to the difficulty of its computation, it is common to use the weighted mean for the construction of index number.

(6) Weight : The importance of different items selected in the construction of index number is not same. The number associated with items in propotion to their importance is called weight. These weights can be of two types : (i) Implicit Weight and (ii) Explicit Weight.

(i) Implict Weight : This is an indirect method of assigning weights. According to this method, the weights are determined as per the number of varieties of different items selected in the construction of index number. This method is called implicit method as the weights can not be accurately quantified.

(ii) Explicit Weight : This is a direct method of assigning weights. The weights of items are expressed numerically in proportion to their importance. In this method, the weights of items are determined according to the consumption, sale, production or the expenditure for that item. Thus, the weights given in accordance with the importance of the items are called explicit weights.

The two methods of assigning explicit weight are as follows :

(1) Method of total expenditure (2) Method of family budget

(1) Method of Total Expenditure : In this method, the expenditure for every item in the base year and current year is found using the consumption of these items and further the total expenditure of all the items is obtained for both the years. The percentage ratio of the total expenditure of the current year with the total expenditure of the base year is called the index number by the method of total expenditure.

Suppose p_0 = price of base year, q_0 = quantity of base year

p_1 = price of current year, q_1 = quantity of current year

If the quantity of the base year is used for finding the total expenditure of the current and base year,

Σp_1q_0 = total expenditure of current year and Σp_0q_0 = total expenditure of base year.

Index number $I = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$. This formula is the formula of Laspeyre's index number.

If the quantity of the current year is used for finding the total expenditure of the current and base year,

Σp_1q_1 = total expenditure of current year and Σp_0q_1 = total expenditure of base year

Index number $I = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100$. This formula is the formula of Paasche's index number.

(1) Method of Family Budget : In this method, the percentage price relative I is found first for every item. Here, $I = \frac{p_1}{p_0} \times 100$ where, p_1 = price of current year and p_0 = price of base year. Then, the expenditure of every item p_0q_0 in the base year is found and it is taken as the weight W for the percentage price relative I. The formula for index number by the method of family budget using the weighted average with weight $W = p_0q_0$ is as follows :

Index number $I = \frac{\Sigma IW}{\Sigma W}$

$$= \frac{\Sigma\left[\frac{p_1}{p_0} \times 100 \times p_0q_0\right]}{\Sigma p_0q_0}$$

$$= \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$$

Thus, the index number obtained by the method of family budget is the Laspeyre's index number.

Illustration 17 : Find the cost of living index number by the family budget method from the following information about index numbers of different groups of items for living and their weights.

Group	Food items	Clothing	Electricity-fuel	House rent	Miscellaneous
Index number	281	177	178	210	242
Weight	46	10	7	12	25

The index numbers of different groups and their weights are given here. Hence, we will use family budget method which is a method of weighted average.

Group	**Index number I**	**Weight W**	IW
Food Items	281	46	12,926
Clothing	177	10	1770
Electricity-fuel	178	7	1246
House Rent	210	12	2520
Miscellaneous	242	25	6050
Total		**100**	**24,512**

$$\text{Index number } I = \frac{\Sigma IW}{\Sigma W}$$

$$= \frac{24512}{100}$$

$$= 245.12$$

Thus, it can be said that there is a rise of (245.12 – 100) = 145.12 % in the total expenditure in the current year as compared to the base year.

Illustration 18 : Calculate the cost of living index number by the total expenditure method and the family budget method for the year 2015 with the base year 2014 using the following data.

Item	Wheat	Rice	Tuver Dal	Oil	Cloth	Kerosene
Unit	Quintal	kg	kg	litre	meter	litre
Quantity of year 2014	35 kg	25 kg	20 kg	10 litre	20 meter	15 litre
Price of year 2014 (₹)	1600	40	60	80	30	28
Price of year 2015 (₹)	1800	45	120	90	45	35

The base year is 2014. We will take p_0 = price of 2014, q_0 = quantity of 2014 and p_1 = price of 2015. We shall make uniform units for the price and quanity of each item.

Method of Total Expenditure

Item	Unit	Year 2014		Year 2015	p_1q_0	p_0q_0
		q_0	p_0	p_1		
Wheat	kg	35	16	18	630	560
Rice	kg	25	40	45	1125	1000
Tuver Dal	kg	20	60	120	2400	1200
Oil	litre	10	80	90	900	800
Cloth	meter	20	30	45	900	600
Kerosene	litre	15	28	35	525	420
Total					**6480**	**4580**

Index number by total expenditure method $= \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$

$= \frac{6480}{4580} \times 100$

$= 141.4847$

$\simeq 141.48$

Thus, there is a rise of (141.48 – 100) = 41.48 % in the total expenditure in the year 2015 as compared to the base year 2014.

Method of Family Budget

Item	Unit	Year 2014		Year 2015	$I = \frac{p_1}{p_0} \times 100$	$W = p_0q_0$	IW
		q_0	p_0	p_1			
Wheat	kg	35	16	18	$\frac{18}{16} \times 100 = 112.5$	560	63,000
Rice	kg	25	40	45	$\frac{45}{40} \times 100 = 112.5$	1000	1,12,500
Tuver dal	kg	20	60	120	$\frac{120}{60} \times 100 = 200$	1200	2,40,000
Oil	litre	10	80	90	$\frac{90}{80} \times 100 = 112.5$	800	90,000
Cloth	meter	20	30	45	$\frac{45}{30} \times 100 = 150$	600	90,000
Kerosene	litre	15	28	35	$\frac{35}{28} \times 100 = 125$	420	52,500
Total						**4580**	**6,48,000**

Index number by family budget method $= \frac{\Sigma IW}{\Sigma W}$

$= \frac{648000}{4580}$

$= 141.4847$

$\simeq 141.48$

Note : We can see here that the index numbers obtained by total expenditure method and family budget method are same.

Illustration 19 : The data referring to worker class of a city are as follows. Find the general index numbers for the years 2014 and 2015. If the wages of these workers in 2014 are increased by 5 % in the year 2015, is this rise in wages sufficient to maintain their standard of living ?

Group	Food	Clothing	Fuel and Electricity	House Rent	Miscellaneous
Weight	48	18	8	12	14
Group index number of 2014	210	220	210	200	210
Group index number of 2015	230	225	220	200	235

Group	**Weight** W	**Group index number of year 2014** I_1	**Group index number of year 2015** I_2	I_1W	I_2W
Food	48	210	230	10,080	11,040
Clothing	18	220	225	3960	4050
Fuel and Electricity	8	210	220	1680	1760
House Rent	12	200	200	2400	2400
Miscellaneous	14	210	235	2940	3290
Total	**100**			**21,060**	**22,540**

Index number for year 2014 $= \frac{\Sigma I_1W}{\Sigma W} = \frac{21060}{100} = 210.60$

Index number for year 2015 $= \frac{\Sigma I_2W}{\Sigma W} = \frac{22540}{100} = 225.40$

There is a rise of $(225.4 - 210.6) = 14.8$ % in the cost of living of workers in the year 2015 than in the year 2014 with reference to the base year.

Thus, the percentage increase in the cost of living index number in the year 2015 is $\frac{14.8}{210.6} \times 100 = 7.03$ as compared to the year 2014. Hence, the rise of 5% in the wages of the year 2014 is not sufficient to maintain the same standard of living of the workers in the year 2015.

Illustration 20 : The following data are obtained from the budget inquiry of middle class families. State the change in the cost of living in the current year 2015 with respect to the base year 2014 by finding the index number. If the average monthly disposable income of a family is ₹ 30,000 during the year 2014 and their average monthly disposable income during the year 2015 is ₹ 35,000 then according to family budget index number, what should be the rise in the average monthly disposable income of the family to maintain the same standard of living of the base year ?

Group	Food	Clothing	Rent	Fuel	Miscellaneous
Weight	45	20	15	10	10
Percentage price relative of the group in year 2015	130	150	120	160	120

The weights W and the percentage price relatives I of the groups for the year 2015 are given here. Hence, we shall calculate the index number by family budget method.

Group	Food	Clothing	Rent	Fuel	Miscellaneous	**Total**
Percentage price relative *I*	130	150	120	160	120	
Weight *W*	45	20	15	10	10	**100**
IW	5850	3000	1800	1600	1200	**13,450**

Index number by family budget method $I = \frac{\Sigma IW}{\Sigma W}$

$= \frac{13450}{100}$

$= 134.5$

Thus, it can be said that there is a rise of (134.5 – 100) = 34.5 % in the cost of living in the year 2015 as compared to the year 2014.

According to the family budget method index number of the year 2015, the average monthly disposable income to maintain the same standard of living as the base year

$= \frac{\text{Index number of current year}}{\text{Index number of base year}} \times \text{income of base year}$

$= \frac{134.50}{100} \times 30,000$

$= ₹\ 40,350$

The required increase in the average monthly disposable income to maintain the standard of living of the family = ₹ 40,350 – ₹ 35,000 = ₹ 5350

1.7.2 Uses and limitations of cost of living index number

This index number is prepared by studying the changes in the cost of living of the people of different classes. Thus, the cost of living index number is used for different objectives as follows :

(1) The changes in the purchasing power of money of a class of people can be measured using the corresponding cost of living index number. If the increase in the price of item is higher than the increase in the income, there is a decrease in the real income of the earning members and subsequently their purchasing power decreases. Thus, the cost of living index number is useful to find the actual purchasing power of money and the real income. The purchasing power of money and real income can be obtained by the following formulae.

(i) Purchasing power of money $= \frac{1}{\text{Cost of living index number}} \times 100$

(ii) Real income $= \frac{\text{Income}}{\text{Cost of living index number}} \times 100$

(2) The cost of living index number for each class shows the real economic condition of the respective class. Hence, this measure is used to suggest the changes in the wage, dearness allowance, bonus, etc. paid to the people of that class.

(3) This index number measures the effect of retail prices on the cost of living of people. Hence, this index number guides the government regarding the items to be controlled under special acts and the items to be kept for free trade.

(4) The cost of living index number helps the government as an indicator to frame the tax policies, policies regarding issues like price-regulation and fare-regulation. Moreover, it is possible to know how the living of people of different classes gets affected by imposing tax on certain items and the tax policy can be planned accordingly.

(5) The government agencies and public institutions use it as a base to determine the necessity of special facilities to elevate the standard of living of people of different classes.

Limitations of cost of living index number :

(1) It is not possible to construct one common cost of living index number for all sections of the society.

(2) The cost of living index number obtained for a certain class of people in a certain region cannot be used for some other region even for the same class of people.

(3) The cost of living index number shows the average percentage changes in the cost of living of a certain class. Thus, it is not possible to measure the changes in the cost of living of an individual.

(4) It is necessary to construct separate index numbers for different classes of people as well as different regions.

(5) The expenditure of the people of any class depends upon the size of the family, life style, liking, habits, etc. There is no uniformity in the expenditure of all the families of the same class.

(6) Its calculation is based on the assumption that the life style of the families does not change in the current year as compared to the base year. In reality, the liking, habits and choices of people change with time. Thus, it is necessary to conduct family budget inquiry at regular intervals of time and change the items and their weights.

EXERCISE 1.4

1. The following data are obtained from the family budget inquiry of middle class people. State the change in the cost of living in the year 2015 with respect to the year 2013 by finding the index number. If the average monthly disposable income of a family in the year 2013 is ₹ 15,000 then obtain the estimate of the necessary average monthly disposable income in the year 2015.

Group	Food	Fuel-Electricity	Rent	Clothing	Miscellaneous
Weight	45	15	10	20	10
Expenditure in 2013 (₹)	3000	1450	1500	600	1600
Expenditure in 2015 (₹)	3900	1850	2400	900	1920

2. Find the index number for the year 2014 by the method of family budget from the following data about prices and consumption of food items and interpret it.

Item	**Year 2010**		**Year 2014**
	Quantity	**Price (₹)**	**Price (₹)**
Wheat	60	15	18
Rice	40	32	40
Bajri	15	12	14
Tuver Dal	25	50	70

3. Compute the cost of living index number by the method of total expenditure from the following data.

Item	*A*	*B*	*C*	*D*	*E*
Unit	Quintal	20 kg	10 litre	dozen	meter
Quantity of year 2014	50 kg	18 kg	12 litre	20 pieces	14 meter
Price of year 2014 (₹)	1200	340	30	15	12
Price of year 2015 (₹)	1700	380	40	24	16

4. Compute the general index number for the production using the following data.

Item	Cotton Cloth	Grains	Sugar	Steel	Copper	Cement
Weight	15	23	15	25	10	12
Index number of production	220	225	190	215	198	220

5. The details of expenditure on clothing for the worker class of a region are as follows. Find the index number for clothing by the total expenditure and family budget method.

Item	Saree	Dhoti	Shirting	Other
Unit	Piece	Piece	Meter	Meter
Quantity in year 2010	5	8	20	15
Price in year 2010 (₹)	300	70	32.40	20.90
Price in year 2014 (₹)	400	100	38	23.80

*

Typical examples :

Illustration 21 : The prices of three items *A*, *B* and *C* among five items have increased in the year 2015 by 90 %, 120 % and 70 % respectively with respect to the year 2010, whereas the prices of two items *D* and *E* have decreased by 2 % and 5 % respectively. Item *A* is four times important than item *B* and item *C* is six times important than item *A*. The importance of items *D* and *E* is two and half times the importance of item *B*. Compute the general price index number of the year 2015 for all the five items.

The percentage increase and decrease in the prices of items is given here. Similarly, the weight W of the items are the numbers showing their relative importance.

Suppose the relative importance of item B is 1.

Then the importance of item A will be 4, importance of C will be 24 and that of D and E will be 2.5 each.

The general price index number will be calculated as follows :

Item	**Percentage increase (+) decrease (–)**	**Index number** I = (100 + increase) = (100 – decrease)	**Weight** W	IW
A	+ 90	100 + 90 = 190	4	760
B	+ 120	100 + 120 = 220	1	220
C	+ 70	100 + 70 = 170	24	4080
D	– 2	100 – 2 = 98	2.5	245
E	– 5	100 – 5 = 95	2.5	237.5
Total			**34**	**5542.5**

$$\text{General price index number} = \frac{\Sigma IW}{\Sigma W}$$

$$= \frac{5542.5}{34}$$

$$= 163.0147$$

$$\simeq 163.01$$

Thus, it can be said that there is a rise of (163.01 – 100) = 63.01 % in the prices in the current year 2015 as compared to the base year 2014.

Illustration 22 : The details of expenses on fuel for a group of workers of a region are as follows.

Item	Base year 2012		Year 2014
	Quantity	Price per unit (₹)	Price per unit (₹)
Coal	5 Kilogram	25	30
Kerosene	20 Litre	40	45
Wood	5 Kilogram	22	25
Match-box	10 Boxes	0.90	1

Prepare the index number of the group of fuel expenditure from these data. If the expenditure for food, clothing, house rent and miscellaneous groups in the year 2015 are 3, 2.5, 4.5 and 3.25 times respectively that of the year 2012, and if the expenditures on these groups are 42 %, 15 %, 10 % and 12 % respectively of the total expenditure then prepare the cost of living index number for the workers.

First of all, we shall prepare the group index number for fuel-expenditure from its details. We will take the base year 2012 and obtain the index number by total expenditure method.

Note : The method of family budget can also be used here for calculation.

Item	Year 2012		Year 2014	p_1q_0	p_0q_0
	q_0	p_0	p_1		
Coal	5	25	30	150	125
Kerosene	20	40	45	900	800
Wood	5	22	25	125	110
Match box	10	0.90	1	10	9
Total				**1185**	**1044**

Index number for fuel expenditure $= \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$

$= \frac{1185}{1044} \times 100$

$= 113.5057$

$\simeq 113.51$

The expenditure for the groups of food, clothing, house rent and miscellaneous are 3, 2.5, 4.5 and 3.25 times respectively than the base year. Hence, the index numbers of these four groups are (3 × 100) = 300; (2.5 × 100) = 250; (4.5 × 100) = 450 and (3.25 × 100) = 325 respectively. The index number of fuel category is obtained as 113.51. We will take the percentage expenditure for all the five groups as the weights for their corresponding index numbers to find the cost of living index number.

The expenditures for the groups of food, clothing, house rent and miscellaneous are given here as 42 %, 15 %, 10 % and 12 % respectively. These will be taken as their respective weights W. Total expenditure is 100 %. Hence, the weight for fuel expenditure index number will be 100 – (42 + 15 + 10 + 12) = 21 %.

The calculation of cost of living index number is as follows.

Group	Food	Clothing	House rent	Miscellaneous	Fuel	Total
Index number I	300	250	450	325	113.5	
Weight W	42	15	10	12	21	**100**
IW	12,600	3750	4500	3900	2383.71	**27,133.71**

$$\text{Cost of living index number} = \frac{\Sigma IW}{\Sigma W}$$

$$= \frac{27133.71}{100}$$

$$= 271.3371$$

$$\simeq 271.34$$

Thus, it can be said that there is a rise of (271.34 – 100) = 171.34 % in the cost of living in the current year as compared to the base year.

Illustration 23 : Find the real wages for the worker class of a city from the following data about their average monthly wage and the cost of living index number (base year 2001). Find the purchasing power of money in the year 2015 by taking the base year 2001 and state the importance of this answer.

Year	2010	2011	2012	2013	2014	2015
Average monthly wage (₹)	15,000	15,600	16,200	17,000	18, 000	20,000
Cost of living index number	192	203	228	268	270	287

The calculation of real wage using the wages and cost of living index numbers will be as follows.

$$\text{Real wage} = \frac{\text{Average monthly wage}}{\text{Cost of living index number}} \times 100$$

Year	Average monthly wage (₹)	Cost of living index number	Real wage (₹)
2010	15,000	192	$\frac{15000}{192} \times 100 = 7812.5$
2011	15,600	203	$\frac{15600}{203} \times 100 = 7684.73$
2012	16,200	228	$\frac{16200}{228} \times 100 = 7105.26$
2013	17,000	268	$\frac{17000}{268} \times 100 = 6343.28$
2014	18,000	270	$\frac{18000}{270} \times 100 = 6666.67$
2015	20,000	287	$\frac{20000}{287} \times 100 = 6968.64$

The purchasing power of money is the reciprocal of the cost of living index number of the current year with the respective base year.

$\therefore$ We can say that the purchasing power of money in the year 2015 with the base year 2001 $= \frac{100}{287} = 0.3484 \simeq 0.35$. Hence, it can be said that if the unit of money is rupee then the value of rupee in the year 2005 is 35 paise with respect to the base year 2001.

Thus, although the actual average monthly wage of the workers of this class in the year 2015 is more than the base year 2001, the real disposable wage in the year 2015 is only ₹ 6968.64 with respect to the base year.

Illustration 24 : Answer the following questions :

(1) The price of wheat was ₹ 1600 per quintal in the year 2014 and it was ₹ 1800 per quintal in the year 2015. Find the price index number of wheat for the year 2015 with the base year 2014 and interpret it.

Price index number of wheat for year 2015 $I = \frac{p_1}{p_0} \times 100$

$= \frac{1800}{1600} \times 100$

$= 112.5$

Thus, it can be said that there is a rise of (112.5 – 100) = 12.5 % in the price of wheat per quintal in the year 2015 as compared to the year 2014.

(2) The Laspeyre's index number is $\frac{8}{9}$ times the Fisher's index number. If the Fisher's index number is 180, find the Paasche's index number.

The Laspeyre's index number is $\frac{8}{9}$ times the Fisher's index number.

$\therefore \quad I_L = \frac{8}{9} \times I_F$

$$\therefore \quad I_L = \frac{8}{9} \times 180$$

$$I_L = 160$$

Now, $I_F = \sqrt{I_L \times I_P}$

$$180 = \sqrt{160 \times I_P}$$

$$(180)^2 = 160 \times I_P$$

$$\therefore \quad I_P = \frac{180 \times 180}{160} = 202.5$$

(3) The production of an item in the year 2015 has increased by 3 times the production in the base year. Find the index number of production for the year 2015.

Consider the index number of the base year as 100. The production has increased by 3 times in the year 2015.

Production index number of year 2015 = index number of base year + increase in index number in current year

$= 100 + (3 \times 100)$

$= 100 + 300 = 400$

(4) If $\Sigma p_1q_0 : \Sigma p_0q_0 = 3:2$ and $\Sigma p_1q_1 : \Sigma p_0q_1 = 5:3$, find I_L, I_P and I_F.

$$\frac{\Sigma p_1q_0}{\Sigma p_0q_0} = \frac{3}{2}$$

$$I_L = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100$$

$$= \frac{3}{2} \times 100 = 150$$

$$\frac{\Sigma p_1q_1}{\Sigma p_0q_1} = \frac{5}{3}$$

$$I_P = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100$$

$$= \frac{5}{3} \times 100 = 166.67$$

$$I_F = \sqrt{I_L \times I_P} = \sqrt{150 \times 166.67} = \sqrt{25000.5} = 158.12$$

(5) If the average monthly disposable income of middle class families in the year 2014 is ₹ 14,400 and if the cost of living index number for the year 2015 with the base year 2014 is 115 then estimate the average monthly disposable income of these families in the year 2015.

The cost of living index number of the middle class families for the year 2015 is 115 with the base year 2014. Hence, the index number has increased by $(115 - 100) = 15\%$ as compared to the base year. Thus, there should be a 15 % rise in the average disposable income of the middle class families.

$\therefore$ Average monthly disposable income of the families $= 14400 + (14400 \times \frac{15}{100})$

$= 14400 + 2160 = 16560$

Hence, the average monthly disposable income of these families in the year 2015 should be ₹ 16,560.

(6) If the cost of living index number of the current year has increased to 180 from the base year index number 100 and if the average income of workers has increased from ₹ 6000 to ₹ 9000, is there an increase or decrease in the purchasing power of the workers ? How much is it ?

The index number has increased to 180 from 100 here which means that there is an increase of 80 %. Hence, the income should also increase by 80 %.

Average income $= 6000 + (6000 \times \frac{80}{100})$

$= 6000 + 4800 = ₹\ 10{,}800$

Hence, the average income of workers should be ₹ 10,800. But the average income of workers has increased to ₹ 9000. Thus, there is a decrease of (10800 – 9000) = ₹ 1800 in the average income with reference to the index number. Hence, it can be said that there is a decrease in their purchasing power.

(7) The wholesale price index numbers of the year 2015 and 2016 are found to be 150.2 and 165.7 respectively. Find the rate of inflation using index numbers of both the years.

The index number of the year 2015 is 150.2 and the index number of current year 2016 is 165.7.

We will use the following formula of the rate of inflation.

$$\text{Rate of inflation} = \frac{\left(\begin{matrix}\text{Wholesale price index}\\ \text{number of current year}\end{matrix}\right) - \left(\begin{matrix}\text{Wholesale price index}\\ \text{number of previous year}\end{matrix}\right)}{\text{Wholesale price index number of previous year}} \times 100$$

$$= \frac{165.7 - 150.2}{150.2} \times 100$$

$$= \frac{15.5}{150.2} \times 100$$

$$= 10.3196$$

$$\simeq 10.32$$

Thus, rate of inflation is 10.32 %.

(8) If the increase in the price relatives of three items are 250 %, 265 % and 300 % respectively and if the ratio of the importance of these items is 8 : 7 : 5, find the general price index number.

The percentage increase in the index numbers (price relatives) I and the relative importance W are given here.

We will calculate the index number.

Item	**Index Number (I)** **(Index Number of base year + increase)**	**Weight** W	IW
A	100 + 250 = 350	8	2800
B	100 + 265 = 365	7	2555
C	100 + 300 = 400	5	2000
Total		**20**	**7355**

General index number $= \frac{\Sigma IW}{\Sigma W} = \frac{7355}{20} = 367.75$

General index number = 367.75

Thus, there is a rise of (367.75 – 100) = 267.75 % in the prices in the current year as compared to the base year.

Summary

- The price, production, demand, supply, quantity, etc. of an item are called the variable for that item.
- The changes taking place in the values of the variable at two different time periods are compared by two methods : (1) Method of absolute measure (difference) and (2) Method of relative measure (ratio)
- The ratio of changes in the values of the variable at two different time periods is called relative change.
- The measure showing the percentage relative change in the prices of an item at different time periods is called price index number.
- The measure showing the percentage relative change in the quantities of an item at different time periods is called quantity index number.
- The average of the percentage change in the values of a variable associated with one or more items for the given period compared to its value in the fixed (base) period is called general index number for the group.
- When the price of an item is compared with the price of the same item in some specific (fixed) year of the past, then that specific year is called the base year.
- The year for which the price of an item is to be compared with the price of the base year is called the current year.
- Two methods of selecting base year : (1) Fixed base method (2) Chain base method.
- The expenditure p_0q_0 is assigned as weight to the price relative $\frac{p_1}{p_0}$ of the items. The formula of weighted average obtained by this method is called the formula of Laspeyre's index number.
- The expenditure p_0q_1 is assigned as weight to the price relative $\frac{p_1}{p_0}$ of the items. The formula of weighted average obtained by this method is called the formula of Paasche's index number.
- The geometric mean of Laspeyre's and Paasche's index numbers is called the Fisher's index number.
- The number showing the percentage of relative changes in the cost of living of the people of a certain section of the society in the current year (period) as compared to the base year (period) is called the cost of living index number.
- The points for the construction of index numbers : purpose, family budget inquiry, availability of prices of items, choice of base year, choice of average and choice of weight.
- In the construction of index number, the number associated with the selected items in proportion to their importance is called weight of that item.
- There are two types of weights : (i) Implicit weights (ii) Explicit weights
- Implicit weights : The weights are included in the selection of items and they cannot be expressed numerically. This indirect method of assigning weight is called implicit weight.
- Explicit weights : The weight to be assigned are determined in proportion to the importance of the item and can be expressed numerically. Such a weight is called explicit weight.
- There are two popular methods of assigning explicit weight : (i) Method of total expenditure (ii) Method of family budget.

List of Formulae

(1) Price relative $= \dfrac{\text{Price of current year (period)}}{\text{Price of base year (period)}}$

$= \dfrac{p_1}{p_0}$

(2) Index number $I = \dfrac{\text{Value of variable in current year (period)}}{\text{Value of variable in base year (period)}} \times 100$

$I = \dfrac{p_1}{p_0} \times 100$

(3) Index number based on price relatives of n items $= \dfrac{\Sigma\left[\dfrac{p_1}{p_0}\right]}{n} \times 100$

(4) Fixed base index number $= \dfrac{\text{Value of variable in current year (period)}}{\text{Value of variable in base year (period)}} \times 100$

$I = \dfrac{p_1}{p_0} \times 100$

(5) Chain base index number $= \dfrac{\text{Value of variable in current year (period)}}{\text{Value of variable in preceding year (period)}} \times 100$

$I = \dfrac{p_1}{p_0} \times 100$

(6) Conversion of fixed base index number into chain base index number :

Chain base index number $= \dfrac{\text{Fixed base index number of current year}}{\text{Fixed base index number of preceding year}} \times 100$

(7) Conversion of chain base index number into fixed base index number :

Fixed base index number $= \dfrac{\text{(Chain base index number of current year)} \times \text{(Fixed base index number of preceding year)}}{100}$

(8) Laspeyre's index number $I_L = \dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$

(9) Paasche's index number $I_P = \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$

(10) Fisher's index number $I_F = \sqrt{I_L \times I_P}$ or

$$I_F = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

(11) Cost of living index number :

[1] Method of total expenditure :

When base year quantity (q_0) is given,

Index number $= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$ (**Note :** This is Laspeyre's index number)

When current year quantity (q_1) is given,

Index number $= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$ (**Note :** This is Paasche's index number)

[2] Method of family budget (weighted average of relative values) :

Index number $= \frac{\Sigma IW}{\Sigma W}$ Where $I = \frac{p_1}{p_0} \times 100$

$W = p_0 q_0$

(12) Purchasing power of money $= \frac{1}{\text{Cost of living index number}} \times 100$

(13) Real wage $= \frac{\text{Wage}}{\text{Cost of living index number}} \times 100$

Exercise 1

Section A

Find the correct option for the following multiple choice questions :

1. Which method is useful to compare the long term variations in the values of the variable ?

(a) Chain base method (b) Laspeyre's method

(c) Fixed base method (d) Paasche's method

2. Which consumption is used in the calculation of Laspeyre's index number ?

(a) Consumption of base year (b) Consumption of current year

(c) Consumption of average year (d) Consumption of any year

3. Which prices are considered in the construction of the cost of living index number ?

(a) Market price (b) Wholesale price (c) Average price (d) Retail price

4. Which expenditure of items is assigned as weights in the method of family budget ?

(a) Expenditure of selected year (b) Average annual expenditure

(c) Expenditure of base year (d) Expenditure of current year

5. Which average is considered as the best average in construction of the index number ?

(a) Harmonic mean (b) Arithmetic mean (c) Weighted mean (d) Geometric mean

6. Which index number gives idea of the standard of living of people ?

(a) Index number of industrial production (b) Quantity index number

(c) Fisher's index number (d) Cost of living index number

7. The price of an item increased by 4.5 times in the current year as compared to the base year. What will be the price index number ?

(a) 45 (b) 450 (c) 550 (d) 350

8. If the purchasing power of money is 0.75 in the year 2016 with respect to the base year 2015 then what will be the price index number for the year 2016 ?

(a) 750 (b) 175 (c) 133.33 (d) 275

9. If the cost of living index number for the people of a class is 200 for the year 2016 with respect to the year 2015, which of the following statements is true ?

(a) There is an average 200 percent rise in the current year prices of the items consumed by that class.

(b) There is an average 100 percent decrease in the current year prices of the items consumed by that class.

(c) Purchasing power of money is ₹ 0.5.

(d) The current year prices of the items consumed by that class are stable.

10. If $I_P = I_F$, which of the following statements is true ?

(a) $I_P = 2I_L$ (b) $I_F = \frac{I_L}{2}$ (c) $I_F = I_P = I_L$ (d) $4I_F = I_L$

11. If the average disposable income of families of a class is ₹ 20,000 in the year 2013 and if the cost of living index number of that class for the year 2015 with the base year 2013 is 130, what should be the average disposable income of the families of this class in the year 2015 ?

(a) ₹ 26,000 (b) ₹ 20,130 (c) ₹ 20,000 (d) ₹ 14,000

12. What weight is assigned as expenditure to the price relatives $\frac{p_1}{p_0}$ of the items to obtain the formula for Paashe's index number ?

(a) p_0q_0 (b) p_1q_1 (c) p_0q_1 (d) p_1q_0

Section B

Answer the following questions in one sentence :

1. What is a price relative ?
2. Which method is more suitable to compare the changes in a variable at two different time periods ?
3. If the quantity index number of an item for a certain year is 130, interpret it.
4. What is a base year ?
5. Write the formula for the conversion of chain base index number into fixed base index number.
6. Define the index number.
7. Define the cost of living index number.
8. What is weight ?
9. What is implicit weight ?
10. Name the important basic tests of index numbers.
11. What is a chain base index number ?
12. 'The price index number of oil is ₹ 500.' State whether this statement is true or false and if false, correct and rewrite it.
13. Which index number is used to find the rate of inflation ? Write the formula to find the rate of inflation.
14. Which index number is used in India to find the rate of dearness allowance ?
15. State the main difference between fixed base method and chain base method.
16. In which method does the base year change each year ?
17. Which is the appropriate average for the construction of index number ?
18. How should be the base year in the calculation of index number ?

Section C

Answer the following questions :

1. Which points should be considered while choosing it ?
2. State the characteristics of index number.

3. What is quantity index number ?

4. What is weight in the construction of index numbers ? State the types of weight.

5. Why is Fisher's index number called an ideal number ?

6. State the main difference between explicit weight and implicit weight.

7. The cost of living index number increased from 280 to 340 during a certain time period and the wage increased from ₹ 13,500 to ₹ 14,750. Find the real gain or loss of the worker.

8. The cost of living index numbers and average monthly wage from the year 2010 to 2013 are given as follows. Find the real wage for each year.

Year	2010	2011	2012	2013
Average monthly wage (₹)	35,000	40,000	42,000	50,000
Cost of living index number	120	150	130	160

9. The wholesale price index numbers for the year 2014 and 2015 are found to be 177.6 and 181.2 respectively. Find the rate of inflation using index numbers of both the years.

10. The percentage increase in the price realtives of three items are 315, 328 and 390 respectively. If the importance of these items has ratio 5 : 7 : 8, find the general price index number.

11. If the average disposable income of family of a class is ₹ 25,000 in the year 2014 and if the cost of living index number of that class for the year 2016 with the base year 2014 is 120, estimate the average disposable income of the family of that class in the year 2016.

12. The average monthly income of a worker was ₹ 16,000 in the year 2015 and it increased to ₹ 20,000 in the year 2016. Find the index number of income for the year 2016 in comparison to the year 2015.

13. If the production of an item increased by $\frac{9}{5}$ times in the year 2016 as compared to the base year, find the index number of production for the year 2016.

14. If $I_L = 221.5$ and $I_F = 222$, find I_P.

Section D

Answer the following questions :

1. State the merits and limitations of fixed base method.

2. State the merits and limitations of chain base method.

3. Differentiate between fixed base and chain base methods.

4. Give the meaning of cost of living index number and state the points to be considered for its construction.

5. State the uses of cost of living index number.

6. State the limitations of cost of living index number.

7. The prices of three items among five fuel items increased by 50 %, 90 % and 110 % in the year 2015 as compared to the base year 2014. The prices of other two items decreased by 5 % and 2 % respectively. If the ratio of importance of these five items is 5 : 4 : 3 : 2 : 1, find the index number of fuel prices for the year 2015.

8. Find the fixed base index numbers from the following data about average annual income of workers in a company from the year 2008 to the year 2014. (Take base year as 2008.)

Year	2008	2009	2010	2011	2012	2013	2014
Average annual income (₹ 10,000)	36	40	48	52	60	80	95

9. The index numbers of average closing prices of shares of a certain company in different months with the base January 2014 are as follows. Find the chain base index numbers.

Month	January '14	February '14	March '14	April '14	May '14	June '14
Fixed base index number	100	104	105	108	109	127

10. Find the fixed base index numbers from the chain base index numbers given below.

Year	2011	2012	2013	2014
Index Number	120	90	140	125

11. Find the chain base index numbers from the following data regarding the price of an item.

Year	2009	2010	2011	2012	2013	2014
Price ₹	40	45	48	55	60	70

12. Find the cost of living index number from the given information for the month of April, 2015 regarding group index numbers and weights of items of living of industrial workers.

Group	*A*	*B*	*C*	*D*	*E*	*F*
Index Number	247	167	259	196	212	253
Weight	44	20	16	6	10	4

13. If $\Sigma p_1q_0 : \Sigma p_0q_0 = 5:3$ and $\Sigma p_1q_1 : \Sigma p_0q_1 = 3:2$, compute the Laspeyre's, Paasche's and Fisher's index numbers.

14. If the ratio of Laspeyre's and Paasche's index number is 4 : 5 and Fisher's index number is 150, find Paasche's index number.

Section E

Solve the following :

1. Find the general index number using the following data about prices of different items in the year 2012 by taking the base year 2010.

Item	*A*	*B*	*C*	*D*	*E*
Unit	Quintal	Kilogram	Dozen	Meter	Litre
Price of year 2010 (₹)	110	50	40	80	20
Price of year 2012 (₹)	120	70	60	90	20

2. Compute the index number for the year 2015 with the base year 2010 by the method of total expenditure using the following data.

Item	*A*	*B*	*C*	*D*
Price of year 2010 (₹)	10	30	40	20
Price of year 2015 (₹)	14	42	80	26
Quantity in year 2010	8	4	4	16

3. Compute the index number by the method of total expenditure for the year 2014 by taking the base year 2013 using the following data.

Item	**Year 2014**		**Year 2013**
	Consumption (Quantity)	**Price (₹)**	**Price (₹)**
Wheat	15 kg	24	20
Rice	10 kg	45	40
Bajri	5 kg	20	16
Tuver Dal	3 kg	90	80

4. Use the following information to find (i) fixed base index numbers with the year 2008 as the base year (ii) the index numbers by taking the average price of the years 2008 and 2009 as the base year price.

Year	2008	2009	2010	2011	2012	2013	2014
Price (₹)	32	38	40	42	45	60	65

5. The index numbers of different groups of industrial output of a city and the weights of these groups are given below. Find the index number of the industrial production.

Group	**Index Number**	**Weight**
Iron	390.2	30
Textile	247.6	31
Chemical	510.2	18
Engineering goods	403.3	17
Cement	624.4	4

6. The price of wheat increased by 70 % and price of rice increased by 40 % in the year 2015 with respect to the year 2010. The price of bajri decreased by 25 %. The price of oil increased by 40 % and the price of ghee decreased by 5 %. If the importance of oil is three times and of rice is double that of ghee and the importance of each of wheat and bajri is double that of rice, find price index number of the group of these five items and interpret it.

7. Calculate the real wages of a worker class from the following data about their monthly wages. Find the purchasing power of money in the year 2015 considering the year 2008 as the base year.

Year	2010	2011	2012	2013	2014	2015
Average monthly wage (₹)	15,000	18,000	19,000	20,000	22,000	25,000
Cost of living index number (Base year 2008)	120	180	205	220	235	260

Section F

Solve the following :

1. Find the Laspeyre's and Paasche's index numbers using the following data for the year 2015 by taking the year 2014 as the base year. Also find the Fisher's index number and interpret it.

Item	Base year 2014		Current year 2015	
	per unit price (₹)	Total expenditure (₹)	per unit price (₹)	Total expenditure (₹)
Wheat	16	224	18	270
Rice	35	140	40	200
Tuver Dal	100	200	120	360
Oil	108	432	120	600

2. The quantity consumed and total expenditure of four different items are as given below. Find Paasche's and Fisher's index numbers for the year 2015 with respect to the year 2013.

Item	Base year 2013		Current year 2015	
	Total expenditure (₹)	Consumption (Quantity)	Total expenditure (₹)	Consumption (Quantity)
A	360	60 kg	375	25 kg
B	160	10 litre	416	20 litre
C	480	15 kg	613.2	6 kg
D	336	3 kg	400	2.5 kg

3. Compute the Fisher's index number from the data given below about six different items.

Item	*A*	*B*	*C*	*D*	*E*	*F*
Unit	20 kg	Quintal	kg	Litre	Meter	Dozen
Year 2013 Quantity	5 kg	10 kg	1200 gm	30 litre	12 meter	20 pieces
Year 2013 Price (₹)	600	1600	60	52	8	30
Year 2015 Quantity	12 kg	12 kg	2000 gm	36 litre	20 meter	16 pieces
Year 2015 Price (₹)	880	2400	75	32	12	36

4. Compute the Laspeyre's, Paasche's and Fisher's index numbers for the year 2015 from the data given below.

Item	Quantity		Price (₹)	
	Year 2014	Year 2015	Year 2014	Year 2015
A	25 kg	32 kg	42	45
B	15 litre	20 litre	28	30
C	10 pieces	20 pieces	30	36
D	8 meter	15 meter	20	25
E	30 litre	36 litre	60	65

5. Compute the index number for the year 2015 by total expenditure method and family budget method and state whether both the index numbers are same.

Item	Unit	Year 2013	Year 2013	Year 2015
		Consumption (Quantity)	Price (₹)	Price (₹)
Wheat	Quintal	100 kg	1800	2400
Rice	20 kg	40 kg	700	800
Sugar	kg	40 kg	30	36
Oil	kg	60 kg	108	120
Pulses	20 kg	40 kg	2000	2400
Ghee	kg	36 kg	400	480

6. The data about index numbers and weights for groups of items for the living of industrial workers in Ahmedabad city in the years 2014 and 2015 are as follows. Find the cost of living index number of industrial workers. If the wages of workers are increased by 5 % in the year 2015 as compare to year 2014 then is this rise sufficient to compensate the rise in price in the year 2015 ?

Group	Food	Fuel and Electricity	Housing	Clothing	Miscellaneous expense
Weight	31	14	22	10	23
Index number of 2014	270	168	205	174	303
Index number of 2015	281	178	210	177	337

7. The following data are given about the index numbers and weights for the items of living of industrial workers in a city in the year 2014. Find the cost of living index number for industrial workers. If the average monthly salary paid to these workers in the year 2012 was ₹ 6,000, what should be the monthly salary in the current year 2014 to maintain the same standard of living ?

Group	Food	Fuel and Electricity	Housing	Clothing	Miscellaneous expense
Price index of 2014 (Base year 2012)	255	174	234	153	274
Weight	42	8	12	18	20

8. The data about the industrial production quantity and weights for the year 2015 are given below. Compute the index number of industrial production and interpret it.

Industry	Unit	Year 2013 production	Year 2015 production	Weight
Mine	Lakh tons	10	15	4
Textile	Crore meters	20	25	6
Engineering	Lakh tons	30	25	30
Chemicals	Hundred tons	40	50	3
Food	Lakh tons	50	60	4

9. The data about per unit price and weights of four different items in the year 2014 and 2015 are as follows. Compute the index number of the year 2015.

Item	Weight	Year 2014	Year 2015
		Price per unit (₹)	Price per unit (₹)
A	40	32	40
B	25	80	100
C	20	24	30
D	15	4	6

10. The index numbers of food and clothing among the different groups of cost of living are 150 and 224.7 respectively for the year 2015. The price of fuel has increased by 220 %. The expense on rent has increased from ₹ 4000 to ₹ 6000 and miscellaneous expenses increased by 1.75 times. If the expenditure for the first four groups are 40 %, 18 %, 12 % and 20 % respectively, find the cost of living index number for the year 2015 and interpret it.

Irving Fisher
(1867 – 1947)

Irving Fisher was an American economist, statistician, inventor and Progressive social campaigner. He was one of the earliest American neoclassical economists. He was described as "The greatest economist the United States has ever produced.

Fisher made important contributions to utility theory and general equilibrium. He was also a pioneer in the rigorous study of intertemporal choice in markets, which led him to develop a theory of capital and interest rates. His research on the quantity theory of money inaugurated the school of macroeconomic thought known as "Monetarism". Fisher was also a pioneer of econometrics, including the development of index numbers. Some concepts named after him include the Fisher Equation, the Fisher Hypothesis, the international Fisher Effect, the Fisher Separation theorem and the Fisher Market.

"Can we say, in this case, that the cause of a cause is the relevant cause ?"

– Johnny Rich

2

Linear Correlation

Contents :

2.1 Introduction

2.2 Meaning and Definition of Linear Correlation

2.3 Correlation and Coefficient of Correlation

2.4 Scatter Diagram Method

2.5 Karl Pearson's Product Moment Method

2.6 Properties of Coefficient of Correlation

2.7 Interpretation of the value of Correlation Coefficient

2.8 Spearman's Rank Correlation Method

2.9 Precautions in the interpretation of Correlation Coefficient

2.1 Introduction

We have studied the characteristics of only one variable in the chapters of measures of central tendency, dispersion, skewness, etc. in 11th Standard. So far our study was restricted to the distribution of one variable. But many situations arise in which it is desirable and necessary to study two or more variables at a time. e.g. If we study the yearly sale of a product of a company, we would also like to know its profit as from that we can find the nature and degree of relation between these two variables 'sale of product' and 'profit'. The yearly rainfall and yield of rice in some region, price and demand for an item, income and expenditure of a family, height of father and height of his son, age of husband and age of wife, etc. are some well-known examples where we see the relation between paired variables. Similarly, we come across many situations in which two variables are related.

We shall study about the relationship between two mutually related variables in this and the next chapter.

Note :

- A distribution is said to be univariate distribution if it is obtained by collecting the information of one variable. e.g. The distribution of marks obtained by students in a subject, monthly income of persons working in a company, age of drivers of state transport (ST) buses.
- The information obtained for two different characteristics of a unit at the same time is called bivariate data and the distribution obtained from it is known as bivariate distribution. e.g. The distribution of price and supply of an item, monthly expenditure and saving of families of a group, age of husband and age of wife.
- The simultaneous changes in the values of more than two variables can be studied by multiple and partial correlation but here we shall study correlation between only two variables.

2.2 Meaning and Definition of Linear Correlation

Let us first understand the meaning of correlation. We know that in many situations, simultaneous changes are seen in the values of two variables. The simultaneous changes in the values of two variables are mainly due to the following reasons.

(1) There is a cause-effect relation between two variables.

(2) The values of two variables change due to the effect of some other factor.

In case of yearly rainfall and yield of rice of a region, usually if rainfall increases (up to some extent), yield of rice also increases and if rainfall decreases, yield of rice also decreases. So, 'rainfall' is a 'cause' and 'yield of rice' is an 'effect'. Similarly, when the income of a person remains more or less same, if his expenditure increases, saving decreases and if his expenditure decreases, saving increases. 'Expenditure' is a 'cause' here and 'savings' is the 'effect'. The changes in two variables in the above two examples indicate cause-effect relationship. Sometimes both the variables may be mutually dependent and therefore neither can be specifically said as the 'cause' and the other the 'effect'. Generally, it happens in case of economic variables. e.g. demand and supply. If demand increases, it is necessary to increase supply (which is not always possible instantly) and when supply increases, price tend to decrease and because of that demand goes up. Thus, demand and supply are interdependent. The ages of husband and age of wife is also an example of such situation.

In case of sale of raincoats and sale of rain shoes, the values of both the variables increase in monsoon. There is no direct cause-effect relation here between two variables but the changes in the sale of raincoats and rain shoes are observed due to the presence of the third variable, namely the monsoon. This is an example of indirect cause-effect relationship.

After discussing the examples of the relationship between two variables, we can define correlation as follows. If there are simultaneous changes in the values of two variables due to direct or indirect cause-effect then it is said that there is **correlation** between two variables. The correlation is said to be **linear correlation**, if the points plotted on the graph paper corresponding to ordered pairs of the values of two correlated variables are on a line or nearer to the line. In other words, when the changes in the values of two correlated variables have nearly constant proportion then we can say that there is linear correlation.

Generally, exact linear relation exists in natural sciences like Mathematics and Physics. i.e. Because of change in one variable, the value of other variable changes in constant proportion.

e.g (1) Radius and Circumference

We know that if the radius of a circle is r units then its circumference is $2\pi r$. So, when radius is multiplied by 2π (constant), we get its circumference. Thus, it can be said that because of change in radius, the circumference of the circle changes in constant proportion. This is clear from the following table :

Radius (r)	2	3	5	10
Circumference $(2\pi r)$	4π	6π	10π	20π

(2) At a constant speed, time taken by an object to travel the distance and the distance travelled by an object.

If an object is travelling at the speed of 50 km per hour then the distance travelled is in constant proportion to the time taken to travel it. We can understand it very easily from the following table.

Time taken by an object to travel (hours)	2	3	6	10
Distance travelled by an object (km)	100	150	300	500

But generally in commerce, economics and social science, the changes in the values of two variables are not in constant proportion. e.g. : If rainfall in the current year increased by 10 % for two consecutive years as compared to the previous year then it is not necessary that yield also increases in same proportion for both the years. This is because both the variables, rainfall and yield change under the effect of some other factors and hence there is an element of 'chance (uncertainty)' in these changes. So, in such cases, the points of ordered pairs of two correlated variables may not be on the line but they may be almost nearer to the line.

We shall study linear correlation in this chapter. Now, we shall refer linear correlation as only 'correlation'.

There are mainly two types of correlation : (1) Positive Correlation (2) Negative Correlation

(1) Positive Correlation : When the changes in the values of two correlated variables are in the same direction, the correlation between them is said to be positive.

The correlation between price and supply of an item, age of husband and age of wife, number of vehicles and number of accidents on a road, sale and profit of an item, rainfall and yield of crop in a region are some examples of positive correlation.

(2) Negative Correlation : When the changes in the values of two correlated variables are in opposite direction, the correlation between them is said to be negative.

The correlation between price and demand of an item, expenditure and saving of a person, minimum day temperature and sale of woollen clothes in winter, altitude and amount of Oxygen in air are some examples of negative correlation.

2.3 Correlation and Coefficient of correlation

We have already discussed the meaning and definition of correlation in the previous section. Now, let us understand 'correlation coefficient' as its measure.

The measure of the strength of a linear correlation between two variables is called the correlation coefficient. It is denoted by r. The correlation coefficient is a numerical measure of correlation which shows the strength or degree of linear correlation between two correlated variables. This measure was first suggested by Karl Pearson.

Methods of studying correlation

The following methods are mainly used to determine the nature and strength of correlation between two variables :

(1) Scatter Diagram Method

(2) Karl Pearson's Product Moment Method

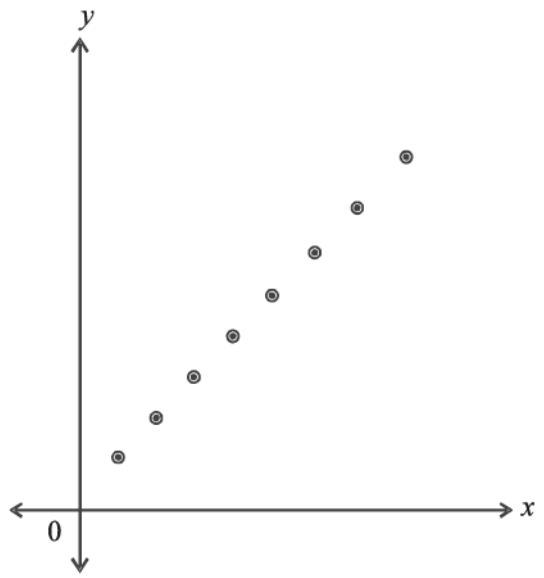
(3) Spearman's Rank Correlation Method

2.4 Scatter Diagram Method

This is a simple method of diagrammatic representation of two related variables to find the nature of correlation. This is a most widely used method for determining the nature of correlation. Moreover, it also gives some idea about the strength of correlation between two variables.

Suppose $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\ldots\ldots, (x_n, y_n)$ are ordered pairs of n values of two variables X and Y. The values of variable X on X-axis and the values of variable Y on Y-axis are plotted on the graph by taking proper scale. The graph showing the plotted points corrosponding to pairs $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$ is called a **Scatter Diagram.**

The pattern of points on a scatter diagram shows the nature or type of correlation and the strength of correlation upto some extent. Now, let us see how the nature and strength of correlation between two variables can be determined by using scatter diagram.



If all the points of a scatter diagram lie on one line and if the line is going in the upward direction from left to right then it shows **perfect positive correlation** between two variables X and Y.

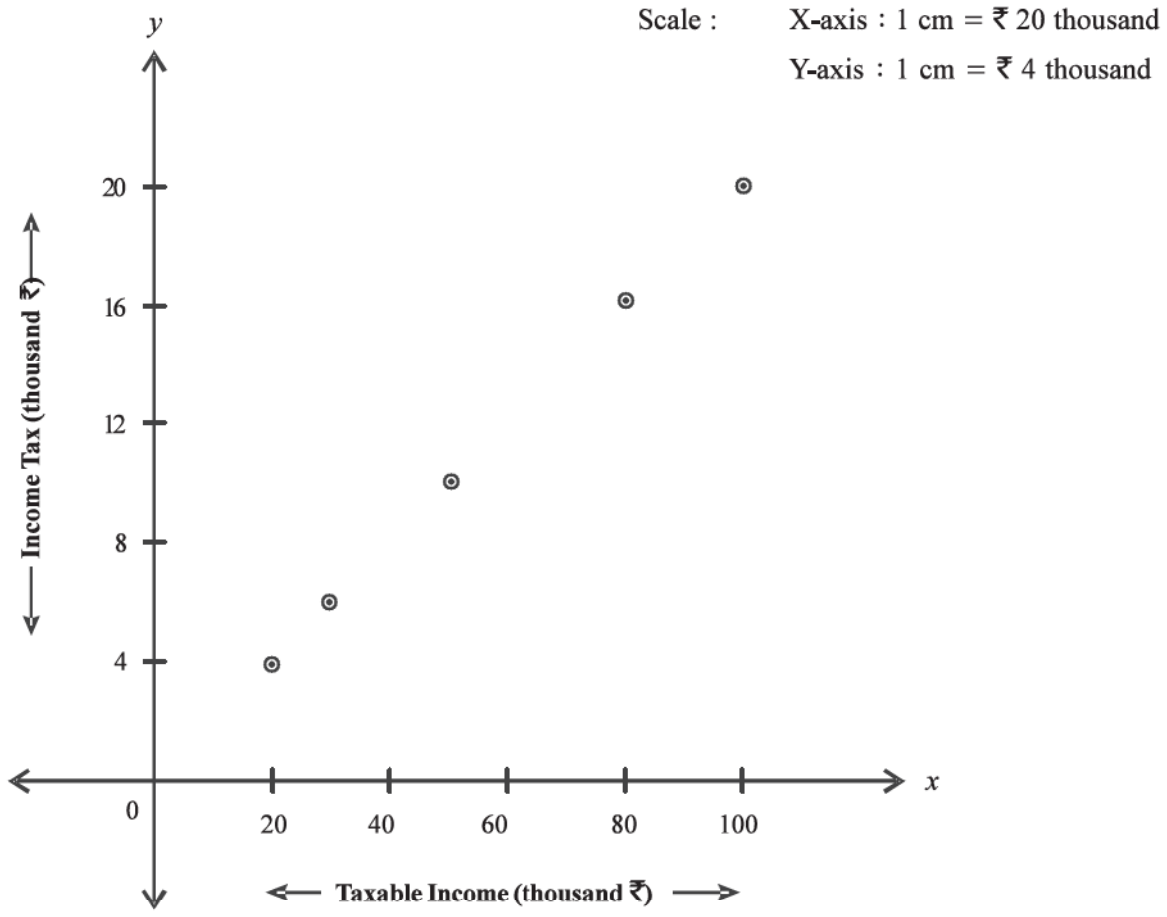
We get such a scatter diagram when the changes in the values of two variables are in the same direction and in the constant proportion. We take one example to understand this case.

Illustration 1 : After standard deduction from total income, 20% income tax is imposed on the remaining income. The information regarding the taxable income and the tax to be paid is given below for five persons.

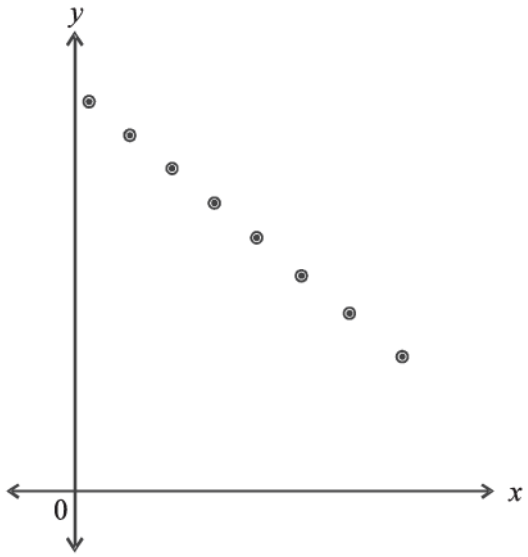
Person	1	2	3	4	5
Taxable Income (thousand ₹) x	50	30	80	20	100
Income Tax (thousand ₹) y	10	6	16	4	20

Draw a scatter diagram from this information and discuss about the correlation.

The following scatter diagram is obtained by plotting the points corresponding to the ordered pairs $(50, 10), (30, 6), (80, 16), (20, 4)$ and $(100, 20)$ of x and y.



We can see that all the points lie on the same line in the scatter diagram. We can also see that as the values of variable X change, the values of variable Y also change in the same direction with a constant proportion. (Check that when the value of variable X is multiplied by 0.2 (20%), the corresponding value of variable Y of the ordered pair is obtained. So, the changes in the two variables X and Y are in same proportion.) Hence, we can say that there is a perfect positive correlation between two variables X and Y.



If all the points of a scatter diagram lie on one line and if the line goes in the downward direction from left to right then it shows **perfect negative correlation** between two variables X and Y.

When the changes in the values of two variables are in opposite direction and in the constant proportion, we get such scatter diagram. We take one example to understand this case.

Illustration 2 : To know the relation between monthly expenditure and monthly savings for middle class families, the information regarding expenditure and savings for 5 families is given below. (The monthly income of each family is ₹ 20,000)

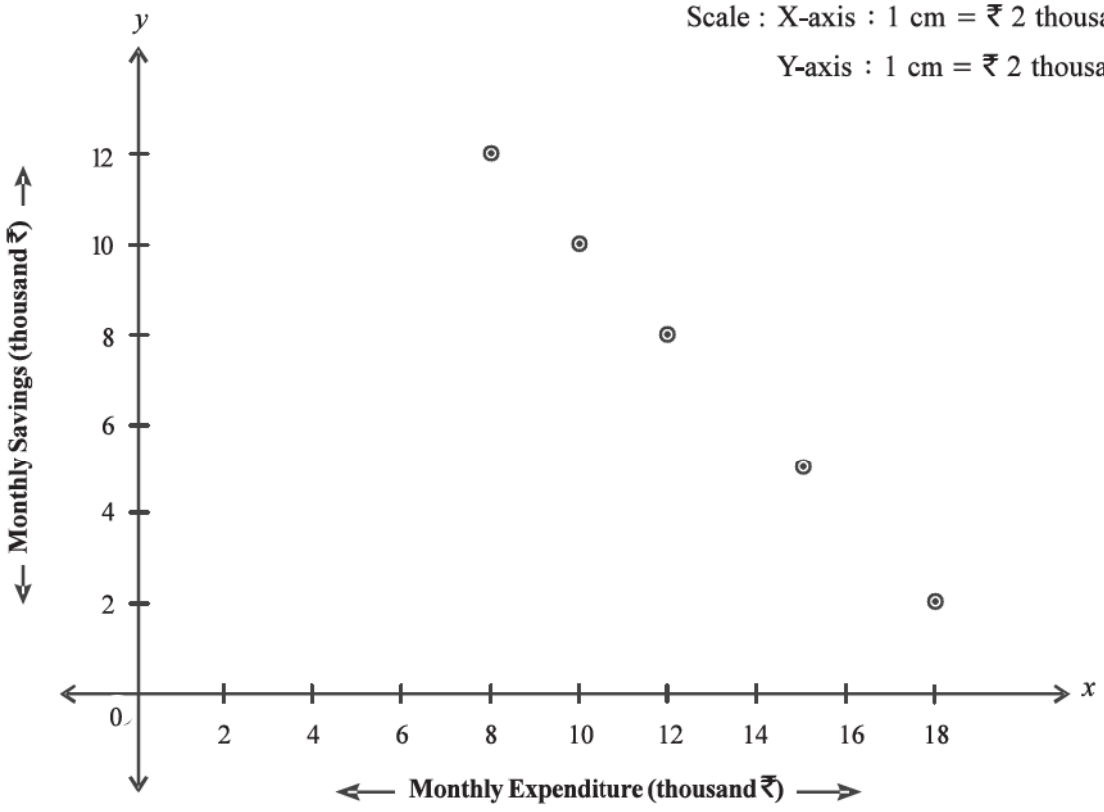
Monthly Expenditure (thousand ₹) x	15	18	8	10	12
Monthly Savings (thousand ₹) y	5	2	12	10	8

Draw a scatter diagram indicating the relation between monthly expenditure and monthly savings from this information and discuss about their correlation.

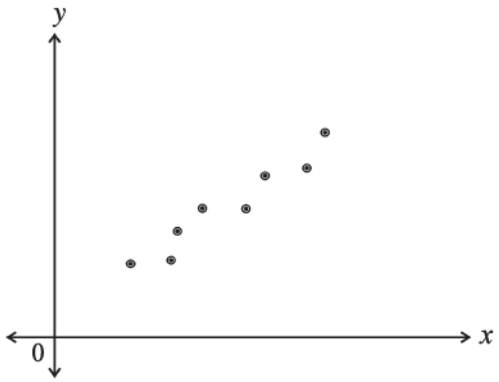
The following scatter diagram is obtained by plotting the points of ordered pairs $(15, 5), (18, 2), (8, 12), (10, 10), (12, 8)$ of X and Y on the graph paper.

Scale : X-axis : 1 cm = ₹ 2 thousand

Y-axis : 1 cm = ₹ 2 thousand



We can see that all the points on the scatter diagram, lie on the same line. We can also see that as the values of variable X change, the values of variable Y also change in opposite direction in constant proportion. (Check that the salary is fixed here for given five months, so the increase (or decrease) in the monthly expenditure results in constant propotionate decrease (or increase) in the monthly savings.) So, we can say that there is a perfect negative correlation between X and Y.



If all the points of scatter diagram are not on one line but lie around a line which is going in upward direction from left to right then it indicates that there is a **partial positive correlation** between two variables X and Y.

When the changes in the values of two variables are in same direction but not in the same proportion, we get such a scatter diagram. We take one example to understand this case.

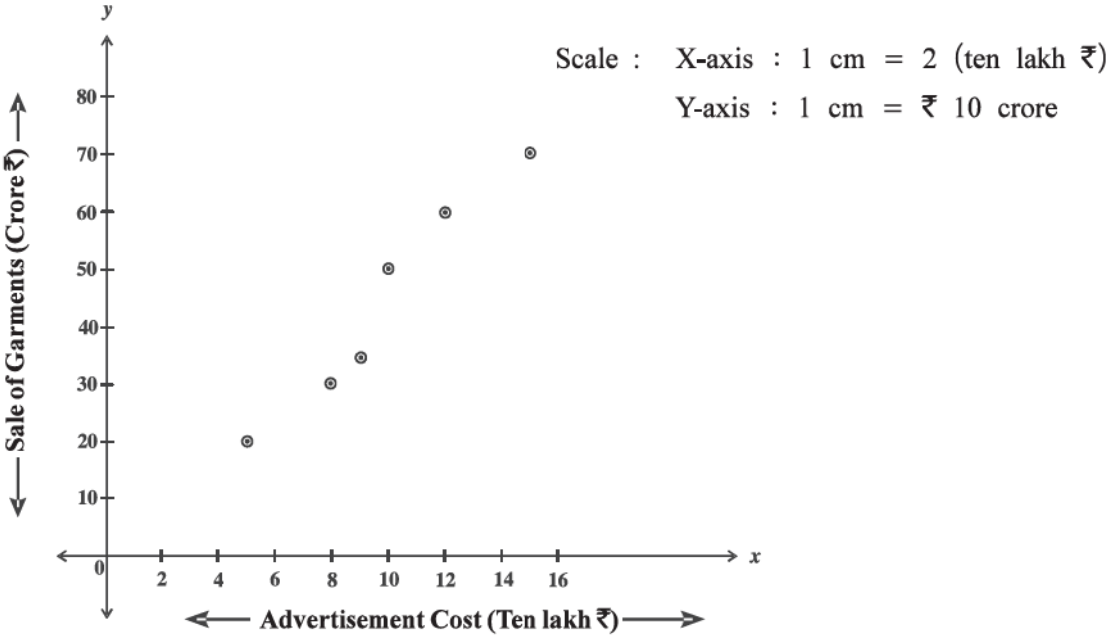
Note : To know the relation between the variables of a population, correlation coefficient is generally obtained by taking a sample of proportionate size. But for simplicity of calculation, we will keep the size of the sample limited and small.

Illustration 3 : A company manufactures readymade garments. The monthly advertisement cost (in ten lakh ₹) and the sale of garments (in crore ₹) for the last six months are given below :

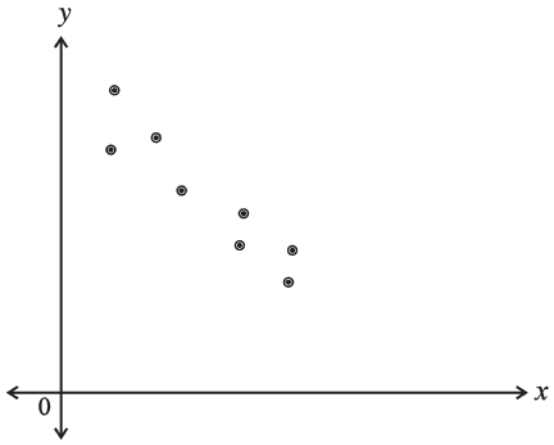
Advertisement Cost (ten lakh ₹) x	5	8	10	15	12	9
Sale of garments (crore ₹) y	20	30	50	70	60	35

Draw a scatter diagram from the data and discuss about the correlation.

The following scatter diagram is obtained by plotting the points of ordered pairs $(5, 20)$, $(8, 30)$, $(10, 50)$, $(15, 70)$, $(12, 60)$ and $(9, 35)$ of two variable X and Y on the graph paper.



We can see that all points on the scatter diagram do not lie on the same line. The changes in advertisement cost and sales are in the same direction but not in the same proportion. Hence, all the points are not on the same line. So, we can say that there is a partial positive correlation between X and Y.



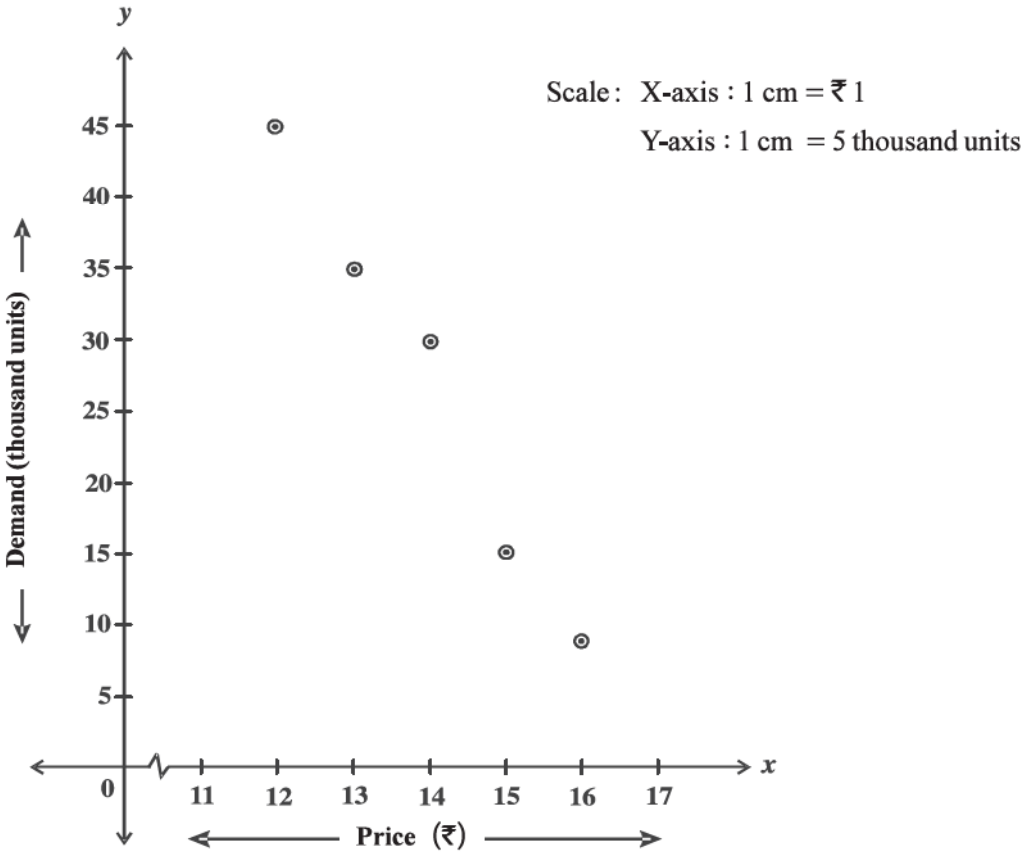
If all the points of scatter diagram are not on one line but lie around some line which is going in downward direction from left to right then it indicates that there is a **partial negative correlation** between two variables X and Y.

When the changes in the values of two variables are in opposite direction and not in same proportion, we get such a scatter diagram. We take one example to understand this case.

Illustration 4 : A company manufacturing spare parts for vehicles fixed different prices of rubber busher for five months to know the effect of price on its demand. Draw the scatter diagram from it and discuss about the nature of the correlation.

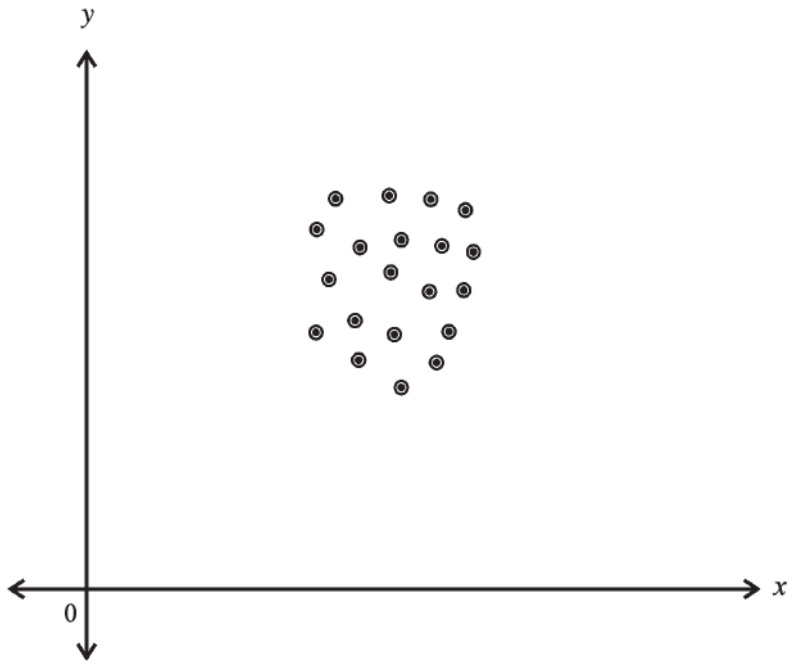
Price x	12	13	14	15	16
Monthly Demand (thousand units) y	45	35	30	15	10

By plotting the points of ordered pairs $(12, 45), (13, 35), (14, 30), (15, 15), (16, 10)$ of two variables x and y on the graph paper, we get the scatter diagram as follows.



We can see that all the points do not lie on the same line in the scatter diagram. The changes in price and demand are in opposite direction but they do not change in same proportion. Hence, all points are not on the same line. So, we can say that there is a partial negative correlation between X and Y.

Note : If the points are nearer to a line, it indicates high degree of correlation and if the points are widely spread around a line, it indicates low degree of correlation.



If the points of the scatter diagram lie randomly without forming any specific pattern then it indicates absence (lack) of linear correlation. When we get such a scatter diagram, we can say that two variables are not linearly related. i.e. there is a lack of linear correlation.

Merits and Limitations of Scatter Diagram Method

Merits :

(1) This is a simple method to know the nature of correlation between two variables.

(2) Less mathematical knowledge is required as the knowledge of plotting the points is only the requirment in this method.

(3) It also gives some idea about strength of correlation between two variables.

(4) The scatteredness of the points suggests whether the correlation is linear or not.

(5) There is no difficulty in judging the nature of correlation even if some pairs of extreme observations are present in the data.

Limitation :

This method gives nature of correlation and some idea about strength of correlation but it does not give exact degree of relationship between two variables.

EXERCISE 2.1

1. A ball pen making company wants to know the relation between the price (in ₹) and supply (in thousand units) of its most selling Gel Pen. The following information is collected for it. Draw a scatter diagram and interpret it.

Price (₹)	14	16	12	11	15	13	17
Monthly Supply (thousand units)	32	50	20	12	45	30	53

2. A company manufactures R.O. plants for the factories. The information about the advertisement cost for its sale and the profit from the sale of R.O. plants is given below.

Advertisement Cost (ten thousand ₹)	5	6	7	8	9	10	11
Profit (lakh ₹)	8	7	9	10	13	12	13

Draw a scatter diagram from this information and state the nature of the relationship between the advertisement cost and profit earned from the sale of R.O. plants.

3. The following information is collected to study the relationship between the minimum day temperature and sale of woollen clothes during a particular day of winter for six different cities.

Minimum day temperature (Celsius)	12	20	8	5	15	24
Sale of woollen clothes (thousand units)	35	10	45	70	20	8

Draw a scatter diagram from this information and interpret it.

*

2.5 Karl Pearson's Product Moment Method

We have already seen that the numerical measure of correlation which shows strength of correlation between two variables is known as correlation coefficient. Since this correlation coefficient was first suggested by statistician Karl Pearson, it is also known as 'Pearson Correlation Coefficient' or 'Product Moment Coefficient'. It is denoted by r.

Let n pairs of observations of a sample on two variables X and Y be $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$. The coefficient of correlation between variables X and Y is denoted by $r(x, y)$ or simply r and it can be obtained as follows :

$$r = \frac{Cov(x, y)}{s_x . s_y} = \frac{\text{Covariance } (x, y)}{(\text{S.D of } x)\ (\text{S.D of } y)}$$

Where,

$$\text{Covariance } (x, y) = Cov(x, y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\text{Standard Deviation of } x = s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$$

$$\text{Standard Deviation of } y = s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n}}$$

$$\text{Mean of } x = \bar{x} = \frac{\Sigma x_i}{n}$$

$$\text{Mean of } y = \bar{y} = \frac{\Sigma y_i}{n}$$

Note : For simplicity, the suffix i is ignored in all subsequent formulae in the study of linear correlation and linear regression and X is replaced by x and Y is replaced by y in the formulae and the calculations.

By substituting these values of $Cov(x, y)$, s_x and s_y in the above formula of r, we get the following form of r.

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

Generally, the above formula of r is used when both the means $\bar{x}$ and $\bar{y}$ are integers.

Now, the following are alternative formulae for $Cov(x, y)$, s_x and s_y.

$$Cov(x, y) = \frac{\Sigma xy - n\bar{x}\bar{y}}{n}$$

$$s_x = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

$$s_y = \sqrt{\frac{\Sigma y^2}{n} - \bar{y}^2}$$

The following are some other forms of formula of correlation coefficient r.

When the measures like $\Sigma(x-\bar{x})\,(y-\bar{y})$, S.D. of x, S.D. of y and n are known, we can use the following formula of r.

$$r = \frac{\Sigma(x-\bar{x})\,(y-\bar{y})}{n \cdot s_x \cdot s_y}$$

When the measures like Σxy, means $\bar{x}$ and $\bar{y}$, S.D. of x and y and n are known, we can use the following formula of r.

$$r = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{n \cdot s_x \cdot s_y}$$

When the measures like Σx, Σy, Σxy, Σx^2, Σy^2 and n are known, we can use the following formula of r.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Generally, the above formula of r is used when mean $\bar{x}$ or $\bar{y}$ or both are fractional.

Assumptions of Karl Pearson's Correlation Coefficient

Karl Pearson's Correlation Coefficient is based on the following assumptions.

(1) There is a linear relationship between two variables.

(2) There is a cause-effect relationship between two variables. Correlation is meaningless if there is no such relationship.

2.6 Properties of Correlation Coefficient

(1) The value of the correlation coefficient lies in the interval -1 to 1.
i.e. $-1 \le r \le 1$

(2) The correlation coefficient is free from unit of measurement. Whatever may be the units of two variables, r does not have any unit of measurement.

(3) The correlation coefficient between variables X and Y is same as the one of between Y and X.
i.e. $r(x, y) = r(y, x)$

(4) The correlation coefficient does not change with the change (transformation) of origin and scale.
Explanation : Suppose the correlation coefficient r is to be obtained between variables X and Y. Now, first we define the new transformed variables u and v as follows.

$$u = \frac{x-A}{c_x} \text{ and } v = \frac{y-B}{c_y}$$

Where, A, B, c_x, c_y are suitable real constants and $c_x > 0$ and $c_y > 0$

Now, this property of r implies that the correlation coefficient between u and v and the one between X and Y are equal. i.e. $r(u, v) = r(x, y) = r$

(5) $r(-x, y) = -r(x, y)$

$r(x, -y) = -r(x, y)$

i.e. If the sign of any one of the two variables X and Y is changed then the sign of the correlation coefficient also changes.

$r(-x, -y) = r(x, y)$

i.e. If the signs of both the variables are changed then the sign of the correlation coefficient remain unchanged.

Additional Information for understanding

When any constant is added or substracted from the values of the variable, it is called the change of origin because by doing so, the position of the origin changes in the corresponding graph. Note that the points of pairs (x, y) change their positions but their relative positions with respect to each other remain same. Hence, the correlation coefficient r remains unchanged after the change of origin.

Similarly, when any positive constant is multiplied or divided to the values of the variable, it is called change of scale because by doing so, the scale per unit of measurement changes on the axis. The relative positions of the points of the pairs (x, y) with respect to each other also remain same. Hence, the correlation coefficient r remains unchanged after the change of scale.

2.7 Interpretation of the value of Correlation Coefficient

We know that the correlation coefficient shows the type (nature) and the extent or strength of the relationship between two variables. It is necessary to interpret the value of the correlation coefficient after obtaining it. The type of correlation can be known by the sign of the correlation coefficient and its strength can be known by its numerical value.

While interpreting the value of r, we should keep in mind that r denotes the type and strength of linear relationship between two variables and it does not indicate the cause-effect relationship between them. In fact, we have tacitly assumed the cause-effect or any other type of relation between them. The value of r does not necessarily indicate the cause-effect relationship between two variables. In view of this fact, we shall now see how the value of r can be interpreted.

Interpretation of $r = 1$:

If $r = 1$ then we can say that there is a perfect positive correlation between two variables. When increase (decrease) in the value of one variable results in increase (decrease) in the value of the other variable in constant proportion then the value of r is 1. For such variables, we get all points on the same straight line in increasing direction in the scatter diagram. (see illustration 1.)

Interpretation of $r = -1$:

If $r = -1$ then we can say that there is a perfect negative correlation between two variables. When increase (decrease) in the value of one variable results in decrease (increase) in the value of the other variable in constant proportion then the value of r is -1. For such variables, we get all points on the same straight line in decreasing direction in the scatter diagram. (see illustration 2.)

Interpretation of $r = 0$:

If $r = 0$ then we can say that there is no linear correlation between two variables. In other words, $r = 0$ shows absence of correlation and hence two variables are said to be linearly uncorrelated. For such variables, we get randomly scattered points (not on or around any line) in the scatter diagram.

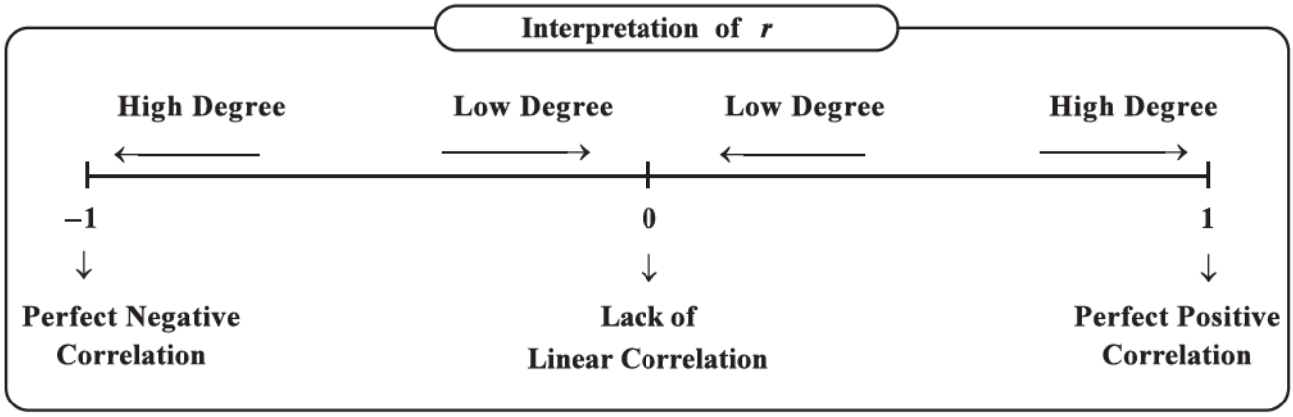
Note that the numerical value of r gives only the strength of linear correlation. So, when $r = 0$, we can only say that linear correlation between two variables is absent but there may be nonlinear

correlation. From the scatteredness of the points in the scatter diagram, we can get some idea about the type of nonlinear correlation.

Interpretation of Partial Correlation :

(Interpretation of $0 < r < 1$ and $-1 < r < 0$) :

If the value of r is between 0 and 1 or between –1 and 0, i.e. if $|r| < 1$ then we can say that there is a partial (imperfect) correlation between two variables. When the value of $|r|$ is nearer to 1, we can say that the relationship between two variables is close to perfect linear correlation and there is high degree of the relationship. In such a case, we can obtain a reliable estimate of changes in one variable corresponding to changes in value of the other variable. When the value of $|r|$ is nearer to 0, we can say that strength of the linear relation is very less and there is almost lack of linear correlation between two variables. In such a case, we cannot obtain a reliable estimate of the changes in the value of one variable corresponding to changes in the value of the other variable.



Note :

(1) Generally, we start the calculation of r by finding $\bar{x}$ and $\bar{y}$. But it is not necessary. We can select any suitable values of the constants A, B and c_x, c_y $(c_x > 0,\ c_y > 0)$ to start the calculation of r. We know that the value of r remains unaffected by any choice of these constants.

(2) For the given data of two variables, the value of the correlation coefficient r remains same by using any form of the formula of r of Karl Pearson's method.

Illustration 5 : The following information is collected by taking a sample of seven candidates having nearly same intellectual ability to know the effect of 'last days preparation' for a competitive examination of general knowledge on the 'result of the examination'.

Reading hours of last three days (hours)	25	38	30	28	34	40	36
Marks obtained in the examination	65	75	68	70	72	79	75

Find the correlation coefficient between reading hours of the last three days and marks obtained in the examination from the data and interpret it.

Here, $n = 7$, mean for reading hours (x) is $\bar{x} = \frac{\Sigma x}{n} = \frac{231}{7} = 33$, mean for marks (y) is

$\bar{y} = \frac{\Sigma y}{n} = \frac{504}{7} = 72$

Since both the means $\bar{x}$ and $\bar{y}$ are integers, we can obtain r as follows.

	Reading Hours x	Marks y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
	25	65	–8	–7	56	64	49
	38	75	5	3	15	25	9
	30	68	–3	–4	12	9	16
	28	70	–5	–2	10	25	4
	34	72	1	0	0	1	0
	40	79	7	7	49	49	49
	36	75	3	3	9	9	9
Total	**231**	**504**	**0**	**0**	**151**	**182**	**136**

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

$$= \frac{151}{\sqrt{182} \cdot \sqrt{136}}$$

$$= \frac{151}{\sqrt{182 \times 136}}$$

$$= \frac{151}{\sqrt{24752}}$$

$$= \frac{151}{157.3277}$$

$$= 0.9598$$

$$\therefore \quad r \simeq 0.96$$

We can see that the value of r is very near to 1. Thus, there is high degree of positive correlation between the reading hours and the marks. Hence, it can be said that generally if last days reading hours are more then more marks can be obtained in the examination.

Illustration 6 : In order to study the relation between the performance of students of a school in terms of marks in the subjects of Gujarati and Statistics, the following data are collected by taking a sample of six students.

Marks in Gujarati x	65	72	66	70	72	69
Marks in Statistics y	90	95	88	92	85	90

Compute the correlation coefficient between the marks obtained by the students in two subjects from this data.

Here, $n = 6$, $\bar{x} = \frac{\Sigma x}{n} = \frac{414}{6} = 69$, $\bar{y} = \frac{\Sigma y}{n} = \frac{540}{6} = 90$

Since both the means $\bar{x}$ and $\bar{y}$ are integers, we can obtain r as follows.

	Marks in Gujarati x	Marks in Statistics y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
	65	90	–4	0	0	16	0
	72	95	3	5	15	9	25
	66	88	–3	–2	6	9	4
	70	92	1	2	2	1	4
	72	85	3	–5	–15	9	25
	69	90	0	0	0	0	0
Total	**414**	**540**	**0**	**0**	**8**	**44**	**58**

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

$$= \frac{8}{\sqrt{44} \cdot \sqrt{58}}$$

$$= \frac{8}{\sqrt{2552}}$$

$$= \frac{8}{50.5173}$$

$$= 0.1584$$

$\therefore \quad r \simeq 0.16$

We can see that the value of r is close to 0. Thus, there is low degree of positive correlation between the marks of students in these two subjects.

Illustration 7 : The details of monthly sale of mobile phones (in thousand units) and its profit (in lakh ₹) for the last six months for a company making mobile phones are given below.

No. of mobile phones sold (thousand units) x	3	8	12	5	7	5
Profit (lakh ₹) y	6	10	15	10	9	8

Find the correlation coefficient between 'number of mobile phones sold' and its 'profit'.

Here, $n = 6$, $\bar{x} = \frac{\Sigma x}{n} = \frac{40}{6} = 6.67$, $\bar{y} = \frac{\Sigma y}{n} = \frac{58}{6} = 9.67$

Since the means $\bar{x}$ and $\bar{y}$ are fractional and the values of X and Y are not very large, we can compute r as follows.

	Number of Mobile phones sold (thousand units) x	Profit (lakh ₹) y	xy	x^2	y^2
	3	6	18	9	36
	8	10	80	64	100
	12	15	180	144	225
	5	10	50	25	100
	7	9	63	49	81
	5	8	40	25	64
Total	**40**	**58**	**431**	**316**	**606**

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{6(431) - (40)(58)}{\sqrt{6(316) - (40)^2} \cdot \sqrt{6(606) - (58)^2}}$$

$$= \frac{2586 - 2320}{\sqrt{1896 - 1600} \cdot \sqrt{3636 - 3364}}$$

$$= \frac{266}{\sqrt{296} \cdot \sqrt{272}}$$

$$= \frac{266}{\sqrt{80512}}$$

$$= \frac{266}{283.7464}$$

$$= 0.9375$$

$$\therefore \quad r \simeq 0.94$$

We can see that the value of r is close to 1. Thus, there is a high degree of positive correlation between sale of mobile phones and its profit.

Illustration 8 : From the following information of weekly minimum temperature (in Celsius) and the sale (in hundred units) of heaters during a week in a city of North India for five weeks, calculate the correlation coefficient between minimum temperature and sale of heaters.

Minimum temperature (Celsius) x	3	4	6	7	9
Demand of heaters (hundred units) y	16	15	14	11	9

Here, $n = 5$, $\bar{x} = \frac{\Sigma x}{n} = \frac{29}{5} = 5.8$, $\bar{y} = \frac{\Sigma y}{n} = \frac{65}{5} = 13$

Since one of the means is fractional and the values of variables X and Y are not very large, we shall compute r as follows.

	Minimum temperature (Celcius) x	**Demand of heaters (hundred units)** y	xy	x^2	y^2
	3	16	48	9	256
	4	15	60	16	225
	6	14	84	36	196
	7	11	77	49	121
	9	9	81	81	81
Total	**29**	**65**	**350**	**191**	**879**

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{5(350) - (29)(65)}{\sqrt{5(191) - (29)^2} \cdot \sqrt{5(879) - (65)^2}}$$

$$= \frac{1750 - 1885}{\sqrt{955 - 841} \cdot \sqrt{4395 - 4225}}$$

$$= \frac{-135}{\sqrt{114} \cdot \sqrt{170}}$$

$$= \frac{-135}{\sqrt{19380}}$$

$$= \frac{-135}{139.2121}$$

$$= -0.9697$$

$\therefore\ r \simeq -0.97$

We can see that the value of r is very close to -1. Thus, there is a high degree of negative correlation between minimum temperature and demand of heaters.

In illustrations (7) and (8), we have seen that both means were not integers and the values of X and Y were not very large and hence we have used the following formula to compute the value of r.

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

But when the values of two variables are numerically large and/or fractional, the computation of xy, x^2, y^2 becomes more difficult and hence the calculation of r becomes tedius. So, in order to make the computation of r simple, a short-cut method is used. This short-cut method is based on the property (No. 4) of r.

According to this property, replacing x by u and y by v in the formula of r, we get the following formula of r by short-cut method.

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

Now, we consider some examples to illustrate the computation of r by the short-cut method.

Illustration 9 : To know the relation between the heights and weights of the students of a school, a sample of six students is taken and the following information is obtained. Find the correlation coefficient between the heights and weights of the students.

Height (cm) x	155	165	158	162	153	160
Weight (kg) y	53	63	56	60	52	60

Here, $n = 6$, $\bar{x} = \frac{\Sigma x}{n} = \frac{953}{6} = 158.83$, $\bar{y} = \frac{\Sigma y}{n} = \frac{344}{6} = 57.33$

Since both the means $\bar{x}$ and $\bar{y}$ are fractional and the values of the variables X and Y are large, we shall prefer the short-cut method for the computation of r.

We take $A = 158$ and $B = 57$ to change the origin and since there are no suitable values for change of scale, we take $c_x = 1$, $c_y = 1$.

We shall define new variables u and v as follows :

$$u = \frac{x-A}{c_x} = \frac{x-158}{1} = x - 158$$

$$v = \frac{y-B}{c_y} = \frac{y-57}{1} = y - 57$$

Note : Since only origin is changed and the scale is not changed, we can also define new variables u and v as follows.

$u = x - A = x - 158$; $v = y - B = y - 57$

	Height (cm) x	**Weight (kg) y**	**$u = x - 158$**	**$v = y - 57$**	**uv**	**u^2**	**v^2**
	155	53	–3	–4	12	9	16
	165	63	7	6	42	49	36
	158	56	0	–1	0	0	1
	162	60	4	3	12	16	9
	153	52	–5	–5	25	25	25
	160	60	2	3	6	4	9
Total	**953**	**344**	**5**	**2**	**97**	**103**	**96**

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{6(97) - (5)(2)}{\sqrt{6(103) - (5)^2} \cdot \sqrt{6(96) - (2)^2}}$$

$$= \frac{582 - 10}{\sqrt{618 - 25} \cdot \sqrt{576 - 4}}$$

$$= \frac{572}{\sqrt{593} \cdot \sqrt{572}}$$

$$= \frac{572}{\sqrt{339196}}$$

$$= \frac{572}{582.4054}$$

$$= 0.9821$$

$\therefore \quad r \simeq 0.98$

We can see that the value of r is very close to 1. Thus, there is a high degree of positive correlation between the height and the weight of the students.

Illustration 10 : To know the relation between the average monthly income (in ₹) and income due to overtime of workers (in ₹), the following information is obtained from six different factories of an area manufacturing similar kind of products. Find the correlation coefficient between the average monthly income and income due to overtime.

Year	2011	2012	2013	2014	2015	2016
Average Monthly Income (₹) x	14,900	15,100	15,000	15,500	15,700	15,800
Income due to overtime (₹) y	100	105	115	160	220	255

Here, $n = 6$, $\bar{x} = \frac{\Sigma x}{n} = \frac{92000}{6} = 15333.33$, $\bar{y} = \frac{\Sigma y}{n} = \frac{955}{6} = 159.17$

We also observe that the values of X are in multiple of 100 and the values of Y are in multiple of 5. So, we shall take $A = 15{,}300$, $B = 160$, $c_x = 100$, $c_y = 5$.

Let us define new variables u and v as follows.

$$u = \frac{x - A}{c_x} = \frac{x - 15300}{100} \text{ and } v = \frac{y - B}{c_y} = \frac{y - 160}{5}$$

Note : Since the values of x are in multiple of 100, we have chosen A (= 15,300) also in multiple of 100. Similarly, as the values of y are in multiple of 5, we have chosen B (= 160) also in multiple of 5.

	Average monthly income (₹) x	Income due to overtime (₹) y	$u = \frac{x-15300}{100}$	$v = \frac{y-160}{5}$	uv	u^2	v^2
	14,900	100	–4	–12	48	16	144
	15,100	105	–2	–11	22	4	121
	15,000	115	–3	–9	27	9	81
	15,500	160	2	0	0	4	0
	15,700	220	4	12	48	16	144
	15,800	255	5	19	95	25	361
Total	**92,000**	**955**	**2**	**–1**	**240**	**74**	**851**

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{6(240) - (2)(-1)}{\sqrt{6(74) - (2)^2} \cdot \sqrt{6(851) - (-1)^2}}$$

$$= \frac{1440 + 2}{\sqrt{444 - 4} \cdot \sqrt{5106 - 1}}$$

$$= \frac{1442}{\sqrt{440} \cdot \sqrt{5105}}$$

$$= \frac{1442}{\sqrt{2246200}}$$

$$= \frac{1442}{1498.7328}$$

$$= 0.9621$$

$\therefore \quad r \simeq 0.96$

We can see that the value of r is very close to 1. Thus, there is a high degree of positive correlation between average monthly income and income due to overtime.

Illustration 11 : For six different cities of Gujarat State, the approximate figures regarding their density of population (per square km) and the death rate (per thousand) are given below.

City	*A*	*B*	*C*	*D*	*E*	*F*
Density (per sq. km) x	200	500	400	700	600	300
Death rate (per thousand) y	10	12	10	15	9	12

Find the correlation coefficient between the density of population and death rate from this information.

Here, $n = 6$, $\bar{x} = \frac{\Sigma x}{n} = \frac{2700}{6} = 450$, $\bar{y} = \frac{\Sigma y}{n} = \frac{68}{6} = 11.33$

We can also see that the values of X are in multiple of 100 and the values of Y are small. So, we shall take $A = 500$, $B = 12$, $c_x = 100$ and $c_y = 1$.

Let us define new variables u and v as follows :

$$u = \frac{x-A}{c_x} = \frac{x-500}{100} \text{ and } v = \frac{y-B}{c_y} = \frac{y-12}{1} = y - 12$$

	Density (per sq. km) x	Death rate (per thousand) y	$u = \frac{x-500}{100}$	$v = y-12$	uv	u^2	v^2
	200	10	–3	–2	6	9	4
	500	12	0	0	0	0	0
	400	10	–1	–2	2	1	4
	700	15	2	3	6	4	9
	600	9	1	–3	–3	1	9
	300	12	–2	0	0	4	0
Total	**2700**	**68**	**–3**	**–4**	**11**	**19**	**26**

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{6(11) - (-3)(-4)}{\sqrt{6(19) - (-3)^2} \cdot \sqrt{6(26) - (-4)^2}}$$

$$= \frac{66 - 12}{\sqrt{114 - 9} \cdot \sqrt{156 - 16}}$$

$$= \frac{54}{\sqrt{105} \cdot \sqrt{140}}$$

$$= \frac{54}{\sqrt{14700}}$$

$$= \frac{54}{121.2436}$$

$$= 0.4454$$

$\therefore \quad r \simeq 0.45$

We can see that the value of r is not very close to 1. Thus, there is a moderate degree of positive correlation between density of population and death rate.

Illustration 12 : In order to study the relation between the sales (in crore ₹) and the profit (in thousand ₹) for truck tyre manufacturing companies, the following information is obtained for the last year.

Sales (crore ₹) x	1.6	2.2	1.9	2.0	2.3	1.7	2.4	1.8	2.1
Profit (thousand ₹) y	4200	5500	6000	6200	6100	4900	5900	5000	6700

Find the correlation coefficient between the sales and the profit from it.

Here, $n = 9$, $\bar{x} = \frac{\Sigma x}{n} = \frac{18}{9} = 2$, and $\bar{y} = \frac{\Sigma y}{n} = \frac{50500}{9} = 5611.11$

Since the values of variable X are fractional and having one digit after decimal point, we shall multiply the values of X by 10 (i.e. divide by $\frac{1}{10} = 0.1$) to make them integer and the values of Y are multiples of 100, we shall take $A = 2$, $B = 5600$, $c_x = 0.1$, $c_y = 100$

Let us now define variables u and v as follows.

$$u = \frac{x-A}{c_x} = 10\,(x - 2.0) = \frac{x-2.0}{0.1} \text{ and } v = \frac{y-B}{c_y} = \frac{y-5600}{100}$$

	Sales (crore ₹) x	**Profit (thousand ₹) y**	$u = 10(x-2.0)$	$v = \frac{y-5600}{100}$	uv	u^2	v^2
	1.6	4200	–4	–14	56	16	196
	2.2	5500	2	–1	–2	4	1
	1.9	6000	–1	4	–4	1	16
	2.0	6200	0	6	0	0	36
	2.3	6100	3	5	15	9	25
	1.7	4900	–3	–7	21	9	49
	2.4	5900	4	3	12	16	9
	1.8	5000	–2	–6	12	4	36
	2.1	6700	1	11	11	1	121
Total	**18**	**50,500**	**0**	**1**	**121**	**60**	**489**

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{9(121) - (0)(1)}{\sqrt{9(60) - (0)^2} \cdot \sqrt{9(489) - (1)^2}}$$

$$= \frac{1089 - 0}{\sqrt{540 - 0} \cdot \sqrt{4401 - 1}}$$

$$= \frac{1089}{\sqrt{540} \cdot \sqrt{4400}}$$

$$= \frac{1089}{\sqrt{2376000}}$$

$$= \frac{1089}{1541.4279}$$

$$= 0.7065$$

$\therefore \quad r \simeq 0.71$

We can see that the value of r is slightly far from 1. So, we can say that there is a moderately high degree of positive correlation between the sales and the profit.

Activity

Calculate the correlation coefficient again from the data given in illustration 12 by taking $A = 1.8$, $B = 6000$, $c_x = 0.05$ and $c_y = 100$. You will see that the value of r will be the same ($= 0.71$).

Illustration 13 : To study the relationship between the marks obtained in Statistics (X) and marks in Economics (Y) of the students of a school, a sample of ten students is taken and the following information is obtained.

$$\Sigma(x-\bar{x})(y-\bar{y}) = 120,\ \Sigma(x-\bar{x})^2 = 80,\ \Sigma(y-\bar{y})^2 = 500$$

Find the value of r.

Here, $n = 10$ and the following formula is suitable to the given data.

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

$$= \frac{120}{\sqrt{80} \cdot \sqrt{500}}$$

$$= \frac{120}{\sqrt{40000}}$$

$$= \frac{120}{200}$$

$\therefore \quad r = 0.6$

Illustration 14 : Find the value of r from the following data.

(1) $n = 20,\ Cov(x, y) = -50,\ s_x = 15,\ s_y = 8$

(2) $n = 10,\ \Sigma(x-\bar{x})(y-\bar{y}) = 60,$ variance of $X = 25,$ variance of $Y = 36$

(3)

Particulars	x	y
No. of observations	25	
Mean	40	50
The sum of squares of deviations taken from mean	120	160
The sum of product of deviations taken from mean	100	

(4) $n = 10,\ \Sigma xy = 1500,$ mean of $X = 12,$ mean of $Y = 15,$ S.D. of $X = 9,$ S.D. of $Y = 5.$

(1) Here, $n = 20,\ Cov(x, y) = -50,\ s_x = 15,\ s_y = 8$

Substituting these values in the following formula,

$$r = \frac{Cov(x, y)}{s_x \cdot s_y}$$

$$= \frac{-50}{(15)\ (8)}$$

$$= \frac{-50}{120}$$

$$= -0.4167$$

$$\therefore\ r \simeq -0.42$$

(2) Here, $n = 10,\ \Sigma(x-\bar{x})(y-\bar{y}) = 60$

Variance of $X = s_x^2 = 25 \quad \therefore\ s_x = 5$

Variance of $Y = s_y^2 = 36 \quad \therefore\ s_y = 6$

Substituting the required values in the following suitable formula,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n s_x s_y}$$

$$= \frac{60}{10(5)\ (6)}$$

$$= \frac{60}{300}$$

$$\therefore\ r = 0.2$$

(3) Here, $n = 25,\ \bar{x} = 40,\ \bar{y} = 50,\ \Sigma(x-\bar{x})^2 = 120,\ \Sigma(y-\bar{y})^2 = 160$ and $\Sigma(x-\bar{x})(y-\bar{y}) = 100$

Substituting all these values in the following suitable formula,

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

$$= \frac{100}{\sqrt{120} \cdot \sqrt{160}}$$

$$= \frac{100}{\sqrt{19200}}$$

$$= \frac{100}{138.5641}$$

$$= 0.7217$$

$$\therefore\ r \simeq 0.72$$

(4) Here, $n = 10,\ \Sigma xy = 1500,\ \bar{x} = 12,\ \bar{y} = 15,\ s_x = 9$ and $s_y = 5$

Substituting all these values in the following suitable formula,

$$r = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{n \cdot s_x \cdot s_y}$$

$$= \frac{1500 - 10\,(12)\,(15)}{10\,(9)\,(5)}$$

$$= \frac{1500 - 1800}{450}$$

$$= \frac{-300}{450}$$

$$= -0.6667$$

$$\therefore\ r \simeq -0.67$$

Illustration 15 : To study the relationship between the sales and the profit of a company, the following information is obtained for the last six years.

X = Annual Sales (lakh ₹)

Y = Annual Profit (ten thousand ₹)

$n = 6,\ \Sigma x = 58,\ \Sigma y = 40,\ \Sigma xy = 431,\ \Sigma x^2 = 606,\ \Sigma y^2 = 316$

Find the correlation coefficient between X and Y.

Substituting the given values in the following formula,

$$r = \frac{n\Sigma xy - (\Sigma x)\,(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{6(431) - (58)\,(40)}{\sqrt{6(606) - (58)^2} \cdot \sqrt{6(316) - (40)^2}}$$

$$= \frac{2586 - 2320}{\sqrt{3636 - 3364} \cdot \sqrt{1896 - 1600}}$$

$$= \frac{266}{\sqrt{272} \cdot \sqrt{296}}$$

$$= \frac{266}{\sqrt{80512}}$$

$$= \frac{266}{283.7464}$$

$$= 0.9375$$

$$\therefore\ r \simeq 0.94$$

Merits and Limitations of Karl-Pearson's Method

Merits :

(1) The type (nature) as well as the degree of relationship between two variables can be known by this method.

(2) This is the most popular method of measuring linear correlation between two variables.

(3) It summarises the degree (high, moderate or low) of correlation in a single number.

Limitations :

(1) It depends on the assumption that two variables under consideration have linear relationship. So, if it is used inspite of not having linear relationship between two variables then the interpretation based on it will be misleading.

(2) The value of correlation coefficient is highly affected by the extreme (too large or too small) observations.

(3) Atmost care is required to interpret this coefficient. Otherwise the relationship between two variables may be misunderstood.

Exercise 2.2

1. From the following information obtained from a sample of 7 families of a society regarding height of father (in cms) and height of his adult son (in cms), calculate the correlation coefficient.

Height of father (cm)	170	169	168	167	166	165	164
Height of son (cm)	172	168	170	168	165	167	166

2. A local cottage industry making various snacks sells each snack in a packet of 100 gms. From a study for price determination regarding a new kind of wafer, the following information is obtained for the price and the demand.

Price (₹)	24	26	32	33	35	30
Demand (thousand units)	27	24	22	20	15	24

Find the correlation coefficient between price of wafer and its demand.

3. From the following information of a sample of six students of a school regarding their marks in two subjects Accountancy and Statistics, find the coefficient of correlation between the marks of two subjects.

Marks in Accountancy	60	80	50	80	95	40	70	40	35	90
Marks in Statistics	50	75	60	85	90	40	65	30	45	70

4. To check the ability of Mathematics and Logic of the students of a city, a private educational institute gives twenty puzzles based on these two subjects to six children selected from various schools. The number of puzzles solved by them is given below.

No. of puzzles solved based on Mathematics	12	8	9	10	8	11
No. of puzzles solved based on Logic	11	10	4	7	13	16

Compute the correlation coefficient between performances of children in two types of puzzles using the given data.

5. Find the correlation coefficient between capital (in crore ₹) invested and the profit (in crore ₹) from the following data.

Company	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Capital investment (crore ₹)	15	22	12	10	17	20	14
Profit (crore ₹)	9	12	8	6	10	9	10

6. The following information is available for five students selected from a school regarding the average number of study hours per day and the average number of sleeping hours.

No. of study hours	10	5	7	5	3
No. of sleeping hours	6	9	7	8	10

Calculate the correlation coefficient between the study hours and sleeping hours.

7. From the following information of the age (in years) and blood pressure (in mm), find the correlation coefficient between age and blood pressure.

Age (years)	58	55	65	52	48	68	62	56
Systolic blood pressure (mm)	130	150	150	130	140	158	155	140

8. An Engineer Association wants to know the relation between the production (thousand units) and the unit production cost of different factories. The information collected from six factories regarding their production and unit production cost is given below.

Production (thousand units)	15	20	35	24	18	31
Cost per unit of production (₹)	95	90	75	80	87	70

Find the correlation coefficient between production and cost per unit of production.

9. Find the correlation coefficient between the yearly per capita income (in ₹) and the price index of the people of six different cities from the following data.

City	*A*	*B*	*C*	*D*	*E*	*F*
Yearly per capita income (₹)	32,000	29,000	40,000	36,000	30,000	39,000
Price index	120	100	250	180	110	220

10. The following data are given to study the relation between the number of persons in a family who drive vehicle and usage of petrol (in Litres) per week.

No. of members per family who drive vehicle	3	5	2	4	3	6	1
Weekly usage of petrol (litre)	11.5	21	14.5	15.5	7	22.5	10

Find the correlation coefficient between the number of members in a family and usage of petrol.

11. The following information is obtained to study the effect of the use of fertilizer on yield of corn in a rural area.

Use of fertilizer (quintal)	1.5	2.1	0.9	1.8	1.1	1.2
Yield of corn per hectare (quintal)	60	95	50	75	45	75

Find the correlation coefficient between use of fertilizer and yield of corn.

12. Find the correlation coefficient from the following information of rainfall (X) (in cm) and yield (Y) (tons per hactare) for the last 10 years of a district.

$n = 10$, $Cov\ (x, y) = 30$, S.D. of $X = 5$ and variance of $Y = 144$

13. The following information is obtained regarding the height (X) and weight (Y) from a sample of ten students of a school.

$\bar{x} = 160$, $\bar{y} = 55$, $\Sigma xy = 90000$, $s_x = 25$, $s_y = 10$

Find the correlation coefficient between the height and weight from it.

14. Determine the value of correlation coefficient from the following data.

(1) $\Sigma(x-\bar{x})^2 = 72$, $\Sigma(y-\bar{y})^2 = 32$, $\Sigma(x-\bar{x})(y-\bar{y}) = 45$

(2) $n = 6$, $\Sigma x = 16$, $\Sigma y = 51$, $\Sigma xy = 154$, $\Sigma x^2 = 52$, $\Sigma y^2 = 471$

15. Find the value of r from the following data.

Particulars	x	y
Average	60	95
The sum of squares of deviations taken from their mean	920	1050
The sum of product of deviations taken from their mean	-545	

*

2.8 Spearman's Rank Correlation Method

We have studied the Karl Pearson's method to find the correlation coefficient between two variables. It is clear that this method is used when two variables are quantitative, i.e. both the variables are numerically measurable. But in problems of business, industry and social science there are some situations when we deal with qualitative variable (attribute). e.g. beauty, honesty, skill, morality, proficiency in elocution, music, dance, etc. In such situations these characteristics (qualitative variables or attributes) are incapable of quantitative measurement, but they can be ranked serially according to their proficiency in such characteristics. The correlation coefficient computed using such ranks of two characteristics (qualitative variables) as suggested by Charles Spearman is known as Spearman's rank correlation coefficient.

Some of the examples in which the Spearman's rank correlation coefficient is to be obtained to know the relationship are as follows :

(1) The study of the relationship between two attributes honesty and punctuality for a group of persons by ranking them according to their proficiency.

(2) The study of the relationship between the judgements of two judges in a beauty contest who rank the contestants according to their performance in the contest.

Sometimes, even if the data consist of quantitative variables (like height, weight), the rank correlation coefficient is obtained by assigning the ranks to the numerical observations according to their magnitudes. Generally, when there is more dispersion among the values of the observations of two variables, rank correlation coefficient is preferred over Karl Pearson's correlation coefficient because rank correlation coefficient is more stable about dispersion of the observations as compared to Karl Perason's correlation coefficient.

Method : Suppose the ranks for n pairs of observations of two attributes X and Y are assigned as follows :

Serial Number of observation	1	2	.....	i	.....	n
Rank based on X	R_{x_1}	R_{x_2}	.....	R_{x_i}	.....	R_{x_n}
Rank based on Y	R_{y_1}	R_{y_2}	.....	R_{y_i}	.....	R_{y_n}

The following formula is used to find the rank correlation coefficient.

$$r = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}$$

Where, $d_i = R_{x_i} - R_{y_i}$, for $i = 1, 2, 3, \ldots\ldots n$

For the sake of convenience, we shall write d in place of d_i, R_x in place of R_{x_i} and R_y in place of R_{y_i}.

So, the formula for rank correlation coefficient can be written as follows.

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Where, $d = R_x - R_y =$ difference between the ranks of X and Y.

$\Sigma d^2 =$ the sum of squares of difference in the ranks of X and Y.

When n pairs of observations are numerical measurements (quantitative variables), generally the highest observation is given the rank 1, the second highest observation is given rank 2 and so on for each variable and the rank correlation coefficient can be obtained from these ranks.

Although we use rank correlation coefficient to find the relationship between two quantitative variables, Karl Pearson's correlation coefficient is more efficient as actual observations are used in the calculations and not the ranks.

Note that the Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between the ranks of two attributes (variables). Therefore, it can be interpreted in the same way as we interpret the Karl Pearson's correlation coefficient.

Generally, the values of the Spearman's rank correlation coefficient and Karl Pearson's correlation coefficient are not equal. But when the values of two variables are some arrangement of first n natural numbers, the correlation coefficients obtained by Karl Pearson's method and Spearman's method are equal.

Illustration 16 : Seven employees are selected from a company. They are judged by two managers from the company in terms of their administrative skills. The ranks given by them are as follows.

Employee	A	B	C	D	E	F	G
Rank by Manager 1	6	7	5	4	3	2	1
Rank by Manager 2	7	6	5	2	4	1	3

Calculate the rank correlation coefficient between the judgments given by two managers.

Here, $n = 7$ and the ranks are already given. So, we prepare the following table to compute the rank correlation coefficient.

Worker	Rank by Manager 1 R_x	Rank by Manager 2 R_y	$d = R_x - R_y$	d^2
A	6	7	–1	1
B	7	6	1	1
C	5	5	0	0
D	4	2	2	4
E	3	4	–1	1
F	2	1	1	1
G	1	3	–2	4
Total	–	–	**0**	**12**

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(12)}{7(49-1)}$$

$$= 1 - \frac{72}{336}$$

$$= 1 - 0.2143$$

$$= 0.7857$$

$$\therefore \ r \simeq 0.79$$

The value of r is near to 1, i.e. it shows partial positive correlation. So, it can be said that there is a similarity between the ranks of the employee given by two managers.

Illustration 17 : The owner of chain of mobile phone shops selling the mobile phones of various brands invited an expert person. The owner asked him to check and rank ten different mobile phones regarding its camera and battery efficiency. The ranks given by the expert to 10 mobile phones of different brands are given below.

Mobile Phone	A	B	C	D	E	F	G	H	I	J
Rank for camera	3	5	8	4	7	10	2	1	6	9
Rank for battery	6	4	9	8	1	2	3	10	5	7

Find the rank correlation coefficient between the efficiency of camera and battery of the mobile phones.

Here $n = 10$ and the ranks are already given, so we prepare a table as follows to compute the rank correlation coefficient.

Mobile phone	For Camera R_x	For Battery R_y	$d = R_x - R_y$	d^2
A	3	6	–3	9
B	5	4	1	1
C	8	9	–1	1
D	4	8	–4	16
E	7	1	6	36
F	10	2	8	64
G	2	3	–1	1
H	1	10	–9	81
I	6	5	1	1
J	9	7	2	4
Total	**–**	**–**	**0**	**214**

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(214)}{10(100-1)}$$

$$= 1 - \frac{1284}{990}$$

$$= 1 - 1.2970$$

$$= -0.2970$$

$$\therefore r \simeq -0.30$$

The value of r is negative and near to 0, i.e. it shows partial negative correlation. So, it can be said that according to expert's review, if camera is efficient then its battery is less efficient and if battery is efficient then its camera is less efficient for the mobile phones checked by him.

Illustration 18 : The principal of a school has conducted a test for five students selected in a sample to judge the relation between the knowledge of Mathematics and memory ability in the subject of History of the students. The ranks given to these five students in the subjects of Mathematics and History are given below. Find the rank correlation coefficient between ranks of two subjects using this information.

Student	A	B	C	D	E
Rank in Mathematics	2	5	1	4	3
Rank in History	4	1	5	2	3

Here, $n = 5$ and the ranks are already given. So, we prepare the following table to compute the rank correlation coefficient.

Student	R_x	R_y	$d = R_x - R_y$	d^2
A	2	4	–2	4
B	5	1	4	16
C	1	5	–4	16
D	4	2	2	4
E	3	3	0	0
Total	–	–	**0**	**40**

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(40)}{5(25-1)}$$

$$= 1 - \frac{240}{120}$$

$$= 1 - 2$$

$$\therefore \quad r = -1$$

Since the performance of five students in Mathematics and History are exactly in opposite order, we get $r = -1$.

Illustration 19 : The singing talent of five singers A, B, C, D and E was judged by two judges in a singing competition. The ranks assigned to five singers are as follows.

Rank	1	2	3	4	5
By Judge 1	C	A	B	E	D
By Judge 2	B	C	D	A	E

Find the similarity between the decisions of the two judges from the rank correlation coefficient.

Here $n = 5$. Let us rearrange the given information in terms of ranks given to five singers as follows.

Singer	A	B	C	D	E
Rank by Judge 1	2	3	1	5	4
Rank by Judge 2	4	1	2	3	5

Singer	R_x	R_y	$d = R_x - R_y$	d^2
A	2	4	–2	4
B	3	1	2	4
C	1	2	–1	1
D	5	3	2	4
E	4	5	–1	1
Total	**–**	**–**	**0**	**14**

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(14)}{5(25-1)}$$

$$= 1 - \frac{84}{120}$$

$$= 1 - 0.7$$

$$\therefore \ r = 0.3$$

The value of r is positive and near to 0. So, there is a partial positive correlation. Thus, it can be said that there is less agreement between the judges. i.e. their opinion differ to some extent.

Illustration 20 : To know the relationship between the abilities of the students of a school in the subjects of Statistics and Accountancy, the teachers of both the subjects have taken a sample of eight students and the following information was collected.

Student	1	2	3	4	5	6	7	8
Marks in Statistics x	78	36	98	25	75	82	90	62
Marks in Accountancy y	84	51	91	60	68	62	86	55

Calculate the rank correlation coefficient between the marks of Statistics and the marks of Accountancy.

Here, $n = 8$. We are given quantitative variables (marks in two subjects). So, first we have to assign the ranks according to the marks obtained by the students in two subjects.

In Statistics, a student with roll no. 3 gets the highest marks 98 and hence is ranked 1; roll no. 7 gets 90 marks (second highest) and hence is ranked 2 and so on. Similarly, we can assign the ranks to the students based on their marks in the subject of Accountancy.

Let us prepare a table as follows.

Roll No.	Statistics		Accountancy		$d = R_x - R_y$	d^2
	Marks	Rank R_x	Marks	Rank R_y		
1	78	4	84	3	1	1
2	36	7	51	8	–1	1
3	98	1	91	1	0	0
4	25	8	60	6	2	4
5	75	5	68	4	1	1
6	82	3	62	5	–2	4
7	90	2	86	2	0	0
8	62	6	55	7	–1	1
Total	–	–	–	–	**0**	**12**

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(12)}{8(64-1)}$$

$$= 1 - \frac{72}{504}$$

$$= 1 - 0.1429$$

$$= 0.8571$$

$$\therefore r \simeq 0.86$$

The value of r is close to 1. So, it can be said that there is a high degree of positive correlation between the ranks obtained from the marks of Statistics and Accountancy. i.e. If a student is having more (less) marks in Statistics then his marks in Accountancy are also more (less). This may not always happen for each student.

Tie in observations (When observations are Equal) :

When values of some observations of variable X or Y or both are equal then a problem arises of assigning ranks to such observations. When values of observations of X or Y are equal, we say that it is a tie. In such a case, average (mean) of corresponding ranks is assigned as rank to each of the repeated observation in the tie and the next observation will get the rank next to the last rank used in calculating the average rank. Let us understand it by taking one example. Suppose observations for one of the variables are 37, 60, 42, 78, 42, 50, 66, 42, 60. The highest observation is 78, so we assign rank 1 to it; next is 66 so we assign rank 2 to it. Now, the next observation is 60 which is repeated two times. So, the average of corresponding ranks (i.e. 3rd and 4th) which is $\frac{3+4}{2} = 3.5$ is assigned as the rank to each such observation (60). Now, the next observation is 50. Since 3rd and 4th ranks are already used, we will assign rank 5 to it. Now, the next observation is 42 and it is repeated three times. So, the average of corresponding ranks (i.e. 6th, 7th, 8th) which is $\frac{6+7+8}{3} = 7$ is assigned as the rank to each such observation (42). Finally, the last observation is 37. So, we assign rank 9 to it as the ranks 6, 7 and 8 are already used. In the same manner, ranks are also given to all observations of other variable.

Now, when a tie occurs (some observations are equal), correction is necessary in the formula of rank correlation coefficient. This correction is to be done by Correction Factor (CF).

To obtain 'CF', we add the factor $\left(\frac{m^3-m}{12}\right)$ to Σd^2 for each repeated observation group. Where m = Number of times an observation is repeated. CF is nothing but the sum of the terms $\left(\frac{m^3-m}{12}\right)$ obtained for such repeated observation groups. i.e. CF = $\Sigma\left(\frac{m^3-m}{12}\right)$

So, for tied ranks (when some observations are equal), the formula for obtaining rank correlation can be written as follows :

$$r = 1 - \frac{6\left[\Sigma d^2 + CF\right]}{n(n^2-1)}$$

Where Correction Factor (CF) = $\Sigma\left(\frac{m^3-m}{12}\right)$

and m = Number of times an observation is repeated.

Illustration 21 : Calculate the rank correlation coefficient using the data given in illustration 5.

We shall assign the ranks to both the variables as per our earlier discussion.

The observation 75 for the variable Y is repeated two times. The highest observation is 79. So, after assigning the rank 1 to it, we shall assign the average rank of rank 2 and rank 3. i.e. $\frac{2+3}{2} = 2.5$ to each of the following observation 75.

Now, let us prepare a table as follows.

	Reading (hours) x	**Rank of x R_x**	**Marks y**	**Rank of y R_y**	$d = R_x - R_y$	d^2
	25	7	65	7	0	0
	38	2	75	2.5	–0.5	0.25
	30	5	68	6	–1	1
	28	6	70	5	1	1
	34	4	72	4	0	0
	40	1	79	1	0	0
	36	3	75	2.5	0.5	0.25
Total	–	–	–	–	**0**	**2.5**

The calculation for obtaining 'CF' is as follows.

Repeated Observation	**No. of times the observation is repeated (m)**	$\left(\frac{m^3-m}{12}\right)$
75	2	$\left(\frac{2^3-2}{12}\right) = 0.5$
–	–	**CF** $= \Sigma\left(\frac{m^3-m}{12}\right) = \mathbf{0.5}$

$$r = 1 - \frac{6\left[\Sigma d^2 + CF\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[2.5 + 0.5\right]}{7(49-1)}$$

$$= 1 - \frac{6(3)}{336}$$

$$= 1 - \frac{18}{336}$$

$$= 1 - 0.0536$$

$$= 0.9464$$

$$\therefore \quad r \simeq 0.95$$

Note : It can be seen here that the value of r obtained by spearman's rank correlation method differs from the value obtained by Karl Pearson's method in illustration 5.

Illustration 22 : To know the relation between understanding of students in the subject of Economics and their dancing skill, a sample of eight students is taken and a test is conducted for them. The marks obtained are given below. Find the rank correlation coefficient between the marks obtained in two subjects.

Marks in Economics	60	30	10	20	30	50	30	40
Marks in dancing skill	80	20	60	40	12	28	20	15

If we assign ranks according to the marks in Economics then the highest marks is 60. So, its rank is 1, rank 2 for the marks 50 and rank 3 for the marks 40. Now, the next is marks 30 which is repeated three times. So, the average rank of the corresponding ranks (rank 4, rank 5, rank 6) i.e. $\frac{4+5+6}{3} = 5$ will be the rank for the marks 30. Now, after mark 30, it is marks 20. So, its rank will be 7 and finally for the lowest marks 10, its rank will be 8. Similarly, if we assign ranks according to the marks in dancing skill then the highest marks is 80. So, its rank will be 1, rank 2 for marks 60, rank 3 for marks 40, rank 4 for marks 28. Now, marks 20 is repeated two times. So, the average rank of the corresponding ranks (rank 5, rank 6) i.e. $\frac{5+6}{2} = 5.5$ will be the rank for the marks 20. Now, next is marks 15, so its rank will be 7 and finally the lowest marks 12 will have rank 8.

Now, let us prepare a table as follows.

	Marks in Economics x	**Rank of X R_x**	**Marks in dancing skill y**	**Rank of Y R_y**	**$d = R_x - R_y$**	**d^2**
	60	1	80	1	0	0
	30	5	20	5.5	–0.5	0.25
	10	8	60	2	6	36
	20	7	40	3	4	16
	30	5	12	8	–3	9
	50	2	28	4	–2	4
	30	5	20	5.5	–0.5	0.25
	40	3	15	7	–4	16
Total	–	–	–	–	**0**	**81.5**

The calculation of Correction Factor (CF) is as follows.

Repeated observation	**No. of times the observation is repeated (m)**	$\left(\frac{m^3-m}{12}\right)$
30	3	$\left(\frac{3^3-3}{12}\right)=2$
20	2	$\left(\frac{2^3-2}{12}\right)=0.5$
–	–	**CF = 2.5**

$$r = 1-\frac{6\left[\Sigma d^2+CF\right]}{n(n^2-1)}$$

$$= 1-\frac{6\left[81.5+2.5\right]}{8(64-1)}$$

$$= 1-\frac{6(84)}{504}$$

$$= 1-\frac{504}{504}$$

$$= 1-1$$

$$\therefore\ r = 0$$

Here, $r = 0$. So, it can be said that there is a lack of linear correlation between the ranks of the marks in both the subjects. i.e. the performances of the students of the given group in the subjects of Economics and dancing skill are independent with reference to the linear relationship.

Illustration 23 : An agency marketing electrical appliances wants to know the relation between the sales and the profits of LED fittings. The following information is obtained about the sales and profits of different electric companies. Find the rank correlation coefficient between the sales (in thousand units) and the profits (in lakh ₹).

Sales (thousand units)	25	58	215	72	58	25	90	162
Profit (lakh ₹)	65	140	500	115	65	65	220	340

Here, $n = 8$ and the observations 25 and 58 in sales are given twice and the observation 65 in profit is given thrice. So, assigning the ranks to both the variables, sales and profit, in the same manner as discussed in the previous example, we prepare a table as follows.

	Sales (thousand units) x	Rank of X R_x	Profit (lakh ₹) y	Rank of Y R_y	$d = R_x - R_y$	d^2
	25	7.5	65	7	0.5	0.25
	58	5.5	140	4	1.5	2.25
	215	1	500	1	0	0
	72	4	115	5	–1	1
	58	5.5	65	7	–1.5	2.25
	25	7.5	65	7	0.5	0.25
	90	3	220	3	0	0
	162	2	340	2	0	0
Total	–	–	–	–	**0**	**6**

The calculation of 'CF' is as follows.

Repeated Observation	No. of times the observation is repeated (m)	$\left(\frac{m^3-m}{12}\right)$
25	2	$\left(\frac{2^3-2}{12}\right) = 0.5$
58	2	$\left(\frac{2^3-2}{12}\right) = 0.5$
65	3	$\left(\frac{3^3-3}{12}\right) = 2$
–	–	**CF = 3**

$$r = 1 - \frac{6\left[\Sigma d^2 + CF\right]}{n(n^2-1)}$$

$$= 1 - \frac{6[6+3]}{8(64-1)}$$

$$= 1-\frac{54}{504}$$

$$= 1-0.1071$$

$$= 0.8929$$

$$\therefore \quad r \simeq 0.89$$

The value of r is near to 1. So, it can be said that there is a high degree of positive correlation between the sales and profit.

Note :

(1) The sum of difference in the ranks R_x and R_y is always zero. i.e. $\Sigma d = \Sigma (R_x - R_y) = 0$

(2) If $R_x = R_y$ for each pairs of the observations of two variables x and y then all corresponding values of d will be zero and hence $\Sigma d^2 = 0$. In this case, the value of r will be 1.

(3) If the ranks R_x and R_y are in exact reverse order of each other (see illustration 18) then $r = -1$.

Activity

Collect the information regarding the marks obtained by any ten students of your class in the subjects of Statistics and Economics. Find the correlation coefficient between the marks of two subjects using Karl Pearson's and Spearman's method and compare them.

Illustration 24 : A transport company wants to know the relation between driving experience and the number of accidents by the drivers. The sum of squares of differences in the ranks given to driving experience and the number of accidents by eight drivers is found to be 126. Find the rank correlation coefficient.

Here, $n = 8$ and the sum of squares of difference in the ranks is 126, i.e. $\Sigma d^2 = 126$.

$$r = 1-\frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1-\frac{6(126)}{8(64-1)}$$

$$= 1-\frac{756}{504}$$

$$= 1-1.5$$

$$r = -0.5$$

Illustration 25 : Ten students selected from various schools of a district were ranked on the basis of their proficiency in Sports and General knowledge. The rank correlation coefficient obtained from the data was found to be 0.2. Later on, it was noticed that the difference in the ranks of the two attributes for one of the students was taken as 3 instead of 2. Find the correct value of rank correlation coefficient.

Here, $n = 10$

Incorrect $d = 3$

Correct $d = 2$

$$\text{Now,} \quad r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$\therefore \quad 0.2 = 1 - \frac{6\Sigma d^2}{10(100-1)}$$

$$\therefore \quad 0.2 = 1 - \frac{6\Sigma d^2}{990}$$

$$\therefore \quad \frac{6\Sigma d^2}{990} = 1 - 0.2$$

$$\therefore \quad \frac{6\Sigma d^2}{990} = 0.8$$

$$\therefore \quad \Sigma d^2 = \frac{0.8 \times 990}{6}$$

$$\therefore \quad \Sigma d^2 = 132$$

Since one difference 2 is wrongly taken as 3, the corrected value of Σd^2 is obtained as follows :

$$\text{Corrected } \Sigma d^2 = 132 - (\text{Wrong } d)^2 + (\text{Correct } d)^2$$

$$= 132 - 3^2 + 2^2$$

$$= 132 - 9 + 4$$

$$= 127$$

$\therefore$ Correct value of the rank correlation coefficient is obtained as follows :

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(127)}{10(100-1)}$$

$$= 1 - \frac{762}{990}$$

$$= 1 - 0.7697$$

$$= 0.2303$$

$$\therefore r \simeq 0.23$$

Merits and Limitations of Spearman's Rank Correlation Method

Merits :

(1) This method is easy to understand.

(2) The calculation of rank correlation method is easier than that of Karl Pearson's correlation coefficient.

(3) It is the only method when the data is qualitative.

(4) When dispersion is more or when the extreme observations are present in the data, Spearman's formula is preferred over Karl Pearson's formula.

Limitations :

(1) Since the ranks are used instead of the actual observations, there is always a loss of some information. So, this method does not provide accurate value of the correlation coefficient as compared to Karl Pearson's method.

(2) Unless the ranks are given, it is tedious to assign ranks when the number of observations is large.

(3) This method can not be used for a bivariate frequeny distribution. (In such a case, Karl Pearson's method is used and you will learn it in your higher studies.)

Exercise 2.3

1. Six companies are ranked by the two market analysts on the basis of their growth in the recent past.

Company	A	B	C	D	E	F
Rank by Analyst 1	5	2	1	4	3	6
Rank by Analyst 2	6	4	3	2	1	5

Find the rank correlation coefficient between the evaluation given by two analysts.

2. An official has ranked nine villages of a sample on the basis of the work done in the area of 'Swachhata Abhiyan' and 'Beti Bachavo Abhiyan' by the villages. The ranks are given below.

Village	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Rank for Swachhata Abhiyan	4	8	7	1	9	5	6	2	3
Rank for Beti Bachavo Abhiyan	6	8	5	1	9	7	3	4	2

Find the rank correlation coefficient between the performances of the villages in two Abhiyans.

3. The following information is obtained by a survey conducted by a town planning committee of a state.

City	A	B	C	D	E
Population (lakh)	57	45	14	18	8
Rate of growth (per thousand)	13	20	10	15	5

Find the rank correlation coefficient between the population of the cities and the rate of growth of the population.

4. The following information is obtained by taking a sample of ten students from the students of a Science college.

Student	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	39	65	62	90	82	75	25	98	36	78
Marks in Statistics	47	53	58	86	62	68	60	91	51	84

Find the rank correlation coefficient between the ability of the students in the subjects of Mathematics and Statistics.

5. From the following information of heights of husband and wife, calculate the rank correlation coefficient between their heights.

Height of husband (cms)	156	153	185	157	163	191	162
Height of wife (cms)	154	148	162	157	162	170	154

6. Two interviewers gave the following scores to the candidates on the basis of their performance in the interview. Find the rank correlation coefficient between the evaluation of two interviewers.

Candidate	A	B	C	D	E	F	G	H
Marks by first interviewer	28	44	10	28	47	35	19	40
Marks by second interviewer	32	45	25	32	41	32	24	38

7. Ten contestants are ranked in a beauty contest by two judges and the sum of squares of differences in their ranks is found to be 214. Find the rank correlation coefficient.

8. The coefficient of rank correlation of the marks obtained by 10 students in two particular subjects was found to be 0.5. Later on, it was found that one of the differences of the ranks of a student was 7 but it was taken as 3. Find the corrected value of the correlation coefficient.

*

2.9 Precautions in the Interpretation of Correlation Coefficient

The coefficient of correlation measures the strength of linear relationship between two variables. An erroneous interpretation of r may lead us to a misunderstanding about the relationship between two variables. The following are some of the points to be kept in mind as a precaution :

(1) Correlation is only a measure of strength of linear relationship between two variables. It gives no indication about presence of cause and effect relationship between them and it does not give any idea about the information that out of the two, which variable is the dependent (effect) and the other as independent (cause). The interpretation of the correlation coefficient depends very much on experience. The investigator must have thorough knowledge about the variables under consideration and the various factors which affect these variables. Several examples can be cited indicating no meaningful correlation between two variables though the value of $|r|$ is very near to 1. Generally, it happens when r is calculated without prior knowledge about cause and effect relationship between the variables. For example, the two series of data relating to the number of persons died in road accidents in a city and the price of Tuver Dal during the same period may exhibit a high correlation (i.e. r may be near to 1). But there can not be meaningful relationship between them. Therefore, this kind of correlation is known as nonsense or spurious correlation.

(2) Sometimes, due to the presence of other factors, the value of $|r|$ between given two variables may be close to 1 though two variables are not correlated. For example, the data relating to the yield of rice and sugarcane show a fairly high degree of positive correlation though there is no connection between these two variables. This may be due to the favourable effect of external factors like weather conditions, irrigation system, fertilizers etc.

(3) When $r = 0$, we can merely say that there is no linear correlation. i.e. there is a lack of linear correlation. But there may be a non linear (quadratic or any other type) relationship between the variables. e.g. :

x	–4	–3	–2	–1	1	2	3	4
y	16	9	4	1	1	4	9	16

If we calculate the Karl Pearson's coefficient of correlation for the above example then the value of r will be 0. So, we may interpret that the two variables are uncorrelated but it is a wrong interpretation. If we observe the values of two variables X and Y then we can see that they have the relation $Y = X^2$. This relation is not linear but it is quadratic. So, though there is a perfect quadratic relationship between the two variables, we get $r = 0$. So, from this example we can understand that $r = 0$ suggests a lack of linear correlation only but there may be other kind of correlation.

(4) If the correlation coefficient computed from bivariate data which is related to a given region or class or given time duration then its interpretation should be limited to that region or class or time duration only. The interpretation of r computed from such data should not be extended or generalised outside the region or class or time duration without proper verification in order to avoid any kind of misunderstanding.

e.g. If a company starts manufacturing a new product and advertises it for its sale then initially by increasing the advertisement cost, sale of the product also increases when the quality of product is good. But after some time limit, sale of the product may not increase even if its advertisement cost increases. Normally there is high degree of positive correlation between the advertisement cost and sales. During initial production period. But after some time that may not be the case. So, the interpretation that there is a high degree of positive correlation between the advertisement cost and sales can not be applied for the data outside its time period.

Some Illustrations :

Illustration 26 : Determine the value of the correlation coefficient from the following results.

$Cov(x, y) : s_x^2 = 3:5$ **and** $s_x : s_y = 1:2$

Here, $Cov(x, y) : s_x^2 = 3:5$ $\quad\therefore\quad \dfrac{Cov(x, y)}{s_x^2} = \dfrac{3}{5}$

and $s_x : s_y = 1:2$ $\quad\therefore\quad \dfrac{s_x}{s_y} = \dfrac{1}{2}$

Now, $r = \dfrac{Cov(x, y)}{s_x\, s_y} = \dfrac{Cov(x, y)}{s_x^2} \times \dfrac{s_x}{s_y}$

$= \dfrac{3}{5} \times \dfrac{1}{2}$

$= \dfrac{3}{10}$

$\therefore\quad r = 0.3$

Illustration 27 : The following results are obtained from a bivariate data.

$n = 10$, $\Sigma(x-\bar{x})(y-\bar{y}) = 72$, $s_x = 3$ and $\Sigma(y-\bar{y})^2 = 160$ Find the correlation coefficient.

From the available results, first we shall find s_y.

$$s_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \sqrt{\frac{160}{10}} = \sqrt{16} = 4$$

Now, substituting the necessary values in the following formula,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{ns_x s_y}$$

$$= \frac{72}{10(3)(4)}$$

$$= \frac{72}{120}$$

$$\therefore \quad r = 0.6$$

Illustration 28 : An educationalist has conducted an experiment to know the relation between the usage of Social Media in mobile phone and the result of the examination. A group of 10 students is selected for this and the following results were obtained regarding, the time spent x (in hours) in last week on Social Media and the marks (y) obtained out of 50 in the examination, taken immediately after it.

$\Sigma x = 133$, $\Sigma y = 220$, $\Sigma x^2 = 2344$, $\Sigma y^2 = 6500$ and $\Sigma xy = 3500$

Later on, it was found that one of the pairs of observations of X and Y was taken as (13, 20) instead of (15, 25). Find the correct value of the correlation coefficient between X and Y.

Here, $n = 10$, $\Sigma x = 133$, $\Sigma y = 220$, $\Sigma x^2 = 2344$, $\Sigma y^2 = 6500$ and $\Sigma xy = 3500$

Incorrect Pair : (13, 20)

Correct Pair : (15, 25)

Now, we find corrected values of these measures as follows :

$\Sigma x = 133 - 13 + 15 = 135$

$\Sigma y = 220 - 20 + 25 = 225$

$\Sigma x^2 = 2344 - (13)^2 + (15)^2 = 2344 - 169 + 225 = 2400$

$\Sigma y^2 = 6500 - (20)^2 + (25)^2 = 6500 - 400 + 625 = 6725$

$\Sigma xy = 3500 - (13 \times 20) + (15 \times 25) = 3500 - 260 + 375 = 3615$

Substituting these corrected values in the following formula,

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{10(3615) - (135)(225)}{\sqrt{10(2400) - (135)^2} \cdot \sqrt{10(6725) - (225)^2}}$$

$$= \frac{36150 - 30375}{\sqrt{24000 - 18225} \cdot \sqrt{67250 - 50625}}$$

$$= \frac{5775}{\sqrt{5775} \cdot \sqrt{16625}}$$

$$= \frac{5775}{\sqrt{96009375}}$$

$$= \frac{5775}{9798.4374}$$

$$= 0.5894$$

$$\therefore \quad r \simeq 0.59$$

Illustration 29 : (1) If the correlation coefficient between two variables X and Y is 0.5, find the value of the following : (i) $r(x, -y)$ (ii) $r(-x, y)$ (iii) $r(-x, -y)$

Here, $r(x, y) = 0.5$

From the property no. 5 of correlation coefficient,

(i) $r(x, -y) = -r(x, y) = -0.5$

(ii) $r(-x, y) = -r(x, y) = -0.5$

(iii) $r(-x, -y) = r(x, y) = 0.5$

(2) If $r(x, y) = 0.8$ then find $r(u, v)$ for the following.

(i) $u = x - 10$ **and** $v = y + 10$

(ii) $u = \frac{x-5}{3}$ **and** $v = 2y + 7$

(iii) $u = \frac{2x-3}{10}$ **and** $v = \frac{10-y}{100}$

(iv) $u = \frac{5-x}{2}$ **and** $v = \frac{5+y}{2}$

(v) $u = \frac{20-x}{3}$ **and** $v = \frac{20-y}{7}$

While defining u and v from the properties (no. 4 and no. 5), the value of $r(u, v)$ will be dependent on the signs of the coefficients of X and Y.

i.e. $r(u, v) = r(x, y)$ or $-r(x, y)$

(i) $r(x - 10, y + 10) = r(u, v) = 0.8$

(ii) $r\left(\frac{x-5}{3}, 2y + 7\right) = r(u, v) = 0.8$

(iii) $r\left(\frac{2x-3}{10}, \frac{10-y}{100}\right) = r(u, v) = -0.8$

(iv) $r\left(\frac{5-x}{2}, \frac{5+y}{2}\right) = r(u, v) = -0.8$

(v) $r\left(\frac{20-x}{3}, \frac{20-y}{7}\right) = r(u, v) = 0.8$

Illustration 30 : A project is conducted by the group of the students of an MBA Institute to know the relation between the results of the final year of school and final year of graduation for the students. The following information is obtained from a sample of 10 students regarding the percentage of marks in standard 12 (x) and the percentage of marks in the final year of graduation (y).

$$n = 10, \Sigma(x-65) = -2, \ \Sigma(y-60) = 2, \ \Sigma(x-65)^2 = 176, \ \Sigma(y-60)^2 = 140, \Sigma(x-65)(y-60) = 141$$

Find the correlation coefficient between the percentages of marks in Standard 12 and the final year of graduation.

Here $\Sigma(x-65) = -2 \neq 0$ $\quad\therefore\ A = 65$

$\Sigma(y-60) = 2 \neq 0$ $\quad\therefore\ B = 60$

(Here, the sum of deviations are not zero, so $65 \neq \bar{x}$ and $60 \neq \bar{y}$)

Now, let us define $u = (x-65)$ and $v = (y-60)$.

So, $\Sigma(x-65) = \Sigma u = -2, \ \Sigma(y-60) = \Sigma v = 2$

$\Sigma(x-65)^2 = \Sigma u^2 = 176, \ \Sigma(y-60)^2 = \Sigma v^2 = 140$

$\Sigma(x-65)(y-60) = \Sigma uv = 141$

Substituting the above values in the following formula,

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{10(141) - (-2)(2)}{\sqrt{10(176) - (-2)^2} \cdot \sqrt{10(140) - (2)^2}}$$

$$= \frac{1414}{\sqrt{1756} \cdot \sqrt{1396}}$$

$$= \frac{1414}{\sqrt{2451376}}$$

$$= \frac{1414}{1565.6871}$$

$$= 0.9031$$

$$\therefore \quad r \simeq 0.90$$

Illustration 31 : To study the relation between the age (X years) of teenage children and their daily requirement of protein (y grams), the following information is obtained from a sample of 10 children taken by the Health Department of State.

$$\Sigma x = 140, \ \Sigma y = 150, \ \Sigma(x-10)^2 = 180, \quad \Sigma(y-15)^2 = 215, \Sigma(x-10)(y-15) = 60$$

Find the correlation coefficient between X and Y.

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{140}{10} = 14, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{150}{10} = 15$

We can see that the deviations are not taken from actual mean $(\bar{x} = 14)$ for the variable X. So, to solve the example, it will be convenient to define $u = (x - A) = (x - 10)$ and $v = (y - B) = (y - 15)$.

We are given the following information.

$\Sigma(x-10)^2 = \Sigma u^2 = 180,\ \Sigma(y-15)^2 = \Sigma v^2 = 215,\ \Sigma(x-10)(y-15) = \Sigma uv = 60$

Now, in order to use an appropriate formula of r, first we need Σu and Σv.

$\Sigma u = \Sigma(x-10) = \Sigma x - \Sigma 10 = \Sigma x - n(10) = 140 - 10(10) = 140 - 100 = 40$

$\Sigma v = \Sigma(y-15) = \Sigma y - \Sigma 15 = \Sigma y - n(15) = 150 - 10(15) = 150 - 150 = 0$

$\{\because\ \Sigma k = \underbrace{k + k + k + \ldots\ldots\ldots + k}_{n \text{ times}} = nk$ where, k = constant$\}$

Substituting the above values in the following formula,

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{10(60) - (40)(0)}{\sqrt{10(180) - (40)^2} \cdot \sqrt{10(215) - (0)^2}}$$

$$= \frac{600 - 0}{\sqrt{1800 - 1600} \cdot \sqrt{2150 - 0}}$$

$$= \frac{600}{\sqrt{200} \cdot \sqrt{2150}}$$

$$= \frac{600}{\sqrt{430000}}$$

$$= \frac{600}{655.7439}$$

$$= 0.9150$$

$\therefore\ r \simeq 0.92$

Illustration 32 : To know the relation between the ability in two different subjects for the students, a sample of seven students is taken from a school. From the information of marks in two subjects for 7 students, it is known that the sum of the squares of differences in the ranks of these marks is 25.5. It is also known that two students got equal marks in one subject and all the remaining marks are different. Find the rank correlation coefficient.

Here, $n = 7$ and $\Sigma d^2 = 25.5$

Two students got equal marks in a subject $(\therefore m = 2)$. So, we can say that there is a tie in assigning the ranks. Therefore, we need to take the term $\left(\frac{m^3 - m}{12}\right)$ only once to obtain CF.

$$CF = \left(\frac{m^3-m}{12}\right) = \left(\frac{2^3-2}{12}\right) = 0.5$$

$$r = 1-\frac{6\left[\Sigma d^2+CF\right]}{n(n^2-1)}$$

$$= 1-\frac{6[25.5+0.5]}{7(49-1)}$$

$$= 1-\frac{6(26)}{336}$$

$$= 1-\frac{156}{336}$$

$$= 1-0.4643$$

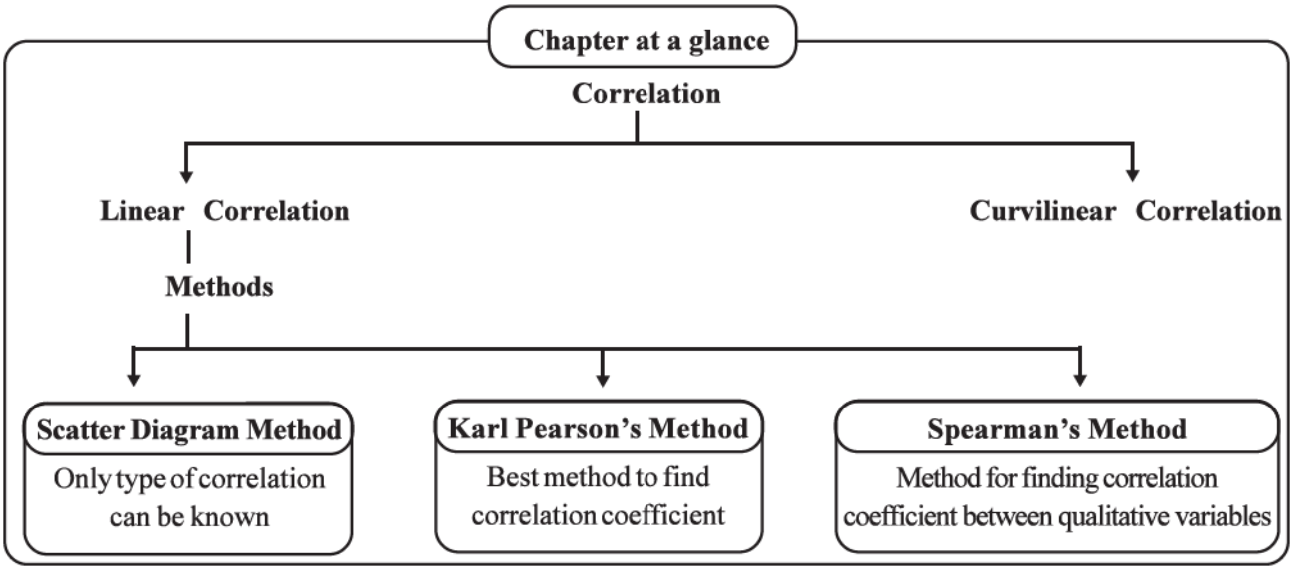
$$= 0.5357$$

$$\therefore \quad r \simeq 0.54$$

Summary

- **Correlation :** Simultaneous change in the values of two variables and direct or indirect cause-effect relationship between them.
- **Linear Correlation :** There are almost constant proportional changes in the values of two variables i.e. the points corresponding to the values of two correlated variables are on or nearer to a line.
- **Positive Correlation :** The changes in the values of two correlated variables are in the same direction.
- **Negative Correlation :** The changes in the values of two correlated variables are in the opposite direction.
- **Correlation Coefficient :** The numerical measure showing the strength of linear correlation between two variables is a correlation coefficient.
- **Scatter diagram :** A simple method for identifying linear correlation and its type (positive or negative).
- **Karl Pearson's Method :** The best method of obtaining type and strength of linear correlation using all observations.
- **Spearman's Rank Correlation Method :** A method for obtaining the correlation coefficient for qualitative variables and also preferable when dispersion is more in quantitative variables.
- The cause and effect relation between two variables cannot be proved but under the assumption that it does exist, the concept of correlation is studied.
- $r = 0$ indicates the absence of linear correlation only but there may be other type of correlation.

Chapter at a glance



List of Formulae

Karl Pearson's Method :

Correlation coefficient $= r$

(1) $r = \dfrac{\text{Covariance}}{(\text{S.D of } X)(\text{S.D of } Y)} = \dfrac{Cov(X,Y)}{s_x \cdot s_y}$

Where, $Cov(X,Y) = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \dfrac{\Sigma xy - n\bar{x}\,\bar{y}}{n}$

$s_x = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n}}$ and $s_y = \sqrt{\dfrac{\Sigma(y-\bar{y})^2}{n}}$

(2) $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2} \cdot \sqrt{\Sigma(y-\bar{y})^2}}$

(3) $r = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

(4) $r = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$ Where, $u = x - A$ or $\dfrac{x-A}{c_x}$, $v = y - B$ or $\dfrac{y-B}{c_y}$

(5) $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n \cdot s_x \cdot s_y}$

(6) $r = \dfrac{\Sigma xy - n\bar{x}\,\bar{y}}{n \cdot s_x \cdot s_y}$

} Specially for short sums

Spearman's Rank Correlation Method

(7) $r = 1 - \dfrac{6\Sigma d^2}{n(n^2-1)}$ When the observations are not repeated

(8) $r = 1 - \dfrac{6\left[\Sigma d^2 + CF\right]}{n(n^2-1)}$ When some of the observations are repeated

Where, d = Rank of x – Rank of y = $R_x - R_y$

CF = Correction Factor = $\Sigma\left(\dfrac{m^3-m}{12}\right)$

m = Number of times an observation is repeated

Exercise 2

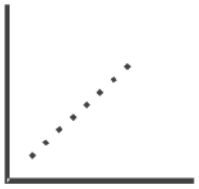
Section A

Find the correct option for the following multiple choice questions :

1. In context with correlation, what do you call the graph, if the points of paired observations (x,y) are shown in a graph ?

(a) Histogram (b) Circle diagram (c) Scatter diagram (d) Frequency curve

2. Which kind of the correlation exists if the following scatter diagram is of two variables X and Y ?



(a) Perfect Positive correlation (b) Partial Positive correlation

(c) Perfect Negative correlation (d) Partial Negative correlation

3. Which kind of the correlation exists if the following scatter diagram is of two variables X and Y ?

(a) Perfect Positive correlation (b) Partial Positive correlation

(c) Perfect Negative correlation (d) Partial Negative correlation

4. What is the value of r, if all the points plotted in a scatter diagram lie on a single line only ?

(a) 0 (b) 1 or -1 (c) 0.5 (d) -0.5

5. What is the range of the correlation coefficient r ?

(a) $-1 < r < 1$ (b) 0 to 1 (c) $-1 \leq r \leq 1$ (d) -1 to 0

6. The measurement unit of a variable 'Weight' is kg. and that of 'Height' is cm. What can you say about the measurement unit of the correlation coefficient between them ?

(a) kg (b) cm (c) km (d) does not have any unit

7. Which kind of the correlation can be obtained if the two variables are varying in opposite direction in constant proportion ?

(a) Partial Positive Correlation (b) Perfect Negative Correlation

(c) Perfect Positive Correlation (d) Partial Negative Correlation

8. What does the numerator indicate in the formula for calculating the correlation coefficient by Karl Perason's method ?

(a) Product of variance of X and Y (b) Covariance of X and Y

(c) Variance of X (d) Variance of Y

9. Which of the following values is not possible as a value of r ?

(a) 0.99 (b) -1.07 (c) -0.85 (d) 0

10. If $u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$, $c_x > 0$, $c_y > 0$ then which of the following statement is correct ?

(a) $r(x, y) \neq r(u, v)$ (b) $r(x, y) > r(u, v)$ (c) $r(x, y) = r(u, v)$ (d) $r(x, y) < r(u, v)$

11. If $r(x, y) = 0.7$ then what is the value of $r(x + 0.2, y + 0.2)$?

(a) 0.7 (b) 0.9 (c) 1.1 (d) –0.7

12. If $r(-x, y) = -0.5$ then what is the value of $r(x, -y)$?

(a) 0.5 (b) –0.5 (c) 1 (d) 0

13. What is the value of the rank correlation coefficient if $\Sigma d^2 = 0$?

(a) 0 (b) –1 (c) 1 (d) 0.5

14. In the method of rank correlation, in usual notations if $R_x = R_y$ for each pair of observations then what is the vlue of the r ?

(a) 0 (b) –1 (c) 1 (d) 0.1

15. In the method of rank correlation, what is the sum of differences of the ranks of two variables ?

(a) 0 (b) –1 (c) 1 (d) Any real number

16. In the method of rank correlation, if the ranks of two variables are exactly in reverse order then what is the value of r ?

(a) $r = 0$ (b) $r = -1$ (c) $r = 1$ (d) $r = 0.1$

17. In usual notations, which term is added in Σd^2 for each repeated observation in the rank correlation ?

(a) $\frac{m^2-1}{12}$ (b) $\frac{m^3-m}{12}$ (c) $\frac{6m^3-m}{12}$ (d) $n\left(n^2-1\right)$

18. Which kind of correlation will you get between the number of units sold and its revenue at constant price ?

(a) Perfect Positive (b) Partial Positive (c) Perfect Negative (d) Partial Negative

Section B

Answer the following questions in one sentence :

1. Define correlation.

2. Define correlation coefficient.

Identify, whether there is a positive correlation or negative correlation between the following pairs of variables (Question 3 to Question 6).

3. The age of an adult person and life insurance premium at the time of taking an insurance under a plan.

4. The sales and profit of last five years for a mostly accepted product of a company.
5. The rate of inflation and the purchase power of common man of a country when income of the common man is stable.
6. Altitude and amount of Oxygen in air.
7. What can be said about the correlation between the annual import of crude oil and the number of marriages during the same time period ?
8. The correlation coefficient between X and Y is 0.4. What will be the value of correlation coefficient if 5 is added in each observation of X and 10 is substracted from each observation of Y ?
9. What is the main limitation of scatter diagram method ?
10. If the value of $n\left(n^2-1\right)$ is six times the value of Σd^2 then what is the value of r ?
11. What will be the sign of r if the value of the covariance is negative ?

Section C

Answer the following questions :

1. Explain the meaning of positive correlation with an illustration.
2. Explain the meaning of negative correlation with an illustration.
3. Write the assumptions of Karl Pearson's method.
4. Define : Scatter Diagram.
5. What is spurious correlation ?
6. Explain the cause and effect relationship.
7. Explain : Perfect positive correlation
8. Explain : Perfect negative correlation
9. When is it necessary to use rank correlation ?
10. In which situation, the values of Karl Pearson's correlation coefficient and Spearman's rank correlation coefficient are equal ?
11. Find the value of r if $Cov(x, y)=120,\ s_x=12,\ s_y=15$.
12. Find the value of r if $\Sigma(x-\bar{x})(y-\bar{y})=-65,\ s_x=3,\ s_y=4$ and $n=10$.
13. For 10 pairs of observations, $\Sigma d^2=120$. Find the value of the rank correlation coefficient.

Section D

Answer the following questions :

1. Explain scatter diagram method.
2. Write merits and limitations of scatter diagram method.
3. Write the properties of correlation coefficient.
4. Write the merits and limitations of Karl Pearson's method.
5. Interpret $r = 1$, $r = -1$ and $r = 0$.
6. Explain Spearman's Rank correlation method.
7. Write merits and limitations of Spearman's rank correlation method.
8. How would you interpret partial correlation ?
9. State the necessary precautions to be taken while interpreting the value of correlation coefficient.
10. The following data is available for two variables rainfall in mm. (X) and yield of crop Qtl/ Hectare (Y).

    $n = 10$, $\bar{x} = 120$, $\bar{y} = 150$, $s_x = 30$, $s_y = 40$ and $\Sigma xy = 189000$. Find the correlation coefficient.
11. The following information is obtained for 9 pairs of observations.

    $\Sigma x = 51$, $\Sigma y = 72$, $\Sigma x^2 = 315$, $\Sigma y^2 = 582$, $\Sigma xy = 408$. Find the correlation coefficient.
12. The information obtained on the basis of ranks given by two judges to eight contestants of a dance competition is given below.

    $\Sigma(R_x - R_y)^2 = 126$

    Where R_x and R_y are the ranks given to a contestant by the two judges respectively. Find Spearman's rank correlation coefficient.
13. The ranks given by two experts on the basis of interviews of five candidates for a job are (3, 5), (5, 4), (1, 2), (2, 3) and (4, 1). Find the rank correlation coefficient from this data.

Section E

Solve the following :

1. The following information is obtained to study the relation between the selling price of nose mask and its demand during an epidemic.

Price (₹)	38	45	40	42	35
Demand (units)	103	92	97	98	100

Find the correlation coefficient between the price and demand of mask by Karl Pearson's method.

2. In order to study the relationship between the abilities in the subjects of Human Resource Management and Personality Development for the students of a post graduate level course, a sample of 5 students is taken and the following information is obtained.

Student	1	2	3	4	5
Marks in HRM	45	25	40	20	45
Marks in PD	47	23	17	35	48

Calculate the Karl Pearson's correlation coefficient between the marks of both the subjects.

3. A vendor wants to display lipsticks of different brands according to their popularity. For that, he invites two experts Preyal and Nishi to rank the lipsticks of different brands.

Lipstick	A	B	C	D	E	F	G
Rank by Preyal	5	6	7	1	3	2	4
Rank by Nishi	5	7	6	2	1	4	3

Find the rank correlation coefficient to know the smilarity in the decision of both the experts.

4. A merchant wants to study the relation between prices of tea and coffee in Ahmedabad city. He obtains the following information about prices of tea and coffee of the last six months.

Price per kg for tea (₹)	340	370	450	320	300	360
Price per 100 grams for coffee (₹)	190	215	200	180	163	175

Calculate the rank correlation coefficient between the price of tea and coffee.

5. The demand of an imported fruit in a local market is very uncertain. To know the relation between the price of the fruit and its supply, a vendor collects the information about the average price and supply for last ten months.

Average price per unit (₹)	65	68	43	38	77	48	35	30	25	50
Supply (hundred units)	52	53	42	60	45	41	37	38	25	27

Find the rank correlation between the average price and the supply.

6. To know the relation between the results of the examinations taken in a span of short time, a teacher has conducted two examinations in last two weeks and the ranks obtained by seven students are as follows.

Student	A	B	C	D	E	F	G
Rank in Test 1	5	1	2	3.5	3.5	7	6
Rank in Test 2	7	1	4	6	5	3	2

Find the rank correlation coefficient to know the similarity between the results of two examinations.

Section F

Solve the following :

1. The information of fertilizer used (in tons) and productivity (in tons) of eight districts is given below.

Fertilizer (tons)	15	18	20	25	29	35	40	38
Productivity (tons)	85	93	95	105	115	130	140	145

Calculate the correlation coefficient by Karl Pearson's method.

2. Find the Karl Pearson's correlation coefficient from the following information of the average weekly hours spent on Video games and the grade points obtained in an examination by 6 children of a big city.

Weekly average hours spent for Video games	43	47	45	50	40	51
Grade points obtained in an examination	5.2	4.9	5.0	4.7	5.4	4.3

3. Find Karl Pearson's correlation coefficient between density of population (per square km) and death rate (per thousand) from the following data.

City	A	B	C	D	E	F	G
Density (per sq. km)	750	600	350	500	200	700	850
Death rate (per thousand)	30	20	15	20	10	25	50

4. The following information is obtained to study the relationship between the advertisement cost and the sales of electric fans of the companies manufacturing electric fans. Find the correlation coefficient between advertisement cost and the sales by Karl Pearson's method.

Company	A	B	C	D	E	F
Advertisement Cost (lakh ₹)	140	120	80	100	80	180
Sales of electric fans (crore ₹)	35	45	15	40	20	50

5. A doctor obtains the following information for the weights of seven mothers and their children from a maternity home for his research to know the relation between the weights of mother and weights of their children at the time of birth.

Weight of mother (kg)	59	72	66	64	77	66	60
Weight of child (kg)	2.5	3.4	3.1	2.7	2.8	2.3	3.0

Find rank correlation coefficient between the weights of mother and child.

6. The following data is obtained to know the relation between maximum day temperature and the sale of ice-cream in Ahmedabad city.

Maximum Temperature (Celsius)	35	42	40	39	44	40	45	40
Sale of ice cream (kg)	600	680	750	630	920	750	900	720

Calculate the rank correlation coefficient.

7. An entrance test required to study abroad is conducted online. The marks obtained in Reasoning Ability and English Speaking in this online test (having negative marking system for wrong answer) by 5 students selected in a sample are given below.

Student	A	B	C	D	E
Marks in Reasoning Ability	5	5	5	5	5
Marks in English Speaking	2	–2	–2	0	2

Find the rank correlation coefficient between Reasoning Ability and ability in English Speaking.

8. Six dancers A, B, C, D, E and F in a dance competition were judged by two dance Gurus. The ranks assigned to the dancers are as follows.

Rank	1	2	3	4	5	6
By Guru 1	B	F	A	C	D	E
By Guru 2	F	A	C	B	E	D

Find the rank correlation coefficient between the judgement of the two Gurus.

9. The following data is obtained for two variables, inflation (X) and interest rate (Y).

$n = 50,\ \Sigma x = 500,\ \Sigma y = 300,\ \Sigma x^2 = 5450,\ \Sigma y^2 = 2000,\ \Sigma xy = 3090$

Later on, it was known that one pair of observation (10, 6) was included additionally by mistake. Find the correlation coefficient by excluding this pair of observations.

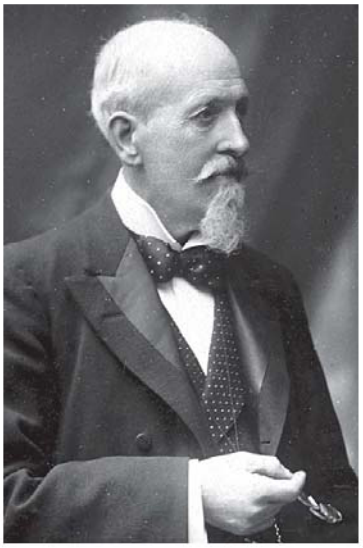
10. The information regarding sales (X) and expenses (Y) of 10 firms is given below.

$\bar{x} = 58,\ \bar{y} = 14,\ \Sigma(x-65)^2 = 850,\ \Sigma(y-13)^2 = 32,\ \Sigma(x-65)(y-13) = 0$

Find the correlation coefficient.

11. Daily calorie intake of ten persons is X and their weight is Y kg. The rank correlation coefficient from this information is 0.6. On subsequent verification, it was noticed that the difference of ranks of X and Y for one of the persons was taken as 2 instead of 4. Find the correct value of rank correlation coefficient.

12. The information of health index x and life expectancy y is obtained for 10 people. These data are ranked to find the rank correlation coefficient and the sum of squares of the ranks was found to be 42.5. It was also observed that health index 70 was repeated three times and life-expectancy 45 was repeated twice in the data. Find the rank correlation coefficient using this information.



Charles Edward Spearman
(1863 –1945)

Charles Edward Spearman was an English psychologist known for work in statistics, as a pioneer of factor analysis and for Spearman's rank correlation coefficient. He also did seminal work on models for human intelligence, including his theory that disparate cognitive test scores reflect a single General intelligence factor and coining the term g-factor.

After serving army for 15 years, he went on to study for a Ph.D. in experimental psychology. Spearman joined University College London and stayed there until he retired in 1931. Initially he was Reader and head of the small psychological laboratory. In 1911 he was promoted to the Grote professorship of the Philosophy of Mind and Logic. His title changed to Professor of Psychology in 1928 when a separate Department of Psychology was created.

His many published papers cover a wide field, but he is especially distinguished by his pioneer work in the application of mathematical methods to the analysis of the human mind and his original studies of correlation in this sphere.

"Prediction is very difficult, especially about the future."

– Niels Bohr

3

Linear Regression

Contents :

3.1 Introduction

3.2 Linear Regression Model

3.3 Fitting of Regression Line

3.3.1 Method of Scatter Diagram

3.3.2 Method of Least Squares

3.4 Utility of the study of Regression

3.5 Regression Coefficient from Covariance and Correlation Coefficient

3.6 Coefficient of Determination

3.7 Properties of Regression Coefficient

3.8 Precautions while using Regression

3.1 Introduction

In the previous chapter 2, we have studied the concept of correlation. We have seen whether the correlation between two variables is positive or negative is known by the correlation coefficient. Moreover, we get numerical measure of the closeness of the variables. But the coefficient of correlation fails to provide the expected value of one variable for the given value of the other variable. When some relation exists between two variables, many times it is necessary to obtain the approximate or estimated value of one variable for a known value of the other variable using this relation.

e.g. We know that there is a correlation between advertisement cost and sale of an item. Now, for some given amount of advertisement cost, if we want to know the corresponding expected sale then it is not possible to obtain it only by correlation. For this, it is necessary to use the concept of regression.

The literal meaning of regression is 'to avert' or 'return to the mean value'. The term regression was first used by a statistician Sir Francis Galton during his study of human inheritance. He had collected the information about the height of 1000 pairs of fathers and sons. He revealed the following interesting results.

(i) Tall fathers have tall sons and short fathers have short sons.

(ii) The average height of sons of a group of tall fathers is less than average height of group of tall fathers.

(iii) The average height of sons of a group of short fathers is greater than average height of group of short fathers.

So, it is clear from the above findings that the heights of sons show regressive tendency with respect to the height of their fathers. The existence of this tendency restricts the humans to split into two races of pigmies and giants. So, Sir Francis Galton has given the name regression to describe such relation.

Regression is a functional relation between two correlated variables. We shall study the concept of regression under the assumption that there exists a cause-effect relationship between two variables.

3.2 Linear Regression Model

A set of one or more equations representing a relation or a problem is called a model. A statistical model which describes the cause and effect relationship between two variables is called a regression model. Generally, out of two variables having cause-effect relationship, the causal variable is denoted by X. We shall call this variable as independent or explanatory variable and effect variable is denoted by Y. We shall call this variable as dependent or explained variable. Let us understand the meaning of independent variable and dependent variable from the following illustrations :

(i) In case of 'advertisement cost' and 'sales', generally, because of increase (decrease) in the 'advertisement cost', corresponding 'sales' also increases (decreases), so we shall take 'advertisement cost' as independent variable X and 'sales' as dependent variable Y.

(ii) In case of 'rainfall' and 'yield of rice' in some region, it is very clear that 'yield of rice' depends on 'rainfall'. So, we shall take 'rainfall' as independent variable X and 'yield of rice' as dependent variable Y.

In a regression model, the dependent variable Y is expressed in the form of an appropriate mathematical function of the independent variable X.

Now, we shall define a linear regression model as follows.

$$Y = \alpha + \beta X + u$$

Where, Y = Dependent Variable

X = Independent Variable

α = Constant

β = Constant

u = Disturbance Variable of the Model

The inadequacy of the linearity between two variables X and Y is shown by u. The perfect linear relation is possible in natural science like mathematics. So, the disturbance variable u obviously becomes 0 in such a case. In other words, when there is a perfect linear correlation between two variables X and Y then the regression model is $Y = \alpha + \beta X$. But we know that exact linear relation between the variables is not always possible in business, economics and social science as these correlated variables are also affected by other factors. Thus, when there is a partial correlation between the variables X and Y then the linear regression model is $Y = \alpha + \beta X + u$. From the above discussion, we can define linear regression in simple words as follows.

"A mathematical or functional relationship between two correlated variables which helps in estimating the value of dependent variable for some given (known) value of independent variable is called **Linear Regression.**''

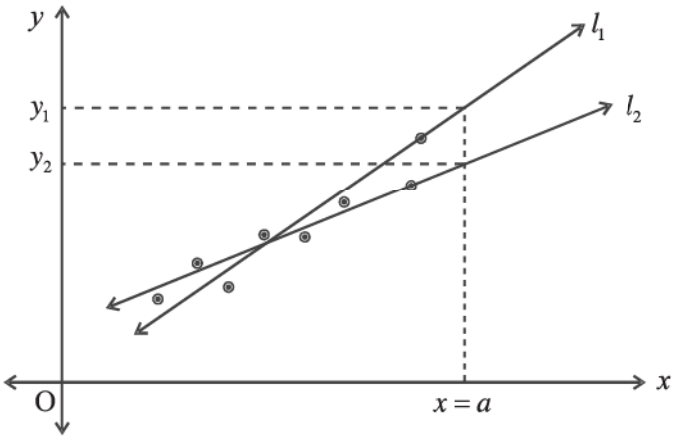
3.3 Fitting of Regression Line

In a scatter diagram of two correlated variables, if the points are clustered around a line, we can say that there is a linear regression. The method of obtaining such a line expressing the relation between two variables is called fitting of a regression line.

There are two methods for fitting a regression line : (1) Method of Scatter Diagram (2) Method of Least Squares.

3.3.1 Method of Scatter Diagram

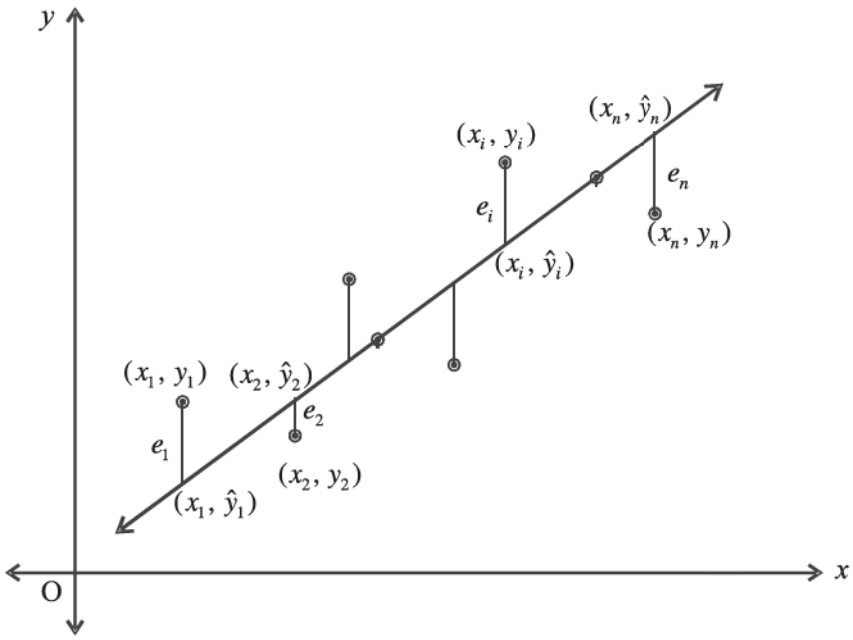
Suppose n ordered pairs of observations of two correlated variables X and Y are $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Using this data, we draw a scatter diagram. Now, a line is drawn in such a way that it is close to almost all the points of the scatter diagram. If Y is a dependent variable and X is an independent variable then such a line is called regression line of Y on X and an approximate value of dependent variable Y can be obtained for any given value of independent variable X from it. Since no computation is required to draw such a line, it is very easy and quick method of fitting a regression line. But there is a problem in doing so. Different persons may draw different lines. As a result, different persons may provide different estimates of the dependent variable Y for the same value of independent variable X. It can be seen very easily from the following scatter diagram.



Two different persons have drawn two different lines l_1 and l_2 in the following scatter diagram of the same data. We can see that for some value 'a' of independent variable X, corresponding estimated value is 'y_1' from line l_1 and it is 'y_2' from the line l_2. Thus we get different estimates for dependent variable Y from different lines for a single value of independent variable X. So, it can be said that this method is subjective. A line of regression drawn by this method is not the best fitted line because it does not guarantee the best estimate of the dependent variable. The method of least square is used to obtain such a best fitted regression line.

3.3.2 Method of Least Squares

Suppose n ordered pairs of observations of two correlated variables X (independent variable) and Y (dependent variable) are $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. We shall draw a scatter diagram for this data to understand the method of least squares.



If an equation of the best fitted line describing the linear regression between the variables X and Y is $\hat{y} = a + bx$ then the constants a and b of this line can be obtained by the method of least squares as follows.

Let the estimated values of variable Y corresponding to values $x_1, x_2, x_3, ..., x_n$ of variable X are $\hat{y}_1, \hat{y}_2, \hat{y}_3, ... \hat{y}_n$ from the line and the corresponding observed values of Y are $y_1, y_2, y_3, ... y_n$ respectively. Now, for some $X = x_i$, estimated value of Y from the line is $\hat{y}_i = a + bx_i$. The vertical distance (i.e. distance parallel to Y-axis) between observed value y_i and the estimated value $\hat{y}_i$ is called an error in the estimation. It is denoted by e_i.

$$\therefore\ e_i = y_i - \hat{y}_i = y_i - (a + bx_i) = y_i - a - bx_i$$

Where, $i = 1, 2, 3, ..., n$

Obviously, the error will be positive for points above the line the error will be negative for points below the line and it will be zero for the points which are on the line.

Now, the values of constants a and b of the fitted line $\hat{y} = a + bx$ (known as regression line of Y on X) are obtained in such a way that the sum of the squares of the errors is minimum.

i.e. $\Sigma e_i^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - a - bx_i)^2$ is minimum.

By ignoring the suffix i for convenience, we can get such values of a and b by a simple algebraic method, which are as follows.

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

And

$$a = \bar{y} - b\,\bar{x}$$

The line $\hat{y} = a + bx$ obtained by this method is a line passing as close as possible to the points of scatter diagram. The sum of squares of the errors is minimised while obtaining the regression line. Therefore, this method is called **'method of least squares'.**

The value of b obtained by this method is called the regression coefficient of the regression line of Y on X. b is also called slope of the regression line and the constant a is called intercept of the regression line.

Interpretation of regression coefficient b

b = the estimated change in the value of Y for a unit change in the value of X.

i.e. when $b > 0$, it means that a unit increase in the value of independent variable X implies an estimated increase of b units in the value of dependent variable Y.

when $b < 0$, it means that a unit increase in the value of independent variable X implies an estimated decrease of $|b|$ units in the value of dependent variable Y.

Note that the regression line obtained by the method of least squares is also known as the line of best fit.

Note : (1) The regression coefficient b can also be denoted by b_{yx}. If not required, generally we shall denote regression coefficient by b only.

(2) If all the points in a scatter diagram are on one line only then error will be zero for all the points. Hence, the estimated value $\hat{y}$ is same as its observed value y. So, the form of the regression line will be $y = a + bx$ in place of $\hat{y} = a + bx$. Naturally, in this situation r is 1 if $b > 0$ and r is -1 if $b < 0$.

Additional Information for understanding

Generally, only 'fitted line' is mentioned for the regression line obtained instead of 'best fitted line'.

Now, let us take some examples to obtain a regression line.

Illustration 1 : The following observations are obtained for life (years of usage) of cars and their average annual maintenance costs of a specific model of car of a particular company.

Life of cars (years)	2	4	6	8
Average annual maintenance cost (thousand ₹)	10	20	25	30

Obtain the regression line of maintenance cost on the life of cars. Also, estimate the maintenance cost if the life of a car is 10 years.

'Life of a car' is an independent variable. So, we shall denote it by variable X and 'maintenance cost' is dependent variable. So, we shall denote it by Y. Considering at the data, we shall prepare the following table for obtaining the regression line.

	Life of car (years) x	**Maintenance cost (thousand ₹) y**	xy	x^2
	2	10	20	4
	4	20	80	16
	6	25	150	36
	8	30	240	64
Total	**20**	**85**	**490**	**120**

$\bar{x} = \frac{\Sigma x}{n} = \frac{20}{4} = 5, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{85}{4} = 21.25$

Let us find the regression coefficient as follows.

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{4(490) - (20)(85)}{4(120) - (20)^2}$$

$$= \frac{1960 - 1700}{480 - 400}$$

$$= \frac{260}{80}$$

$$= 3.25$$

$\therefore \quad b = 3.25$

Now, putting the values of $\bar{x}, \bar{y}$ and b in the formula of a,

$$a = \bar{y} - b\,\bar{x}$$

$$= 21.25 - 3.25(5)$$

$$= 21.25 - 16.25$$

$\therefore \quad a = 5$

So, the regression line of 'maintenance cost' (Y) on 'life of car' (X) is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = 5 + 3.25\,x$

Putting $X = 10$,

$$\hat{y} = 5 + 3.25(10)$$

$$= 5 + 32.5 = 37.5$$

$\therefore \quad \hat{y} = 37.5$

So, when the life of a car is 10 years then its estimated maintenance cost is ₹ 37.5 thousand **Note :** Since $b = 3.25$, we can say that every year (one unit change in X), the maintenance cost of the car increases by approximately ₹ 3.25 thousand (change in Y).

Illustration 2 : The monthly sale of different types of laptops (in hundred units) and its profit (in lakh ₹) for the last six months for a company is given below.

Month	1	2	3	4	5	6
No. of laptops sold (hundred units) x	5	7	5	12	8	3
Profit (lakh ₹) y	8	9	10	15	10	6

Obtain the regression line of Y on X. Also find the error in estimating Y for X = 7.

	No. of laptops sold (hundred units) x	Profit (lakh ₹) y	xy	x^2
	5	8	40	25
	7	9	63	49
	5	10	50	25
	12	15	180	144
	8	10	80	64
	3	6	18	9
Total	**40**	**58**	**431**	**316**

$\bar{x} = \frac{\Sigma x}{n} = \frac{40}{6} = 6.67$; $\bar{y} = \frac{\Sigma y}{n} = \frac{58}{6} = 9.67$

Let us find the regression coefficient b as follows.

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{6(431) - (40)(58)}{6(316) - (40)^2}$$

$$= \frac{2586 - 2320}{1896 - 1600}$$

$$= \frac{266}{296}$$

$$= 0.8986$$

$$\simeq 0.90$$

$\therefore b \simeq 0.90$

By putting the values of $\bar{x}, \bar{y}$ and b in the formula of a,

$a = \bar{y} - b\bar{x}$

$= 9.67 - 0.90(6.67)$

$= 9.67 - 6.003$

$= 3.667$

$\therefore a \simeq 3.67$

So, regression line of Y on X is

$\hat{y} = a + bx$

$\therefore \hat{y} = 3.67 + 0.9x$

Now, to find the error for $X = 7$, first we obtain the estimated value of Y corresponding to it.

Putting $X = 7$,

$\hat{y} = 3.67 + 0.9(7)$

$= 3.67 + 6.3$

$\therefore\ \hat{y}$ = ₹ 9.97 lakh

Now, we can see from the available data that observed value of Y when $X = 7$ is 9.

$\therefore$ Error $e = y - \hat{y}$

$= 9 - 9.97$

$\therefore\ e$ = ₹ -0.97 lakh

Illustration 3 : In order to study the relationship between the repairing time of accident damaged cars and the cost of repair, the following information is collected.

Repairing time of a car (man hours)	32	40	25	29	35	43
Repairing cost (thousand ₹)	25	35	18	22	28	46

Obtain the regression line of Y (repairing cost) on X (repairing time). If the time taken to repair a car is 50 hours, find an estimate of the repairing cost.

Here, $n = 6, \bar{x} = \frac{\Sigma x}{n} = \frac{204}{6} = 34$ અને $\bar{y} = \frac{\Sigma y}{n} = \frac{174}{6} = 29$

	Repairing time (man hours) x	**Repairing cost (thousand ₹) y**	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	32	25	–2	–4	8	4
	40	35	6	6	36	36
	25	18	–9	–11	99	81
	29	22	–5	–7	35	25
	35	28	1	–1	–1	1
	43	46	9	17	153	81
Total	**204**	**174**	**0**	**0**	**330**	**228**

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$= \frac{330}{228}$

$= 1.4474$

$\simeq 1.45$

$\therefore\ b \simeq 1.45$

Now, by putting the value of $\bar{x}$, $\bar{y}$ and b in the formula of a,

$a = \bar{y} - b\bar{x}$

$= 29 - 1.45(34)$

$= 29 - 49.3$

$\therefore\ a = -20.3$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = -20.3 + 1.45x$$

Putting $X = 50$,

$$\hat{y} = -20.3 + 1.45(50)$$

$$= -20.3 + 72.5$$

$$\therefore \quad \hat{y} = 52.2$$

So, when the repairing time is 50 hours, the estimated repairing cost is ₹ 52.2 thousand.

Exercise 3.1

1. From the following data of price (in ₹) and demand (in hundred units) of a commodity, obtain the regression line of demand on price. Also estimate the demand when price is 20 ₹.

Price (₹)	12	14	15	16	18	21
Demand (hundred units)	18	12	10	8	7	5

2. To study the relationship between the time of usage of cars and its average annual maintenance cost, the following information is obtained :

Car	1	2	3	4	5	6
Time of usage of a car (years) x	3	1	2	2	5	3
Average annual maintenance cost (thousand ₹) y	10	5	8	7	13	8

Obtain the regression line of Y on X. Find an estimate of average annual maintenance cost when the usage time of a car is 5 years. Also find its error.

3. The information for a year regarding the average rainfall (in cm) and total production of crop (in tons) of five districts is given below :

Average rainfall (cm)	25	32	38	29	31
Crop (tons)	84	90	95	88	93

Find the regression line of production of crop on rainfall and estimate the crop if average rainfall is 35 cm.

4. The following data gives the experience of machine operators and their performance ratings.

Operator	1	2	3	4	5	6	7	8
Experience (years) x	12	5	10	3	18	4	12	16
Performance rating y	83	75	80	78	89	68	88	87

Calculate the regression line of performance ratings on the experience and estimate the performance rating of an operator having 7 years of experience.

*

3.4 Utility of the Study of Regression

The following are some utilities of regression :

(1) We can determine a functional relation between two correlated variables.

(2) Once the functional relation is established, it can be used to predict the unknown value of dependent variable Y on the basis of known value of independent variable X.

(3) We can determine the approximate change in the value of dependent variable Y for a unit change in the value of independent variable X.

(4) We can determine the error in the estimation of dependent variable obtained by a regression line.

Regression is very useful for economists, planners, businessmen, administrators, researchers, etc.

Short-cut Method for computing Regression coefficient

When the values of X and Y are relatively large and / or fractional, it is difficult to calculate the terms like x^2, xy, etc. In such cases, an alternative formula can be used. It is based on the following property of regression coefficient.

Property : The regression coefficient is independent of change of origin but not of change of scale.

If $b = b_{yx}$ is a regression coefficient of a regression line of Y on X then using the above property, the following formulae can be written for the regression coefficient by short-cut method.

(1) If $u = x - A$ and $v = y - B$ then

$$b = b_{yx} = b_{vu} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2}$$

(2) If $u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$ then

$$b = b_{yx} = b_{vu} \cdot \frac{c_y}{c_x} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

Here, A, B, c_x and c_y are constants and $c_x > 0, c_y > 0$.

Illustration 4 : In order to determine the relationship between monthly income (in thousand ₹) and monthly expenditure (in thousand ₹) of people of a group, a sample of seven persons is taken from that group and the following information is obtained.

Person	1	2	3	4	5	6	7
Monthly income (thousand ₹)	60	70	64	68	62	65	72
Monthly expenditure (thousand ₹)	50	59	57	50	53	58	60

Obtain the regression line of monthly expenditure on monthly income. If a person of the group has monthly income of ₹ 75 thousand, estimate his monthly expenditure.

Since the regression line of monthly expenditure on monthly income is to be obtained, we shall take 'monthly expenditure' as variable Y and 'monthly income' as variable X.

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{461}{7} = 65.86$ and $\bar{y} = \frac{\Sigma y}{n} = \frac{387}{7} = 55.29$

So, by taking $A = 65$ and $B = 55$, we can define u and v as follows.

$u = x - A = x - 65$ and $v = y - B = y - 55$

	Monthly income (thousand ₹) x	Monthly expenditure (thousand ₹) y	$u = x - 65$	$v = y - 55$	uv	u^2
	60	50	–5	–5	25	25
	70	59	5	4	20	25
	64	57	–1	2	–2	1
	68	50	3	–5	–15	9
	62	53	–3	–2	6	9
	65	58	0	3	0	0
	72	60	7	5	35	49
Total	**461**	**387**	**6**	**2**	**69**	**118**

b can be obtained by short-cut method as follows.

$$b = b_{yx} = b_{vu} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2}$$

$$= \frac{7(69) - (6)(2)}{7(118) - (6)^2}$$

$$= \frac{483 - 12}{826 - 36}$$

$$= \frac{471}{790}$$

$$= 0.5962$$

$$\therefore \quad b \simeq 0.60$$

Now, $a = \bar{y} - b\bar{x}$

$$= 55.29 - 0.60(65.86)$$

$$= 55.29 - 39.516$$

$$= 15.774$$

$$\therefore \quad a = 15.77$$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = 15.77 + 0.60x$$

Putting $X = 75$,

$$\hat{y} = 15.77 + 0.60(75)$$

$$= 15.77 + 45$$

$$= 60.77$$

$$\therefore \quad \hat{y} = 60.77$$

So, if a person has monthly income of ₹ 75 thousand, his approximate monthly expenditure is ₹ 60.77 thousand.

Illustration 5 : For the data given in illustration 1, obtain the regression line of maintenance cost (Y) on the life of cars (X) by using short-cut method.

Life of car (years) x	**2**	**4**	**6**	**8**
Annual maintenance cost (thousand ₹) y	**10**	**20**	**25**	**30**

All the values of X here are divisible by 2 and that of Y are divisible by 5. Moreover $\bar{x} = 5$ and $\bar{y} = 21.25$. So, we shall take $A = 4, B = 20, c_x = 2, c_y = 5$.

Now, let us define u and v as follows :

$$u = \frac{x-A}{c_x} = \frac{x-4}{2} \quad \text{and} \quad v = \frac{y-B}{c_y} = \frac{y-20}{5}$$

	x	y	$u = \frac{x-4}{2}$	$v = \frac{y-20}{5}$	uv	u^2
	2	10	−1	−2	2	1
	4	20	0	0	0	0
	6	25	1	1	1	1
	8	30	2	2	4	4
Total	**20**	**85**	**2**	**1**	**7**	**6**

$$b = b_{vu} \cdot \frac{c_y}{c_x} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

$$= \frac{4(7) - 2(1)}{4(6) - (2)^2} \times \frac{5}{2}$$

$$= \frac{28-2}{24-4} \times \frac{5}{2}$$

$$= \frac{26}{20} \times \frac{5}{2}$$

$$b = 3.25$$

Now, $a = \bar{y} - b\bar{x} = 21.25 - 3.25(5) = 21.25 - 16.25 = 5$

$\therefore$ The regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = 5 + 3.25x$$

Note : We can see that, $b_{vu} = \frac{26}{20} = 1.3$ but when it is multiplied by $\frac{c_y}{c_x} = \frac{5}{2}$ then we get $b = 1.3 \times \frac{5}{2} =$ 3.25 (as obtained in illustration 1). So, we can understand that when the scale of variable X and/or Y are changed, it is necessary to multiply b_{vu} by $\frac{c_y}{c_x}$ to obtain b.

Illustration 6 : A sample of seven students is taken from the students coming from abroad in the current year to study in university of the Gujarat State. The information regarding their I.Q. and the marks obtained in an examination of 75 marks is given below.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I.Q. x	85	95	100	90	110	125	70
Marks y	46	50	50	45	60	70	40

Obtain the regression line of Y on X and estimate the marks of a student whose I.Q. is 120. Also find the error in estimation when I.Q. is 100.

Here, $n = 7$, $\bar{x} = \frac{\Sigma x}{n} = \frac{675}{7} = 96.43$, $\bar{y} = \frac{\Sigma y}{n} = \frac{361}{7} = 51.57$

Since the values of X and Y are large, means are fractional and all the values of X are divisible by 5, we shall use short-cut method.

By taking $A = 95, B = 50, c_x = 5, c_y = 1$, we define u and v as follows.

$u = \frac{x - A}{c_x} = \frac{x - 95}{5}$ and $v = \frac{y - B}{c_y} = \frac{y - 50}{1} = y - 50$

	I.Q. x	**Marks y**	**$u = \frac{x-95}{5}$**	**$v = y - 50$**	**uv**	**u^2**
	85	46	–2	–4	8	4
	95	50	0	0	0	0
	100	50	1	0	0	1
	90	45	–1	–5	5	1
	110	60	3	10	30	9
	125	70	6	20	120	36
	70	40	–5	–10	50	25
Total	**675**	**361**	**2**	**11**	**213**	**76**

$$b = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

$$= \frac{7(213) - (2)(11)}{7(76) - (2)^2} \times \frac{1}{5}$$

$$= \frac{1491 - 22}{532 - 4} \times \frac{1}{5}$$

$$= \frac{1469}{528} \times \frac{1}{5}$$

$$= \frac{1469}{2640}$$

$$= 0.5564$$

$\therefore \ b \simeq 0.56$

Now, $a = \bar{y} - b\bar{x}$

$$= 51.57 - 0.56(96.43)$$

$$= 51.57 - 54.0008$$

$$= -2.4308$$

$\therefore \ a \simeq -2.43$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = -2.43 + 0.56x$

Putting $X = 120$,

$$\hat{y} = -2.43 + 0.56(120)$$

$$= -2.43 + 67.2$$

$\therefore \quad \hat{y} = 64.77$ marks.

So, when the I.Q. of a student is 120 then his marks are approximately 65.

Now, to obtain the error when I.Q. (X) = 100, first we have to find the estimate of Y i.e. $\hat{y}$.

$$\hat{y} = -2.43 + 0.56x$$

Taking $X = 100$,

$$\hat{y} = -2.43 + 0.56(100)$$

$$= -2.43 + 56$$

$\therefore \quad \hat{y} = 53.57$ marks

But the observed value of Y for $X = 100$ is 50. (See the given data)

$\therefore$ Error $e = y - \hat{y}$

$= 50 - 53.57$

$\therefore \quad e = -3.57$ marks

Note : It is necessary to keep in mind that the error can be obtained only for those values of independent variable (X), for which the observed value of dependent variable (Y) are known.

In this example, we can not obtain the error in estimating Y for $X = 120$ because the observed value of Y when $X = 120$ is not known.

Illustration 7 : From the data and calculation of illustration 12 of the chapter of linear correlation, obtain the regression line of profit on the sales. Estimate the profit when sales is ₹ 3 crore.

From the illustration, we know that

$$u = \frac{x-A}{c_x} = \frac{x-2}{0.1} \text{ and } v = \frac{y-B}{c_y} = \frac{y-5600}{100}$$

$\therefore c_x = 0.1$ and $c_y = 100$

Note that c_x is the divisor of $(x - A)$. So, though we have multiplied $(x - A)$ by 10 for simplicity of calculation, c_x is $\frac{1}{10} = 0.1$.

($\because$ To multiply by 10 is same as to divide by $\frac{1}{10} = 0.1$)

Now $b = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \dfrac{c_y}{c_x}$

$= \dfrac{9(121) - (0)(1)}{9(60) - (0)^2} \times \dfrac{100}{0.1}$

$= \dfrac{1089}{540} \times \dfrac{100}{0.1}$

$= \dfrac{108900}{54}$

$= 2016.6667$

$\therefore b \simeq 2016.67$

Now, $a = \bar{y} - b\bar{x}$

$= 5611.11 - 2016.67(2)$

$= 5611.11 - 4033.34$

$\therefore \quad a = 1577.77$

So, the regression line of profit (Y) on the sales (X) is

$\hat{y} = a + bx$

$\therefore \quad \hat{y} = 1577.77 + 2016.67x$

Putting $X = 3$,

$$\hat{y} = 1577.77 + 2016.67(3)$$

$$= 1577.77 + 6050.01$$

$$\therefore \quad \hat{y} = 7627.78$$

So, when sales is ₹ 3 crore then the estimated profit is 7627.78 (thousand ₹).

Activity

Collect the information of monthly income and monthly expenditure of your family from June to December of a year in which you are studying in standard 12. Obtain the regression line of monthly expenditure on the monthly income. Estimate the monthly expenditure of January of the successive year. Check the actual expenditure at the end of January and find the error in your estimation.

3.5 Regression coefficient from covariance and correlation coefficient

When the summary measures like mean, standard deviation (or variance), covariance, correlation coefficient are known for bivariate data of two variables X and Y, regression coefficient and the line of regression can be obtained as follows.

(1) When the measures like $\bar{x}, \bar{y}, s_x^2$ (or s_x), s_y^2 (or s_y) and Cov(x, y) are known,

$$b = \frac{\text{Covariance}(x, y)}{\text{Variance of } x} = \frac{Cov(x, y)}{s_x^2}$$

and $a = \bar{y} - b\bar{x}$

where, $Cov(x, y) = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \dfrac{\Sigma xy - n\,\bar{x}\,\bar{y}}{n}$

$$s_x^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$$

$$s_y^2 = \frac{\Sigma(y-\bar{y})^2}{n} = \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2 = \frac{\Sigma y^2}{n} - \bar{y}^2$$

(2) When the measures like $\bar{x}, \bar{y}, r, s_x$ (or s_x^2), and s_y (or s_y^2) are known,

$$b = r \cdot \frac{\text{S.D. of } y}{\text{S.D. of } x} = r \cdot \frac{s_y}{s_x}$$

and $a = \bar{y} - b\bar{x}$

The regression line of Y on X i.e. $\hat{y} = a + bx$ can be obtained by putting the values of a and b.

Now, we consider some examples in which some summary measures are known and the regression line is to be obtained.

Illustration 8 : The following measures are obtained to study the relation between rainfall in cm (X) and yield of Bajri in Quintal per Hectare (Y) in ten different regions during monsoon.

$n = 10, \bar{x} = 40, \bar{y} = 175, s_x = 12, Cov(x, y) = 360$

Obtain the regression line of yield Y on rainfall X.

Here, $Cov(x, y) = 360$ and $s_x = 12 \therefore s_x^2 = 144$

$$b = \frac{Cov(x, y)}{s_x^2}$$

$$= \frac{360}{144}$$

$\therefore \quad b = 2.5$

and $a = \bar{y} - b\bar{x}$

$$= 175 - 2.5(40)$$

$$= 175 - 100$$

$\therefore \quad a = 75$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = 75 + 2.5x$

Illustration 9 : To study the relation between two variables, yearly income (X) of a family and their yearly investment (Y) in mutual funds, the following information is shown for a sample of 100 families of a city.

$X =$ **Annual income of a family (lakh ₹)**

$Y =$ **Annual investment in mutual fund of a family (thousand ₹)**

$\bar{x} = 5.5, \bar{y} = 40.5, s_x = 1.2, s_y = 12.8, r = 0.65$

Obtain the regression line of annual investment in mutual fund of a family on their annual income. Estimate the annual investment in mutual fund of a family whose annual income is ₹ 4.5 lakh.

Here, $n = 100, \bar{x} = 5.5, \bar{y} = 40.5$

$s_x = 1.2, s_y = 12.8$ and $r = 0.65$

Now, $b = r \cdot \frac{s_y}{s_x}$

$$= 0.65 \times \frac{12.8}{1.2}$$

$$= 6.9333$$

$\therefore \quad b \simeq 6.93$

And $a = \bar{y} - b\bar{x}$

$= 40.5 - 6.93\ (5.5)$

$= 40.5 - 38.115$

$= 2.385$

$\therefore \quad a \simeq 2.39$

So, the regression line of Y on X is

$\hat{y} = a + bx$

$\therefore \quad \hat{y} = 2.39 + 6.93x$

Putting $X = 4.5$,

$\hat{y} = 2.39 + 6.93(4.5)$

$= 2.39 + 31.185$

$= 33.575$

$\therefore \quad \hat{y} \simeq 33.58$

So, when annual income of a family is ₹ 4.5 lakh then estimated investment in mutual fund is ₹ 33.58 thousand.

Illustration 10 : The information of price (in ₹) of a ballpen and the supply of ballpen (in units) at the end of each month of a year for a company making ball pen is given below. Estimate the supply of ballpen when its price is ₹ 40.

Detail	Price (x)	Supply (y)
Average	30	500
Variance	25	10,000
$r = 0.8$		

Here, $\bar{x} = 30, \bar{y} = 500, s_x^2 = 25, s_y^2 = 10000$ and $r = 0.8$

Since $s_x^2 = 25$, $s_x = 5$

Since $s_y^2 = 10000$, $s_y = 100$

Since the supply Y is to be estimated for the price $X = 40$, we shall obtain the regression line of Y on X.

$b = r \cdot \frac{s_y}{s_x}$

$= 0.8 \times \frac{100}{5}$

$\therefore \quad b = 16$

$a = \bar{y} - b\bar{x}$

$= 500 - 16\ (30)$

$= 500 - 480$

$\therefore \quad a = 20$

So, the regression line of Y on X is

$\hat{y} = a + bx$

$\therefore \quad \hat{y} = 20 + 16x$

Putting $X = 40$,

$\hat{y} = 20 + 16(40)$

$= 20 + 640$

$\therefore \quad \hat{y} = 660$ units

So, when the price is ₹ 40, the estimate of supply is 660 units.

Illustration 11 : A person in a state of South India produces spoons from eatable materials. It can be eaten after using it. He launched such spoons for the purpose of selling in a state on an experimental level. The following results are obtained for the average price (in ₹) and its demand (in hundred units) for the last six months.

$\boldsymbol{n = 6, \Sigma x = 45, \Sigma y = 122, \Sigma x^2 = 439, \Sigma xy = 605}$

Obtain the regression line of the demand (Y) of spoons on the price (X) and estimate the demand of spoons when the price of a spoon is ₹ 10.

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{45}{6} = 7.5$, $\bar{y} = \frac{\Sigma y}{n} = \frac{122}{6} = 20.33$

$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

$= \frac{6(605) - (45)(122)}{6(439) - (45)^2}$

$= \frac{3630 - 5490}{2634 - 2025}$

$= \frac{-1860}{609}$

$= -3.0542$

$\therefore \quad b \simeq -3.05$

$a = \bar{y} - b\bar{x}$

$= 20.33 - (-3.05)\ (7.5)$

$= 20.33 + 22.875$

$= 43.205$

$\therefore \quad a \simeq 43.21$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = 43.21 - 3.05x$$

Putting $X = 10$,

$$\hat{y} = 43.21 - 3.05(10)$$

$$= 43.21 - 30.5$$

$$\therefore \quad \hat{y} = 12.71$$

So, when the price is ₹ 10, estimated demand is 12.71 (hundred units).

Illustration 12 : The electricity is generated by windmill manufactured by a company. The following information is obtained by recording five observations regarding the velocity of wind (km per hour) and generation of electricity (in Watts) by a unit of the company.

Velocity of Wind = X km per hour

Electricity Generation = Y Watts

$$\bar{x} = 20, \ \bar{y} = 186, \ \Sigma xy = 23200, \ s_x^2 = 50$$

Obtain the regression line of electricity generation (Y) on velocity of wind (X). Estimate the electricity generation if the velocity of wind is 25 km per hour.

Here, $n = 5, \ \Sigma xy = 23200, \ \bar{x} = 20, \ \bar{y} = 186$ and $s_x^2 = 50$

Now, $b = \dfrac{Cov(x, y)}{s_x^2}$

$$= \frac{\Sigma xy - n\bar{x}\,\bar{y}}{n \cdot s_x^2}$$

$$= \frac{23200 - 5(20)(186)}{5(50)}$$

$$= \frac{23200 - 18600}{250}$$

$$= \frac{4600}{250}$$

$$\therefore \quad b = 18.4$$

$$a = \bar{y} - b\bar{x}$$

$$= 186 - 18.4 \ (20)$$

$$= 186 - 368$$

$$\therefore \quad a = -182$$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = -182 + 18.4x$

Putting $X = 25$,

$$\hat{y} = -182 + 18.4(25)$$

$$= -182 + 460$$

$\therefore \quad \hat{y} = 278$

So, when the velocity of wind is 25 km per hour, approximately 278 watts electricity is generated.

Exercise 3.2

1. The following information is obtained from a study to know the effect of use of fertilizer on the yield of cotton.

Consumption of fertilizer (10 kg) x	28	35	25	24	20	25	20
Yield of cotton per hectare (Quintals) y	128	140	115	120	105	122	100

Obtain the regression line of Y on X and estimate the yield of cotton per hectare if 300 kg fertilizer is used.

2. To know the relationship between the heights of father and sons, obtain the regression line of height of son on the height of father from the following information of eight pairs of fathers and adult sons.

Height of father (cm) x	167	169	171	168	173	166	167	165
Height of son (cm) y	158	170	169	172	170	168	164	167

Estimate the height of a son whose father's height is 170 cm.

3. From the following information of altitude and the amount of effective Oxygen in air at the place, obtain the regression line of amount of effective Oxygen (Y) on the altitude (X). (305 meter $\simeq$ 1000 feet)

Altitude (305 meter) x	0	1	2	3	4	5	6
Effective Oxygen (%) y	20.9	20.1	19.4	17.9	17.9	17.3	16.6

If the altitude of a place is 7 units (1 unit = 305 meter), estimate the percentage of effective Oxygen in air at that place.

4. The following information is obtained to study the relation between the carpet area in a house and its monthly rent in a city.

Carpet area (square meter) x	55	60	75	80	100	120	140
Monthly rent (₹) y	18,000	19,000	20,000	20,000	25,000	30,000	50,000

Obtain the regression line of Y on X. Estimate the monthly rent of a house having carpet area of 110 square meter.

5. The following sample data is obtained to study the relation between the number of customers visiting a mall per day and the sales (ten thousand ₹).

No. of customers x	50	70	100	70	150	120
Sales (ten thousand ₹) y	2.0	2.0	2.5	1.4	4.0	2.5

Obtain the regression line of Y on X. Estimate the sales of a mall if 80 customers have visited the mall on a particular day.

6. The following information is given for ten firms running business of clothes in a city regarding their average annual profit (in lakh ₹) and average annual administrative cost (in lakh ₹).

Particulars	**Profit (in lakh ₹)** x	**Administrative Cost (in lakh ₹)** y
Mean	60	25
Standard Deviation	6	3
Covariance = 10.4		

Obtain the regression line of Y on X.

7. The following information is obtained to study the relationship between average rainfall (in cm) and the yield of maize (in quintal per hectare) in different talukaa of Gujarat.

Particulars	**Rainfall (cm)** x	**Yield of Maize (Quintal per Hectare)** y
Mean	82	180
Variance	64	225
Correlation coefficient = 0.82		

Estimate the yield of maize when the rainfall is 60 cm.

8. The following results are obtained to study the relation between the price of battery (cell) of wrist watch in rupees (X) and its supply in hundred units (Y).

$n = 10,\ \Sigma x = 130, \Sigma y = 220,\ \Sigma x^2 = 2288,\ \Sigma xy = 3467$

Obtain the regression line of Y on X and estimate the supply when price is ₹ 16.

9. The information regarding maximum temperature (X) and sale of ice-cream (Y) of six different days in summer for a city is given below.

Maximum Temperature = X (in Celsius)

Sale of Ice-cream = Y (in lakh ₹)

$\bar{x} = 40,\ \bar{y} = 1.2,\ \Sigma xy = 306,\ s_x^2 = 20$

Obtain the regression line of sale of ice-cream on maximum temperature. Estimate the sale of ice-cream if the maximum temperature on a day is 42 Celsius.

*

3.6 Coefficient of Determination

We know that regression is a functional relation between two correlated variables and it is useful to estimate the value of dependent variable for some given value of independent variable. The coefficient of determination is a measure to find the reliability of such an estimate.

Suppose the regression line of Y on X is $\hat{y} = a + bx$, then the square of the correlation coefficient between observed values of dependent variable y obtained from the observations and its estimated values $\hat{y}$ which are obtained from the regression line is called the **coefficient of determination.** It is denoted by R^2.

$$\therefore R^2 = \left[r\left(y, \hat{y}\right)\right]^2$$

It can be easily checked that R^2 is same as $r^2(x, y)$ or r^2.

$$R^2 = \left[r\left(y, \hat{y}\right)\right]^2$$

$$= \left[r\left(y, a + bx\right)\right]^2$$

$$= \left[r\left(y, x\right)\right]^2$$

$$= \left[r\left(x, y\right)\right]^2$$

$$\therefore R^2 = r^2$$

$\because$ r is independent of change of origin and scale, so from variable $\hat{y}(= a + bx)$, substracting a and then dividing by b, the value of r will not change.

Since $R^2 = r^2$, we can say that the reliability of an estimate of dependent variable Y largely depends on the correlation coefficient r between two variables X and Y.

If $r = \pm 1$ then $R^2 = r^2 = 1$ and there is a perfect linear correlation between X and Y. So, we can say that the estimates of Y obtained from the regression line are 100 % reliable. But if $r = 0$ then $R^2 = r^2 = 0$ and there is no linear correlation between X and Y. So, we can say that the estimates of Y obtained from the regression line are not reliable.

It is clear from the above discussion that high value of R^2 shows that a good linear correlation exists between two variables X and Y. So, we can check whether the linearity assumption of regression is valid or not from the measure of coefficient of determination (R^2). If the value of R^2 is nearer to 1, the assumption of linearity of regression is valid. But if the value of R^2 is nearer to 0, the assumption of linearity of regression between X and Y is not valid.

How much variation in the dependent variable Y can be explained by the regression line, can be obtained from the coefficient of determination. e.g., If $r = 0.9$ for some data, then coefficient of determination = $(0.9)^2 = 0.81$ and therefore $r^2 \times 100\ \% = 81\ \%$. So, it can be said that out of total variation in variable Y, the explanation of 81 % variation is obtained from the regression line. So, we can say that the regression model used for the given data is suitable.

Illustration 13 : The following table shows the experience of technicians (in years) employed at various companies and their monthly salary (in thousand ₹).

Experience (years) x	**12**	**8**	**16**	**20**	**5**	**14**	**10**
Monthly Salary (thousand ₹) y	**22**	**15**	**25**	**30**	**12**	**24**	**20**

Calculate the coefficient of determination and check the validity of the linearity assumption of regression between the years of experience and the monthly salary.

Here, $n = 7,\ \bar{x} = \frac{\Sigma x}{n} = \frac{85}{7} = 12.14,\ \bar{y} = \frac{\Sigma y}{n} = \frac{148}{7} = 21.14$

	Experience (year) x	Monthly salary (thousand ₹) y	xy	x^2	y^2
	12	22	264	144	484
	8	15	120	64	225
	16	25	400	256	625
	20	30	600	400	900
	5	12	60	25	144
	14	24	336	196	576
	10	20	200	100	400
Total	**85**	**148**	**1980**	**1185**	**3354**

Now, $R^2 = r^2 = \left[\frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}\right]^2$

$= \left[\frac{7(1980) - (85)(148)}{\sqrt{7(1185) - (85)^2} \cdot \sqrt{7(3354) - (148)^2}}\right]^2$

$= \frac{[13860 - 12580]^2}{[8295 - 7225] \cdot [23478 - 21904]}$

$= \frac{(1280)^2}{(1070) \cdot (1574)}$

$= \frac{1638400}{1684180}$

$= 0.9728$

$\therefore\ R^2 \simeq 0.97$

The value of R^2 is very near to 1. So, we can say that the linearity assumption of regression between the years of experience and the monthly salary is valid.

Note : For the above example, we can also compute R^2 by taking $u = x - A$ and $v = y - B$. Here, A and B are suitable constants.

Illustration 14 : In order to study the relationship between the density of population and the number of persons suffering from skin diseases, the following information is obtained for six cities regarding their density of population (per sq. km) and persons suffering from skin diseases (per thousand).

Density (per sq. km) x	**12,000**	**14,500**	**19,000**	**17,500**	**13,500**	**16,000**
Number of patients (per thousand) y	**80**	**60**	**90**	**80**	**40**	**30**

Obtain the regression line of Y on X. Estimate the number of patients suffering from skin diseases if density of population of a city is 15000 (per sq.km). Examine the reliability of this regression model.

Here, $n = 6, \bar{x} = \frac{\Sigma x}{n} = \frac{92500}{6} = 15416.67$; $\bar{y} = \frac{\Sigma y}{n} = \frac{380}{6} = 63.33$

We can see that the values of variable X are multiple of 500 and that of variable Y are of 10. So, by taking A = 15000, B = 60, c_x = 500, c_y = 10, we shall use short-cut method. Let us define u and v as follows.

$$u = \frac{x-A}{c_x} = \frac{x-15000}{500} \text{ and } v = \frac{y-B}{c_y} = \frac{y-60}{10}$$

	Density (per sq. km) x	**Number of patients (per thousand) y**	**$u = \frac{x-15000}{500}$**	**$v = \frac{y-60}{10}$**	**uv**	**u^2**	**v^2**
	12000	80	–6	2	–12	36	4
	14500	60	–1	0	0	1	0
	19000	90	8	3	24	64	9
	17500	80	5	2	10	25	4
	13500	40	–3	–2	6	9	4
	16000	30	2	–3	–6	4	9
Total	**92500**	**380**	**5**	**2**	**22**	**139**	**30**

$$b = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

$$= \frac{6(22) - (5)(2)}{6(139) - (5)^2} \times \frac{10}{500}$$

$$= \frac{132 - 10}{834 - 25} \times \frac{1}{50}$$

$$= \frac{122}{809} \times \frac{1}{50}$$

$$= \frac{122}{40450}$$

$\therefore b \simeq 0.003$

$$a = \bar{y} - b\bar{x}$$
$$= 63.33 - 0.003\,(15416.67)$$
$$= 63.33 - 46.25$$

$\therefore\ a = 17.08$

$\therefore$ The regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore\ \hat{y} = 17.08 + 0.003\,x$

Putting $X = 15000$,

$$\hat{y} = 17.08 + 0.003(15000)$$
$$= 17.08 + 45$$

$\therefore\ \hat{y} = 62.08$

So, when the density of a city is 15,000 then approximately $62.08 \simeq 62$ patients are suffering from skin diseases.

Now, reliability of a regression model can be examined by coefficient of determination R^2. So, we obtain it.

$$R^2 = r^2 = \left[\frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\cdot\sqrt{n\Sigma v^2 - (\Sigma v)^2}}\right]^2$$

$$= \frac{[6(22) - (5)(2)]^2}{\left[6(139) - (5)^2\right]\left[6(30) - (2)^2\right]}$$

$$= \frac{(122)^2}{(809)(176)}$$

$$= \frac{14884}{142384}$$

$$= 0.1045$$

$\therefore\ R^2 \simeq 0.10$

As the value of R^2 is very near to 0, it can not be said that the regression model is reliable.

3.7 Properties of Regression Coefficient

(1) Correlation coefficient r and regression coefficient b are either both positive or both negative. ($\because$ We know that standard deviations s_x and s_y are always non-negative and $-1 \le r \le 1$. So, from $b = r \cdot \frac{s_y}{s_x}$ it can be understood that the sign of b will be same as that of r.)

(2) Regression coefficient is independent of change of origin but not independent of change of scale. (This property is discussed in detail in the explanation of the short-cut method of calculation of regression coefficient.)

Note : The regression line of Y on X always passes through the point $(\bar{x}, \bar{y})$.

Illustration 15 : Six pairs of father-son are selected in a sample of an experiment to know the relation between the heights of fathers in cm (X) and the heights of their adult sons in cm (Y).

The following results are obtained from it.

$$\Sigma x = 1020,\ \Sigma y = 990,\ \Sigma(x-170)^2 = 60,\ \Sigma(y-165)^2 = 105$$

$$\Sigma(x-170)(y-165) = 45$$

Obtain the regression line of the heights of sons (Y) on the heights of fathers (X). Also verify the reliability of the regression model.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1020}{6} = 170$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{990}{6} = 165$$

$$\therefore \Sigma(x-170)^2 = \Sigma(x-\bar{x})^2 = 60$$

$$\Sigma(y-165)^2 = \Sigma(y-\bar{y})^2 = 105$$

$$\Sigma(x-170)(y-165) = \Sigma(x-\bar{x})(y-\bar{y}) = 45$$

$$\therefore b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$= \frac{45}{60}$$

$$\therefore b = 0.75$$

$$a = \bar{y} - b\bar{x}$$

$$= 165 - 0.75(170)$$

$$= 165 - 127.5$$

$$\therefore a = 37.5$$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \hat{y} = 37.5 + 0.75x$$

Now, to verify reliability of the regression model, let us obtain the coefficient of determination R^2.

$$R^2 = \left[\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\cdot\sqrt{\Sigma(y-\bar{y})^2}}\right]^2$$

$$= \frac{(45)^2}{(60)(105)}$$

$$= \frac{2025}{6300}$$

$$= 0.3214$$

$$\therefore R^2 \simeq 0.32$$

Since the value of R^2 is nearer to 0, it can not be said that the regression model is reliable.

Illustration 16 : (i) If the regression line of Y on X is $\hat{y} = 12 - 1.5x$ and the mean of X is 6, find the mean of Y. (ii) If the regression line of Y on X is $\hat{y} = 11.5 + 0.65x$ and $\bar{y} = 18$, find $\bar{x}$.

(i) We know that the regression line always passes through a point $(\bar{x}, \bar{y})$. So, the $\hat{y}$ obtained by putting $\bar{x}$ in place of x in the regression line is $\bar{y}$ or the x obtained by putting $\bar{y}$ in place of $\hat{y}$ is $\bar{x}$.

Putting $\bar{x} = 6$ in place of x in $\hat{y} = 12 - 1.5x$,

$\hat{y} = 12 - 1.5(6)$

$\therefore \hat{y} = 12 - 9$

$\therefore \hat{y} = 3$, so $\bar{y} = 3$

Therefore, the mean of Y is 3.

(ii) As per the above discussion, the value of x obtained by putting $\bar{y} = 18$ in place of $\hat{y}$ in $\hat{y} = 11.5 + 0.65x$, we get $\bar{x}$.

By putting $\hat{y} = \bar{y} = 18$ in $\hat{y} = 11.5 + 0.65x$,

$18 = 11.5 + 0.65\,x$

$\therefore 6.5 = 0.65x$

$\therefore x = \frac{6.5}{0.65}$

$\therefore x = 10$ So $\bar{x} = 10$

Therefore, the mean of X is 10.

Illustration 17 : (i) If $\bar{x} = 5, \bar{y} = 11$ and $b = 1.2$, obtain the regression line of Y on X. (ii) If $\bar{x} = 60, \bar{y} = 75$ and $s_x^2 : Cov(x, y) = 5:3$, obtain the regression line of Y on X and estimate y for $X = 65$ from it.

(i) Here, $b = 1.2$, $\bar{x} = 5$ and $\bar{y} = 11$.

Now, $a = \bar{y} - b\bar{x}$

$\therefore a = 11 - 1.2(5)$

$= 11 - 6$

$\therefore a = 5$

We get the regression line of Y on X as follows.

$\hat{y} = a + bx$

$\therefore \hat{y} = 5 + 1.2x$

(ii) Here, $\bar{x} = 60,\ \bar{y} = 75$ and $s_x^2 : Cov(x, y) = 5:3$

$s_x^2 : Cov(x, y) = 5:3$

$\therefore \quad \frac{s_x^2}{Cov(x, y)} = \frac{5}{3}$ hence, $\frac{Cov(x, y)}{s_x^2} = \frac{3}{5}$

Now, $b = \frac{Cov(x, y)}{s_x^2} = \frac{3}{5} = 0.6$

and $a = \bar{y} - b\bar{x}$

$= 75 - 0.6\,(60)$

$= 75 - 36$

$\therefore \ a = 39$

We get the regression line of Y on X as follows.

$\hat{y} = a + bx$

$\therefore \ \hat{y} = 39 + 0.6\,x$

Putting $X = 65$,

$\hat{y} = 39 + 0.6\,x$

$= 39 + 0.6\,(65)$

$= 39 + 39$

$\therefore \ \hat{y} = 78$

So, for $X = 65$, the estimated value of Y is 78.

Illustration 18 : The fitted regression line of Y on X is $\hat{y} = 50 + 3.5x$. If an observation $(16, 108)$ is used in fitting of the line, find the error in estimating Y for $X = 16$. (ii) If one observation $(10, 30)$ is used in the fitting of the line $\hat{y} = 22 + 0.8x$, find the error in estimating Y for $X = 10$. What can you deduce from the value of the error ?

(i) Putting $x = 16$ in $\hat{y} = 50 + 3.5x$,

$\hat{y} = 50 + 3.5(16)$

$\therefore \ = 50 + 56$

$\therefore \ \hat{y} = 106$

And for $X = 16$, corresponding $Y = 108$ is observed.

$\therefore$ Error $e = y - \hat{y}$

$= 108 - 106$

$\therefore \ e = 2$

So, the error in estimating Y for $X = 16$ is 2.

(ii) Putting $X = 10$ in $\hat{y} = 22 + 0.8x$,

$$\hat{y} = 22 + 0.8(10)$$

$$= 22 + 8$$

$$\therefore \hat{y} = 30$$

And for $X = 10$, corresponding $Y = 30$ is observed.

$$\therefore \text{ Error } e = y - \hat{y}$$

$$= 30 - 30$$

$$\therefore e = 0$$

So, the error in estimating Y for $X = 10$ is 0.

Since value of the error is 0, we can say that the point $(10, 30)$ lies on the fitted line $\hat{y} = 22 + 0.8x$.

Note : For a regression line obtained by the method of least squares, the error is positive for the points above the line, negative for the points below the line and it is zero for the points which are on the line.

Illustration 19 : (i) If the regression line of Y on X is $\hat{y} = 25 + 3x$ and $Cov(x, y) = 48$, find the standard deviation of X. Also find coefficient of determination if the standard deviation of Y is 15. (ii) For the regression line given in the above question, how many units should be increased in the value of X to increase approximately 15 units in Y ?

(i) By comparing the regression line of y on x, $\hat{y} = 25 + 3x$ with its general form $\hat{y} = a + bx$, we get regression coefficient $b = 3$. Since $Cov(x, y) = 48$ is given,

$$b = \frac{Cov(x, y)}{s_x^2}$$

$$\therefore \quad 3 = \frac{48}{s_x^2}$$

$$\therefore \quad s_x^2 = 16$$

$$\therefore \quad s_x = 4$$

So, the standard deviation of X is 4.

Now, the standard deviation of Y, $s_y = 15$ is given.

So, coefficient of determination $R^2 = \left[\frac{Cov(x, y)}{s_x \cdot s_y}\right]^2$

$$\therefore \quad R^2 = \left[\frac{48}{4 \times 15}\right]^2 = (0.8)^2 = 0.64$$

Second Method :

$$b = r \cdot \frac{s_y}{s_x}$$

$$\therefore 3 = r \cdot \frac{15}{4}$$

$$\therefore r = \frac{3 \times 4}{15}$$

$$\therefore r = 0.8$$

$$\therefore R^2 = r^2 = (0.8)^2 = 0.64$$

(ii) $\hat{y} = 25 + 3x$ and regression coefficient $b = 3$. It indicates that if the value of X is increased by one unit then estimated value of Y is increased by 3 units. So, if the value of Y is to be increased approximately by 15 units then the value of X should be increased by $\frac{15}{3} = 5$ units.

Illustration 20 : (i) If the regression line is $\hat{y} = \frac{x}{2} + 5$ and $s_y : s_x = 5:8$, find the coefficient of determination. (ii) If the regression line of Y on X is $4x + 5y - 65 = 0$, find the value of regression coefficient b.

(i) By comparing the regression line $\hat{y} = \frac{x}{2} + 5 = \frac{1}{2} \cdot x + 5$ with its general form $\hat{y} = a + bx$, we get $b = \frac{1}{2}$.

Now, $s_y : s_x = 5:8$

$$\therefore \frac{s_y}{s_x} = \frac{5}{8}$$

and $b = r \cdot \frac{s_y}{s_x}$

$$\therefore \frac{1}{2} = r \cdot \frac{5}{8}$$

$$\therefore r = \frac{1}{2} \times \frac{8}{5}$$

$$\therefore r = 0.8$$

$\therefore$ The coefficient of determination $R^2 = r^2 = (0.8)^2 = 0.64$.

(ii) The regression line of Y on X, $4x + 5y - 65 = 0$ is given.

Now, we convert it into its general form.

$$4x + 5y - 65 = 0$$

$$\therefore 5y = 65 - 4x$$

$$\therefore y = \frac{65 - 4x}{5}$$

$$\therefore y = \frac{65}{5} - \frac{4x}{5}$$

$$\therefore y = 13 - 0.8x$$

By comparing it with $\hat{y} = a + bx$, we get $b = -0.8$.

Illustration 21 :

(i) If $b_{yx} = 0.85$, $u = x - 15$ and $v = y - 20$, find the value of b_{vu}.

(ii) If $u = \frac{x-5}{3}$, $v = \frac{y-8}{5}$ and $b_{yx} = 0.9$, find the value of b_{vu}.

(iii) If $u = 10(x - 4.5)$, $v = \frac{y-50}{10}$ and $b_{yx} = 0.25$, find the value of b_{vu}.

(iv) If $u = 5(x - 40)$, $v = 2(y - 18)$ and $b_{yx} = 1.6$, find the value of b_{vu}.

For the solution of all the questions given above, we shall use the following property of the regression coefficient.

- If $u = x - A$ and $v = y - B$ then $b_{yx} = b_{vu}$

- If $u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$ then $b_{yx} = b_{vu} \cdot \frac{c_y}{c_x}$

(i) Since $u = x - 15 = x - A$ and $v = y - 20 = y - B$

$\therefore\ b_{vu} = b_{yx} = 0.85$

(ii) Since $u = \frac{x-5}{3} = \frac{x-A}{c_x}$ and $v = \frac{y-8}{5} = \frac{y-B}{c_y}$

$b_{yx} = b_{vu} \cdot \frac{c_y}{c_x}$ $\therefore\ b_{vu} = b_{yx} \cdot \frac{c_x}{c_y} = 0.9 \times \frac{3}{5} = 0.54$

(iii) Since $u = 10(x - 4.5) = \frac{x-4.5}{\frac{1}{10}} = \frac{x-A}{c_x}$ and $v = \frac{y-50}{10} = \frac{y-B}{c_y}$

$b_{yx} = b_{vu} \cdot \frac{c_y}{c_x}$ $\therefore\ b_{vu} = b_{yx} \cdot \frac{c_x}{c_y} = 0.25 \times \frac{\left(\frac{1}{10}\right)}{10} = 0.25 \times \frac{1}{100} = 0.0025$

(iv) Since $u = 5(x - 40) = \frac{x-40}{\frac{1}{5}} = \frac{x-A}{c_x}$ and $v = 2(y - 18) = \frac{y-18}{\frac{1}{2}} = \frac{y-B}{c_y}$

$b_{yx} = b_{vu} \cdot \frac{c_y}{c_x}$ $\therefore\ b_{vu} = b_{yx} \cdot \frac{c_x}{c_y} = 1.6 \times \frac{\left(\frac{1}{5}\right)}{\left(\frac{1}{2}\right)} = 1.6 \times \frac{2}{5} = 0.64.$

3.8 Precautions while using Regression

We know that regression is a functional relation between two correlated variables and hence we can estimate the value of dependent variable from it. The regression analysis is very useful in decision making in the practical fields like economics, trade, industry, education, psychology, sociology, medicine, planning etc. Inspite of vast application of regression analysis, some precautions are necessary while using it.

(1) The reliability of the estimate can be verified by the coefficient of determination $\left(R^2\right)$. So, we should use the estimate only after ascertaining the linearity of regression by the coefficient of determination.

(2) Another point which is necessary to keep in mind while using the regression analysis is, the regression relation obtained by the scatter diagram or by the method of least squares should not be used for the values which are very far from the given values of the independent variable.

e.g. If for some data, there is a high degree of correlation between the rainfall and the yield of wheat, we can say that as rainfall increases, yield of wheat also increases. Now, using the regression relation obtained from the given data, if for some value of rainfall, corresponding yield of wheat is to be estimated then the estimate of yield of wheat can be proper only when the value of rainfall is around the given values of rainfall. If there is heavy rain then the crop may get damaged and the yield of wheat may also decrease. In such a case, the dependent variable (yield) estimated from the above mentioned regression relation may be wrong.

Summary

- The concept of regression is studied under the assumption that two variables under study have cause-effect relationship.
- Regression : Functional relation between two related variables.
- Linear Regression : Functional relation between two related variables in which the change in the values of the variables are approximately in constant proportion and this relationship can be determined by a straight line.
- The value of the dependent variable can be estimated for some known value of the independent variable by using regression.
- Regression coefficient : The approximate change in the value of dependent variable for a unit change in the value of independent variable. It is also known as slope of the regression line.
- Error : Mistake occurring in estimating the value of the dependent variable.
- Coefficient of Determination : It is the square of the correlation coefficient between the observed value of dependent variable Y and its estimated values. In case of two variables, it is same as square of the coefficient of correlation between independent variable X and dependent variable Y.
- Using the coefficient determination, how much variation in the dependent variable Y is explained by the regression line can be known and the reliability of the regression model can also be known.
- The regression relation should not be used for the values which are very far from the given values of the independent variable.

List of Formulae :

Equation of Regression Line

$\hat{y} = a + bx$

Where, $b = b_{yx}$ = Regression Coefficient

(1) $b = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$

(2) $b = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

(3) $b = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2}$ Here, $u = x - A$ and $v = y - B$

(4) $b = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \dfrac{c_y}{c_x}$ Here, $u = \dfrac{x-A}{c_x}$ and $v = \dfrac{y-B}{c_y}$

(5) $b = r \cdot \dfrac{s_y}{s_x}$

(6) $b = \dfrac{Cov(x, y)}{s_x^2}$

(7) $a = \bar{y} - b\bar{x}$

(8) Coefficient of Determination $R^2 = \left[r(y, \hat{y})\right]^2 = \left[r(x, y)\right]^2 = r^2$

Exercise 3

Section A

Find the correct option for the following multiple choice questions :

1. Which of the following indicates the functional relation between the two variables ?

(a) Correlation (b) Regression (c) Mean (d) Variance

2. The best fitted line of regression can be obtained by which method ?

(a) Least Square Method (b) Karl Pearson's Method

(c) Maximum Square Method (d) Bowley's Method

3. In usual notation, what is b_{yx} ?

(a) Intercept (b) Dependent Variable

(c) The approximate change in the value of Y for a unit change in the value of X.

(d) The approximate change in the value of X for a unit change in the value of Y.

4. Which of the following is correct ?

(a) $b_{yx} = r \cdot \frac{s_x}{s_y}$ (b) $b_{yx} = r \cdot \frac{s_y^2}{s_x^2}$ (c) $b_{yx} = \frac{Cov(x, y)}{s_y^2}$ (d) $b_{yx} = r \cdot \frac{s_y}{s_x}$

5. The regression line always passes through which point ?

(a) $(\bar{x}, \bar{y})$ (b) $(0, \bar{y})$ (c) $(\bar{x}, 0)$ (d) $(0, 0)$

6. What is error e in estimation in case of line of regression of Y on X ?

(a) $y - \hat{y}$ (b) $\hat{x} - \hat{y}$ (c) $x - \hat{x}$ (d) $\hat{y} - \hat{x}$

7. Which regression line is used if the sale of a commodity depends on its advertisement cost ?

(a) Regression line of advertisement cost on sale

(b) Regression line of advertisement cost on advertisement cost

(c) Regression line of sales on advertisement cost

(d) Regression line of sales on sales

8. Which of the following is a regression line of Y on X ?

(a) $\hat{y} = a + bx + cx^2$ (b) $\hat{x} = c + by$ (c) $\hat{y} = a + bx$ (d) $\hat{y} = a + bx^2$

9. For which value of the correlation coefficient (r), the regression coefficient becomes zero ?

(a) 1 (b) –1 (c) $\frac{1}{2}$ (d) 0

10. What is coefficient of determination in the study of regression for two variables ?

(a) Product of two standard deviations (b) Square of correlation coefficient

(c) Square of covariance (d) Product of two variances

11. If the regression line is $\hat{y} = 10 + 3x$, what is the estimate of Y for $X = 20$?

(a) 13 (b) 60 (c) 70 (d) 203

12. What is the value of b_{yx} if the regression line is $2x + 3y - 50 = 0$?

(a) $\frac{3}{2}$ (b) $-\frac{3}{2}$ (c) $-\frac{2}{3}$ (d) 2

13. The regression line of Y on X is $\hat{y} = 30 - 1.5x$. What is the value of $\bar{y}$ if $\bar{x} = 10$?

(a) 28.5 (b) 20 (c) 15 (d) 45

14. If $u = \frac{x-15}{10}$ and $v = \frac{y-50}{2}$ and $b_{yx} = 7.5$, What is the value of b_{vu} ?

(a) 7.5 (b) 1.5 (c) 37.5 (d) 150

15. If $r = 0.8$, how much part of the total variation in the dependent variable can be explained by the regression model ?

(a) 80 % (b) 64 % (c) 36 % (d) 20 %

Section B

Answer the following questions in one sentence :

1. Define : Linear Regression
2. Define : Regression Coefficient
3. State the Linear Regression model.
4. What is an error in context with a regression line ?
5. Give the name of a method to obtain the best fitted regression line.
6. The regression coefficient is independent of which transformation ?
7. The regression coefficient is not independent of which transformation ?
8. What is the value of error if a sample point is on the fitted line ?
9. Will the regression coefficient change if the values of both the variables are doubled with the help of transformation of scale ?
10. If $r = 0.5, s_x = 2, s_y = 4$, what is the value of b_{yx} ?
11. If a regression line is $\hat{y} = 31.5 + 1.85x$, estimate Y for $X = 10$.
12. If Y and X have the relation $y = a + bx$, where $b > 0$ then what is the value of r ?
13. If $y = 5 - 3x$ is the relation between Y and X then what is the value of r ?

Section C

Answer the following questions :

1. What are the constants a and b in the regression line $\hat{y} = a + bx$?
2. The fitted regression line of Y on X is $\hat{y} = 23.2 - 1.2x$ and one of the observations used in fitting of the line is (6, 17). Find the error in estimating Y for $X = 6$.
3. If $\bar{x} = 30, \bar{y} = 20$ and $b = 0.6$, find the intercept of the regression line of Y on X and write equation of the line.
4. Interpret $b_{yx} = 5$.
5. If $b = 1.5, r = 0.8$ and standard deviation of X is 1.6, find the standard deviation of Y.
6. If the regression coefficient of the regression line of Y on X is 0.6 and the standard deviations of X and Y are 5 and 3 respectively, find the coefficient of determination.
7. If the regression line of Y on X is $\hat{y} = 35 + 2x$ and $Cov(x, y) = 50$, find the standard deviation of X.
8. For the regression line given in the previous question (7), if the value of Y is to be increased by 10 units, how many units should be increased in the value of X ?
9. If $\bar{x} = 10, \bar{y} = 25, \Sigma(x-10)(y-25) = 120$ and $\Sigma(x-10)^2 = 100$, find the values of a and b for the regression line of Y on X.
10. If $b_{yx} = 0.75, u = 6(x-20)$ and $v = 2(y-15)$ for the data in the study of a regression line then find the value of b_{vu}.

Section D

Answer the following questions :

1. Explain the statement, "There is a cause and effect relationship between two variables" by giving a suitable example. Also define independent variable and dependent variable.
2. Explain the method of scatter diagram for fitting a line of regression and state its limitation.
3. Explain the method of least square for fitting a regression line.
4. State the utility of regression.
5. State properties of regression coefficient. Also state the point through which a regression line always passes.
6. Explain : coefficient of determination
7. State precautions which are necessary while using the regression.
8. For two related variables X and Y, $\Sigma(x-\bar{x})^2 = 80$, $\Sigma(x-\bar{x})(y-\bar{y}) = 60$, $\bar{x} = 8$, $\bar{y} = 10$. Obtain the regression line of Y on X.
9. If $\bar{x} = 30$, $\bar{y} = 50$, $r = 0.8$ and the standard deviations of X and Y are 2 and 5 respectively, obtain the regression line of Y on X.
10. If the regression line of Y on X is $\hat{y} = 11 + 3x$ and $s_x : s_y = 3:10$, find the coefficient of determination.
11. In usual notations, $n = 7$, $\Sigma u = 2$, $\Sigma v = 25$, $\Sigma u^2 = 160$ and $\Sigma uv = 409$. Obtain the regression coefficient of a regression line of Y on X and interpret it.
12. If $b_{yx} = 0.8$ then find the value of b_{vu} for the following u and v.

    (i) $u = x - 105$ and $v = y - 90$

    (ii) $u = \frac{x-1400}{100}$ and $v = \frac{y-750}{50}$

    (iii) $u = 10(x - 4.6)$ and $v = y - 75$

13. The following results are obtained for a bivariate data.

Particulars	x	y
No. of observations	8	
Mean	100	100
The sum of squares of deviations taken from mean	130	145
The sum of product of deviations taken from mean	115	

Obtain the regression line of Y on X.

Section E

Solve the following :

1. A manager of the an I.T. company has collected the following information regarding the years of job and monthly income of seven marketing executives.

Years of job	10	6	8	5	9	7	11
Monthly income (ten thousand ₹)	11	7	9	5	6	8	10

Obtain the regression line of the monthly income on the years of job of the marketing executives.

2. The information collected regarding price (in ₹) of a commodity and its supply (in hundred units) is as follows.

Price (₹)	59	60	61	62	64	57	58	59
Supply (hundred units)	78	82	82	79	81	77	78	75

Obtain the regression line of the supply on the price.

3. The following information is obtained for monthly advertisement cost and the sales of the last year for a company providing online shopping.

Particulars	**Advertisement cost (ten thousand ₹)**	**Sales (lakh ₹)**
Mean	10	90
Standard Deviation	3	12
$r = 0.8$		

Obtain the regression line of the sales on the advertisement cost.

4. The following results are obtained from the information of average rain and yield of a crop per acre in the last ten years of an arid region.

Particulars	**Rainfall (cm)**	**Yield of crop (kg)**
Mean	18	970
Standard Deviation	2	38
Correlation Coefficient = 0.6		

Estimate the yield of the crop if it rains 20 cms.

5. The information of investment (in lakh ₹) and its market price (in lakh ₹) after six months in share market in the last seven years for a Mutual Fund Company is obtained as follows.

Particulars	Investment (lakh ₹) x	Market price after six months (lakh ₹) y
Mean	40	50
Variance	100	256
Covariance = 80		

Obtain the regression line of Y on X and estimate the market price in the share market after six months if there is an investment of ₹ 45 lakh in a year.

Section F

Solve the following :

1. Obtain the regression line of the demand on the price using the following information collected for the demand and the price of a commodity. Estimate the demand of the commodity if price is ₹ 40.

Price (₹)	38	36	37	37	36	38	39	36	38
Demand (hundred units)	12	18	15	12	17	13	13	15	12

2. The information regarding the experience (in years) of eight workers on a machine and their performance ratings based on the nondefective units they manufactured in every 100 units is as follows.

Experience of worker (years)	5	12	15	8	20	18	22	25
Performance rating	80	82	85	81	90	90	95	97

Obtain the regression line of the performance rating on the experience and estimate the performance rating if a worker has an experience of 17 years.

3. The information regarding daily income (in ₹) and expenditure (in ₹) of five labour families earning by daily work.

Daily income (₹)	200	300	400	600	900
Expenditure (₹)	180	270	320	480	700

Obtain the regression line of the expenditure on the daily income. Estimate the expenditure of a family having daily income of ₹ 500.

4. The following information is collected by a firm to know the effect of an advertisement campaign.

Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Advertisement cost (ten thousand ₹)	12	15	15	23	24	38	42	48
Sales (crore ₹)	5	5.6	5.8	7	7.2	8.8	9.2	9.5

Obtain the regression line of sales on the advertisement cost. Estimate the sales when the advertisement cost is ₹ 5,00,000.

5. The information of eight construction companies regarding the number of contracts received in a year and the annual profit is as follows.

No. of contracts	2	5	9	12	6	4	8	10
Annual profit (lakh ₹)	100	300	700	1000	350	250	700	750

Obtain the regression line of the annual profit on the number of contracts. Verify the reliability of the regression model.

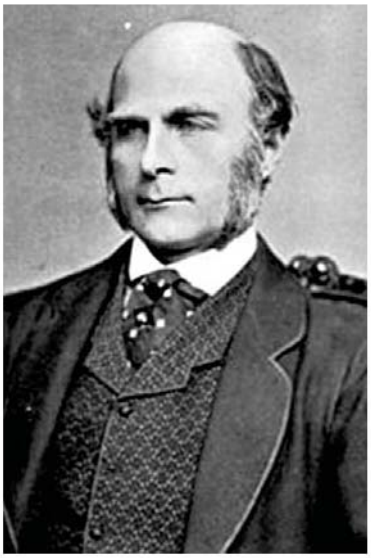
6. Obtain the regression line of Y on X from the following data and estimate Y for $X = 30$.

$n = 10,\ \Sigma x = 250,\ \Sigma y = 300,\ \Sigma xy = 7900,\ \Sigma x^2 = 6500$

7. The following results are obtained for a data.

$n = 12,\ \Sigma x = 30,\ \Sigma y = 5,\ \Sigma x^2 = 670,\ \Sigma xy = 344$

Later on, it was known that one pair (10, 14) was wrongly taken as (11, 4). By correcting the above measures, obtain the regression line of Y on X. Estimate Y for $X = 5$.



Sir Francis Galton
(1822 –1911)

Sir Francis Galton was an English Victorian statistician, progressive, polymath, sociologist, psychologist, anthropologist, eugenicist, tropical explorer, geographer, inventor, meteorologist, protogeneticist and psychometrician. He was knighted in 1909.

Galton produced over 340 papers and books. He also created the statistical concept of correlation and widely promoted regression towards the mean. He was the first to apply statistical methods to the study of human differences and inheritance of intelligence, and introduced the use of questionnaires and surveys for collecting data on human communities, which he needed for genealogical and biographical works and for his anthropometric studies.

He was a pioneer in eugenics, coining the term itself and the phrase "nature versus nature". His book Hereditary Genius (1869) was the first social scientific attempt to study genius and greatness.

As an investigator of the human mind, he founded psychometrics (the science of measuring mental faculties) and differential psychology and the lexical hypothesis of personality. He devised a method for classifying fingerprints that proved useful in forensic science.

"Imperfect prediction, despite being imperfect can be valuable for decision making process."

– Michael Kattan

4

(Time Series)

Contents :

4.1 Time Series : Introduction, meaning, importance, definition and utility

4.2 Components of time series

4.3 Time series – Trend, methods of measuring trend

4.3.1 Graphical method

4.3.2 Method of least squares

4.3.3 Method of moving averages

4.1 Time Series

Introduction

Two related variables are studied by different methods in Statistics. A special method is applied to study the dependent variable among these related variables by taking time as an independent variable. The data related to values of the variable changing with time are studied in Economics, Sociology and Business Statistics. For example, population of a country, agricultural production, wholesale price index, unemployment statistics, import-export information, annual production of a certain factory, data from share market, bank interest rates, the hourly temperature measured in a city, etc. are presented with respect to time. These data are called time series as they are dependent on time.

Meaning of Time Series

The statistical data collected at specific intervals of time and arranged in a chronological order is called Time Series. Time series consists of values of a variable associated with time. The estimated value of this variable for the future can be obtained if the values of such a variable are studied over a long period of time. These forecasts are very useful for future planning. For example, the direction, proportion and pattern of variations in the population of a certain region can be known by studying its time series. The necessary infrastructure, medical facilities, employment opportunities, education can be planned for the people of this region in future. The fluctuations in the prices of shares can be known by studying the time series of share prices of different companies and the investors can decide about buying or selling shares. The temperature measured at different places and time as well as the data of rainfall indicate the global changes in the weather which is useful to form policies for conservation of environment. In recent times, time series is extensively used in different methods of business analytics.

The data regarding changes in a variable at specific time interval are shown in the time series. The unit of time is dependent on the variable under study. For example, population data are obtained every ten years, data about total sales tax collected are available annually, quarterly interests are calculated in banks, monthly profits of shops are given, the time for bacterial growth is in hours, etc.

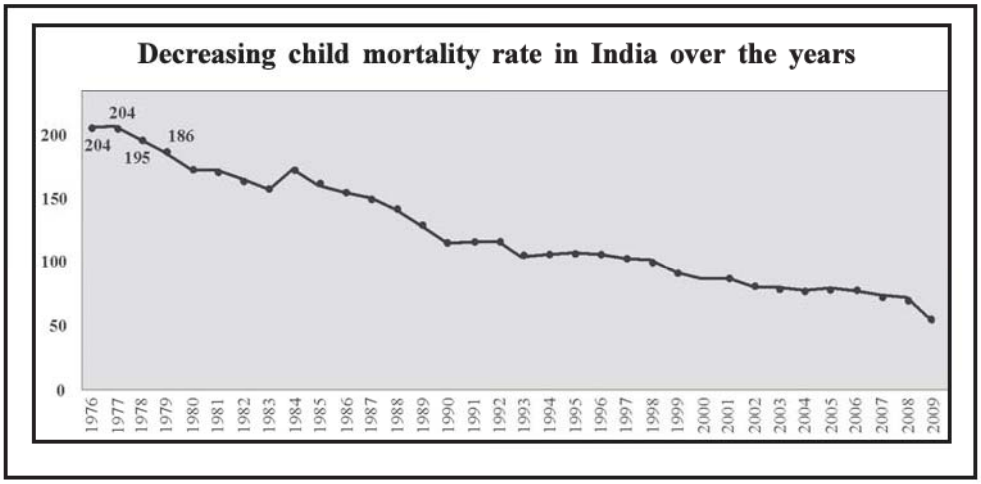
We will see the following illustrations of time series :

(1)

MACRO FUNDAMENTALS (in %)				
Year	**GDP Growth**	**Investment Growth**	**Average WPI Inflation**	**CAD (As % of GDP)**
2002-03	4.0	–0.4	3.4	0
2003-04	8.1	10.6	5.5	0
2004-05	7.0	24.0	6.5	0.4
2005-06	9.5	16.2	4.5	1.2
2006-07	9.6	13.8	6.6	1.0
2007-08	9.3	16.2	4.7	1.3
2008-09	6.7	3.5	8.1	2.3
2009-10	8.6	7.7	3.8	2.8
2010-11	9.3	14.0	9.6	2.8
2011-12	6.2	4.4	8.9	4.2
2012-13	5.4*	2.3*	7.6**	4.7*
* April-September ** April-December				

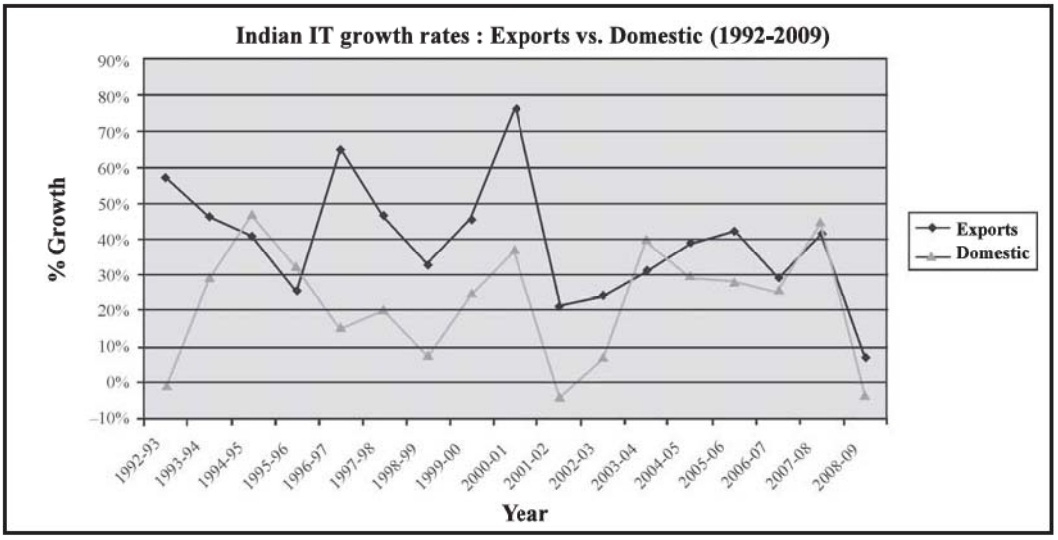
This time series gives information about macro fundamentals of different years which includes growth in Gross Domestic Product (GDP) along with percentage investment growth, Wholesale Price Index (WPI) and Current Account Deficit (CAD).

(2)



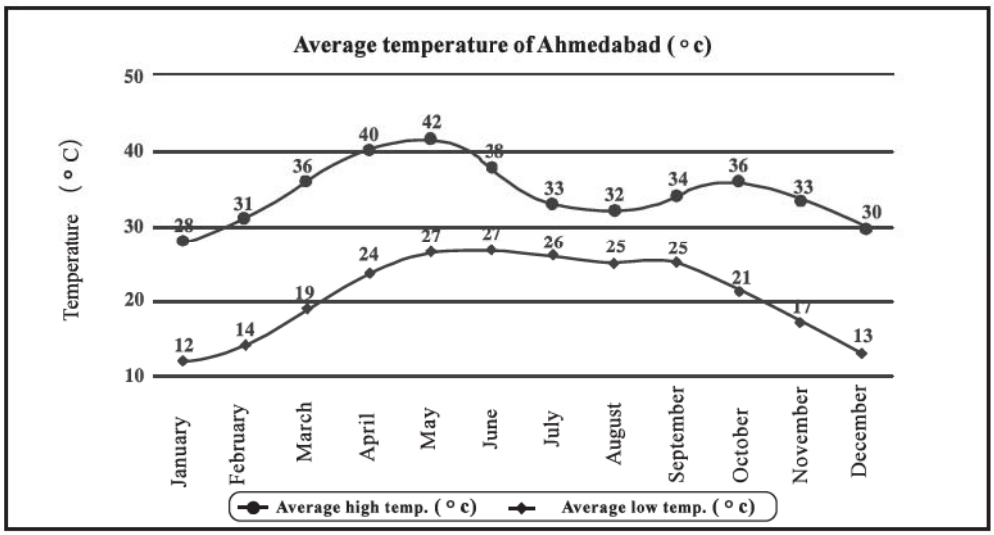
Time Series showing death rate among new-born children in India over different years.

(3)



A comparative study of two time series showing growth in IT sector.

(4)



Time Series showing average monthly maximum and minimum temperatures in Ahmedabad.

Importance of Time Series

The information collected in the form of time series is extremely essential in modern era due to increasing uncertainties in trade and business activities. The study of time series gains importance due to following reasons.

(1) The direction and pattern of variations in the values of the series can be known from the past data.

(2) The variations in the future can be estimated from the extent of variation in the values of the series.

(3) Important decisions can be taken from the estimated values of the future and industrial as well as government policies can be easily framed.

(4) Two or more industrialists or government institutions can make a comparative study from the data of the time series obtained by them.

Definition of Time Series

Time series is defined as follows :

'A time series is a set of observations taken at specified time periods.'

Usually these observations are taken at equal intervals of time.

The time is taken as an independent variable in the time series which will be denoted by t and the dependent variable associated with it will be denoted by y_t. Thus, we shall represent the time series for different units of time as follows :

Time t	1	2	3	...	n
Variable y_t	y_1	y_2	y_3	...	y_n

Uses of Time Series

The variations in the variable of a time series which changes with time are not caused by one specific reason. The variable of a time series is influenced by various factors and all these factors have an effect on the given variable. For example, the price of wheat in the wholesale market changes with time due to various different reasons such as the production of wheat at that time, demand of wheat, the cost of transporting the production to the market, etc. Each of these factors is dependent on other forces. For instance, the production of wheat is affected by various factors like the weather at that time, irrigation facility, the quality of seeds. It is necessary to study the various ways in which these factors affect the variable of the time series. Such a detailed study conducted for the time series is called as the analysis of time series which is done in the following two stages :

(1) To identify the various factors affecting the variable of the time series.

(2) To segregate these factors and determine the extent of effect of each factor on the given variable.

The analysis of time series done in this way is useful in trade, science, social and political fields as follows :

(1) It is possible to know the past situation and use it to obtain the type and measure of the variation.

(2) It is possible to estimate the value of the variable in future using statistical methods.

(3) Proper decisions can be taken for the future using the estimated values and activities can be planned accordingly.

(4) A comparative study can be carried out for the variations in the given variable at different places or time intervals.

(5) The estimates obtained from the past data can be compared with the present values and the reasons for the discrepancies between them can be investigated.

Activity

Collect the information with the help of your teachers about the percentage of students passing 12th standard from your school in the last 10 years and present it in the form of a time series.

4.2 Components of Time Series

We saw that there is a composite effect of many factors on the variable of the time series which brings fluctuations in the values of the variable. After observing different time series, it is known that the variations in the variable exhibit a specific pattern. The time series can be decomposed in the following components based on this pattern :

(1) Long-term Component or Trend : The variation seen in the variable of a time series over a very long period of time is the effect of long-term component or trend. The variable of a time series is generally found to have a continuous increase or decrease. This phenomenon is due to trend. For example, decreasing value of rupee in the international market, increasing usage of mobile phones, rising population of the country, decreasing death rates, etc. The intermittent short-term variations are ignored as the long-term variations in the time series are studied in the trend. The overall changes taking place in the variable of the series are considered here. The factors responsible for these changes produce a very slow pace of variation. For example, the number of literates are increasing in India but this change is taking place slowly in the last 60-70 years. Generally, the causes of such variations are the changing customs in the society, changes in tastes and choices of the people, technological changes in the industry, etc.

The trend of a time series is experienced after a very long time where 'long time' is a completely relative term. An interval of 10-15 years is required to know the trend in agricultural yield or industrial production whereas it may be clear within 4-5 years in the sale of electronic goods. The trend in the series having almost constant increase or decrease is called a linear trend, which is generally observed in most of the time series. The data in economics, commerce and trade have series in which the rate of increase or decrease of the values does not remain constant. The rate of increase in the values of such series is very slow initially which goes on increasing slowly. The values stabilize after a certain interval of time and then start decreasing gradually. The trend of the series in such a situation is said to be Non-linear or Curvi-linear.

We will denote the component showing trend in the variable y_t of the time series as 'T_t'.

(2) Seasonal Component : The variation occurring in the time series variable almost regularly over a very short period of time is the effect of seasonal componont. The period of oscillation of such variations is usually less than a year. It is necessary to record the short-tem values in the time series to study these variations. It is not possible to get the information about the seasonal component if the yearly values of the given variable are available. The seasonal component affects the time series as follows :

(i) Effect of natural factors : The variations in the values of the time series occur in association with the seasons or weather fluctuations. Such variations occur at almost regular intervals. For example, the demand of fans, coolers or A.C. increases during summer whereas the demand of woollens increases in winter, the market prices reduce when the new crop is ready, etc.

(ii) Effect of man-made factors : The variations occurring regularly in less than one year period are caused by the festivals, customs in the society, habits of people, etc. For example, the purchase of ornaments increases during marriage season, kites are in demand during Uttarayan, the number of customers increases at the restaurants or theatres during weekends, increase in the purchase of clothing and gift articles during festivals, etc.

As the period of oscillation of these types of variations is almost certain, they are called regular variations. If the time and measure of these variations are known then the traders, producers are benefitted as higher profit can be earned by a control over their inventory.

We shall denote this short-term component of the time series by 'S_t'.

(3) Cyclical Component : The variations occurring in the time series variable at approximately regular intervals of more than one year are the effects of cyclical component. The variations occurring due to this component are less regular as compared to seasonal component. The period of oscillation of these variations can be 2 to 10 years and in specific circumstances it can also be 10-15 years. The cyclical component is also considered as a short-term component as the time interval of variations due to this component is less than that of the time for the whole series which can be 40-50 years or even more. The cycles of boom and recession are examples of these variations. These cycles pass through the four stages namely, depression, recovery, boom, recession. These variations are found in the time series of trade and financial matters such as production, price of an item, prices of shares in share market, investment, etc. The traders can plan suitably with the help of estimate of time and measure of these variations.

This component of the time series is denoted by 'C_t'.

(4) Random or Irregular Component : The effect of irregular or random component is also seen in the variations of the variable of time series in addition to the approximately regular short-term components like seasonal and cyclical which also give short-term effect. If the values in the series change due to sudden and unpredictable causes, the changes are called as random variations. The time-interval and effect of this variation is not certain. The variation which cannot be attributed to any one of the trend, seasonal component, cyclical component is the effect of random component. These fluctuations appear completely unexpected and are irregular. It cannot be predicted, it does repeat regularly and cannot be controlled. This variation occurs due to natural disasters like earthquake, floods or due to man-made predicaments like war, strike, political upheaval. The estimates can contain error due to this component.

This component is denoted by 'R_t'.

The value of the variable of the time series y_t based on time t is determined with the combined effect of trend (T_t), seasonal component (S_t), cyclical component (C_t) and random component (R_t). This relationship is shown as follows in the additive model of the time series :

$$y_t = T_t + S_t + C_t + R_t$$

The seasonal component (S_t) does not appear if the yearly values of the variable are given for the time series. To find the effect of each component, trend (T_t) is found first using the given values y_t. After substracting it from y_t, the residual variation shows short-term components (S_t, C_t, R_t). Then seasonal component (if available) and cyclical components are found. Random component is found in the end as $R_t = y_t - (T_t + S_t + C_t)$. The future estimate of the variable $(\hat{y}_t)$ is found by estimating the trend value and then adding the effect of each component at the given time (t) as mentioned above.

Activity

Prepare a time series of units of electricity consumption for the past one year from the electricity bills of your house. Identify the component of time series showing its effect on the variation in the variable of this series.

4.3 Methods for Determining Trend

Trend is an important component of time series. We will study the following methods to estimate it :

4.3.1 Graphical Method

This is the easiest method to find trend. The points are plotted on the graph paper by taking the independent variable, time (t), on X-axis and the dependent variable y_t on Y-axis. These points are joined in their order by line segments. This shows the variation in the values of the variable. A smooth curve is then drawn through the middle of the points by personal judgement. This curve shows the trend by ignoring the short-term fluctuations in the series. The future estimates are obtained by extending the curve thus drawn.

The merits and limitations of graphical method are as follows.

Merits :

(1) This method is easy to understand and use.

(2) The trend can be found without any mathematical formula or calculations.

(3) This method can be used even if the trend is not linear.

(4) The judgement about the type of curve to be fitted for obtaining trend can be given by this method.

Limitations :

(1) It is possible that different people draw different curves. Hence, the uniformity is not maintained in the trend and its estimates.

(2) The estimates cannot be accurate as this is not a mathematical method and it is not possible to know the reliability of the estimates.

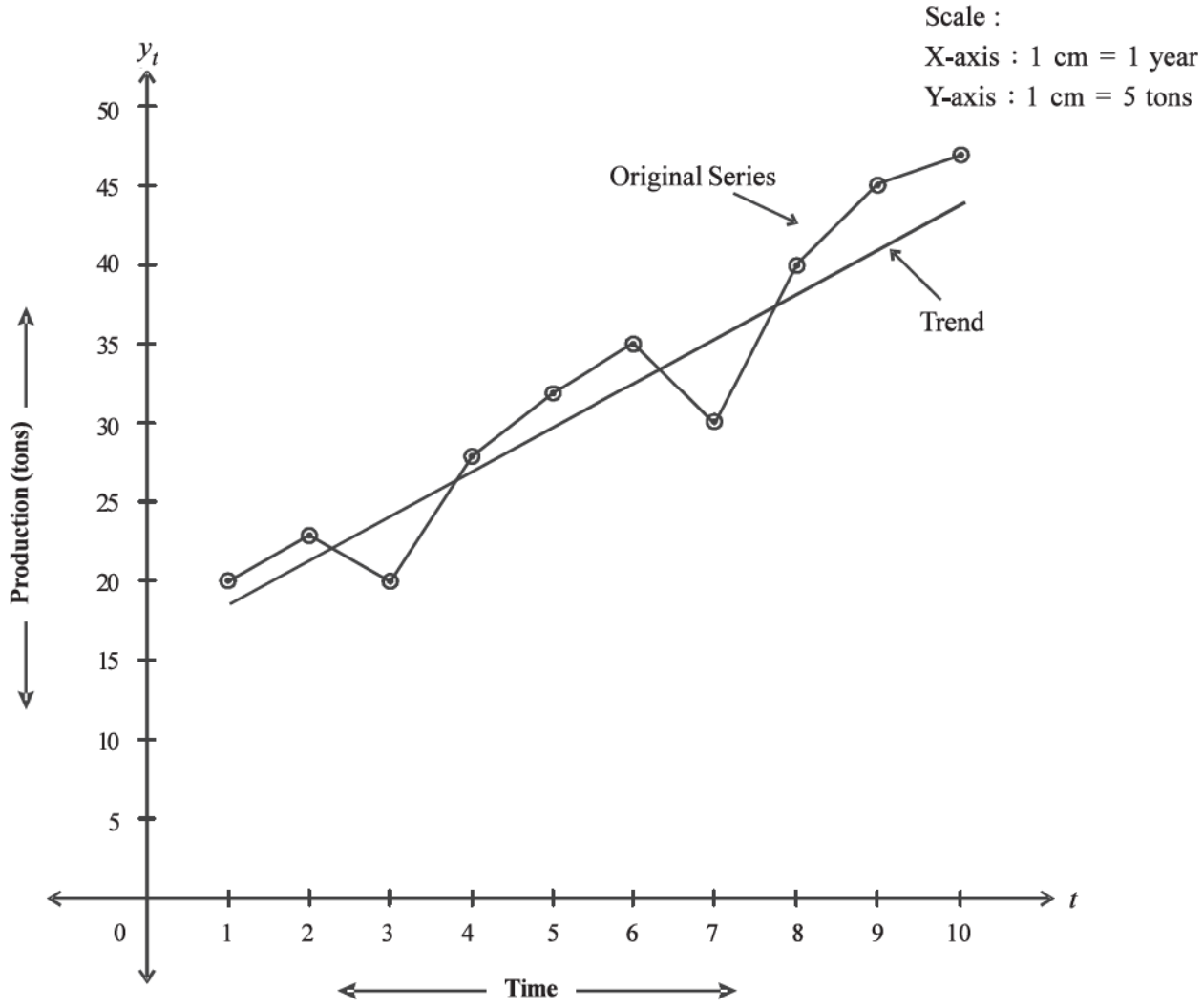
Illustration 1 : The yearly production (in tons) of a factory is as follows. Obtain the trend using graphical method.

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Production (tons)	20	23	20	28	32	35	30	40	45	47

We will represent these data as the following time series :

Time t	1	2	3	4	5	6	7	8	9	10
Production (ton) y_t	20	23	20	28	32	35	30	40	45	47

We will plot these points on a graph by taking t on X-axis and production y_t on Y-axis. The pattern of points indicates that linear trend is more suitable.



The line passing through the middle of the points shows trend.

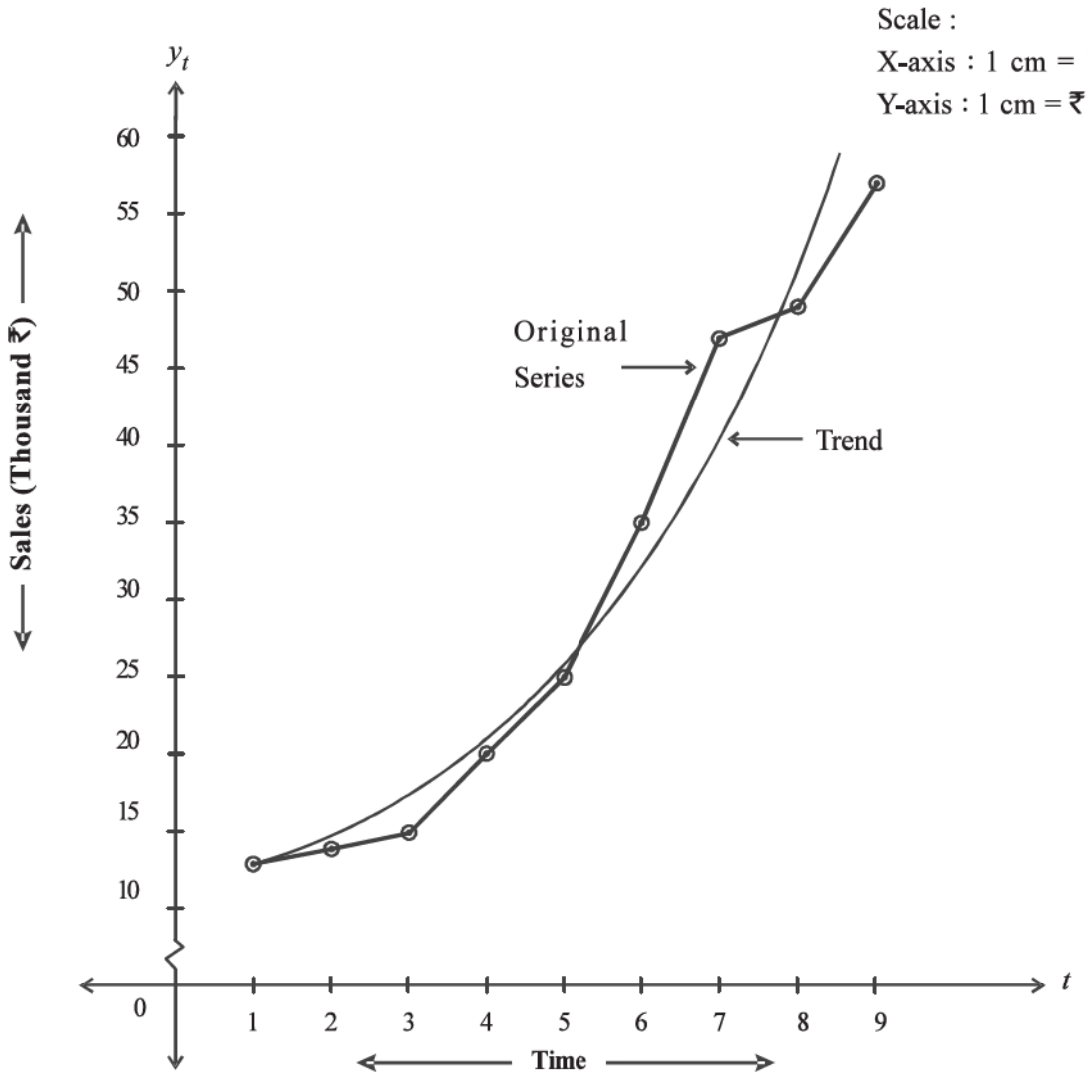
Illustration 2 : The data about monthly sales (in thousand ₹) of a company are given in the following table. Obtain trend using grapical method.

Month	Jan.	Feb.	March	April	May	June	July	August	Sept.
Sales (thousand ₹)	13	14	15	20	25	35	47	49	57

We will take the following time series for these data :

Time t	1	2	3	4	5	6	7	8	9
Sale (thousand ₹) y_t	13	14	15	20	25	35	47	49	57

We will plot these points on a graph by taking t on X-axis and sales y_t on Y-axis. It indicates that the non-linear trend is more suitable for these data.



The curve passing through the middle of the points on the graph shows trend.

EXERCISE 4.1

1. The information about the capacity (in lakh tons) to load ships at a port each year is given below. Find the linear trend using graphical method.

Year	2008	2009	2010	2011	2012	2013	2014	2015
Capacity (lakh tons)	90	97	108	111	127	148	169	200

2. The number of tourists (in thousand) visiting a certain tourist place is as follows. Find the trend using a suitable graph.

Year	2010	2011	2012	2013	2014	2015	2016
No. of tourists (thousand)	5	7	10	14	30	41	50

3. The data regarding number of girls (y_t) per 1000 boys in the age group 0-6 years of a state are given in the following table. Obtain the linear trend using graphical method.

Year	1961	1971	1981	1991	2001	2011
y_t	956	948	947	928	883	890

4. The data about the closing prices of shares of a company for 10 days are given in the following table. Obtain the trend using graphical method.

Day	1	2	3	4	5	6	7	8	9	10
Price of share (₹)	297	300	304	299	324	320	318	324	329	328

*

4.3.2 Method of Least Squares

As a limitation of the graphical method we saw that if a mathematical technique is not used then the trend and the estimates obtained from it change from person to person and its reliability cannot be known. If we want to find the linear trend of the time series by a mathematical method, we will require a specific linear equation which represents trend. We have studied the method of least squares in the chapter of regression to fit a linear equation to the given data which will be used to find the linear trend in a time series.

Suppose the values of the variable y_t in the time series are available based on time t. We shall use the linear model $y_t = \alpha + \beta t + u_t$ (where u_t is disturbance variable) to represent the relation between them. The estimated values $\hat{y}_t$ of y_t can be found by fitting this model using the method of least squares. We will use the equation $\hat{y}_t = a + bt$ for this as shown in chapter 3.

We will ignore the suffix t in y_t for simplicity and consider $\hat{y} = a + bt$. The dependent variable is y for the independent variable t.

The constants a and b are obtained by the method of least squares as follows :

$$b = \frac{n\Sigma ty - (\Sigma t)(\Sigma y)}{n\Sigma t^2 - (\Sigma t)^2} \quad \text{and} \quad a = \bar{y} - b\bar{t},$$

where n = no. of observations

The linear equation thus obtained is the best linear equation for the given data.

The estimate of trend for the future is obtained using this linear equation.

Note : Other equations like polynomial, exponential equations can also be fitted besides the linear equation to find trend.

The merits and limitations of the method of least squares are as follows.

Merits :

(1) This method is absolutely mathematical and hence the future estimates do not change subjectively with the person.

(2) The trend estimates can be obtained by this method for each of the given values of t.

(3) The trend estimates can also be obtained for intermediate periods as the trend values are obtained using an equation. For example, the trend estimate for the period in the centre of the second and third year can be found by taking $t = 2.5$.

Limitations :

(1) This method requires extensive calculations to find trend.

(2) The reliability of the estimated values obtained by this method is less if an appropriate type of trend curve and its suitable equation is not fitted.

Illustration 3 : The profit earned (in lakh ₹) by a company making computers is as follows. Find the linear equation for the trend from these data by least square method and estimate the profit for the year 2017.

Year	2011	2012	2013	2014	2015
Profit (Lakh ₹)	31	35	39	41	44

The values of profit are given for $n = 5$ years. We will thus denote the given years as $t = 1, 2, \ldots, 5$ respectively.

Calculation for fitting linear trend

Year	**Profit** y	t	t^2	ty
2011	31	1	1	31
2012	35	2	4	70
2013	39	3	9	117
2014	41	4	16	164
2015	44	5	25	220
Total	**190**	**15**	**55**	**602**

$$\bar{t} = \frac{\Sigma t}{n} = \frac{15}{5} = 3, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{190}{5} = 38$$

$$b = \frac{n\Sigma ty - (\Sigma t)(\Sigma y)}{n\Sigma t^2 - (\Sigma t)^2}$$

$$= \frac{5 \times 602 - 15 \times 190}{5 \times 55 - (15)^2}$$

$$= \frac{3010 - 2850}{275 - 225}$$

$$= \frac{160}{50}$$

$$= 3.2$$

$$a = y - b\bar{t}$$

$$= 38 - 3.2 \times 3$$

$$= 38 - 9.6$$

$$= 28.4$$

Equation for trend $\hat{y} = a + bt$

$$\therefore \hat{y} = 28.4 + 3.2\,t$$

We take $t = 7$ for the year 2017.

$\therefore \quad \hat{y} = 28.4 + 3.2 \times 7$

$= 28.4 + 22.4$

$= 50.8$

$\therefore \quad \hat{y} =$ ₹ 50.8 lakh

Thus, the estimated trend value of profit for the year 2017 is ₹ 50.8 lakh.

Illustration 4 : The dropout rate of students of standard 1 to 5 from primary schools of a district is as follows :

Year	2009–10	2010–11	2011–12	2012–13	2013–14	2014–15	2015–16
Dropout rate	3.24	2.98	2.29	2.20	2.09	2.07	2.04

Estimate the dropout rate for students from standard 1 to 5 for the year 2016-17 and 2017-18 by fitting a linear equation for trend.

The data are given for $n = 7$ years. We will thus denote the given years as $t = 1, 2, \ldots, 7$ respectively.

Calculations for fitting linear trend

Year	**Dropout rate y**	t	t^2	ty
2009-10	3.24	1	1	3.24
2010-11	2.98	2	4	5.96
2011-12	2.29	3	9	6.87
2012-13	2.20	4	16	8.80
2013-14	2.09	5	25	10.45
2014-15	2.07	6	36	12.42
2015-16	2.04	7	49	14.28
Total	**16.91**	**28**	**140**	**62.02**

$$\bar{t} = \frac{\Sigma t}{n} = \frac{28}{7} = 4, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{16.91}{7} = 2.4157 \simeq 2.42$$

$$b = \frac{n\Sigma ty - (\Sigma t)(\Sigma y)}{n\Sigma t^2 - (\Sigma t)^2}$$

$$= \frac{7 \times 62.02 - 28 \times 16.91}{7 \times 140 - (28)^2}$$

$$= \frac{434.14 - 473.48}{980 - 784}$$

$$= \frac{-39.34}{196}$$

$$= -0.2007$$

$$\simeq -0.2$$

$$a = \bar{y} - b\bar{t}$$

$$= 2.42 - (-0.2) \times 4$$

$$= 2.42 + 0.8$$

$$= 3.22$$

Equation for trend $\hat{y} = a + bt$

$$\therefore \hat{y} = 3.22 + (-0.2)\, t$$

$$= 3.22 - 0.2\, t$$

We take $t = 8$ for the year 2016-17.

$$\therefore \hat{y} = 3.22 - 0.2 \times 8$$

$$= 3.22 - 1.6$$

$$= 1.62$$

We take $t = 9$ for the year 2017-18.

$$\therefore \hat{y} = 3.22 - 0.2 \times 9$$

$$= 3.22 - 1.8$$

$$= 1.42$$

Thus, the estimates of dropout rates for the students of standard 1 to 5 in this district for the years 2016-17 and 2017-18 are 1.62 and 1.42 respectively.

Illustration 5 : The data of population (in lakh) of a taluka are given in the following table. Fit a linear equation for the data and find the trend value for each year. Also find the trend estimate for the population in the year 2021.

Year	1951	1961	1971	1981	1991	2001	2011
Population (lakh)	15.1	16.9	18.7	20.1	21.6	25.7	27.1

The data about population are given which are associated with each decade. We will take $t = 1, 2, \ldots, 7$ respectively for the given years. Hence, we get $n = 7$.

Calculation for fitting linear trend

Year	**Population (lakh) y**	t	t^2	ty	**Trend values $\hat{y} = 12.66 + 2.02\ t$**
1951	15.1	1	1	15.1	14.68
1961	16.9	2	4	33.8	16.7
1971	18.7	3	9	56.1	18.72
1981	20.1	4	16	80.4	20.74
1991	21.6	5	25	108	22.76
2001	25.7	6	36	154.2	24.78
2011	27.1	7	49	189.7	26.8
Total	**145.2**	**28**	**140**	**637.3**	

$$\bar{t} = \frac{\Sigma t}{n} = \frac{28}{7} = 4\ , \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{145.2}{7} = 20.7429 \simeq 20.74$$

$$b = \frac{n\Sigma ty - (\Sigma t)(\Sigma y)}{n\Sigma t^2 - (\Sigma t)^2}$$

$$= \frac{7 \times 637.3 - 28 \times 145.2}{7 \times 140 - (28)^2}$$

$$= \frac{4461.1 - 4065.6}{980 - 784}$$

$$= \frac{395.5}{196}$$

$$= 2.0179$$

$$\simeq 2.02$$

$$a = \bar{y} - b\bar{t}$$

$$= 20.74 - 2.02 \times 4$$

$$= 20.74 - 8.08$$

$$= 12.66$$

Equation for trend $\hat{y} = a + bt$

$$\therefore\ \hat{y} = 12.66 + 2.02\ t$$

We take $t = 1, 2, \ldots, 7$ for each of the given year respectively to find the values of trend.

Taking $t = 1$,

$$\hat{y} = 12.66 + 2.02 \times 1$$

$$= 12.66 + 2.02$$

$$= 14.68$$

$\therefore \quad \hat{y} = 14.68$ lakh

Similarly we will take $t = 2, 3, \ldots, 7$ to find the remaining values of trend and show them in the table.

It can be seen here that the values of $\hat{y}$ increase successively by 2.02.

We take $t = 8$ for the year 2021

$$\hat{y} = 12.66 + 2.02 \times 8$$

$$= 12.66 + 16.16$$

$$= 28.82$$

$\therefore \quad \hat{y} = 28.82$ lakh

Thus, the estimate for trend value for the population for the year 2021 is 28.82 lakh.

Illustration 6 : The data about monthly sales (in thousand ₹) of a company are given in the following table. Fit a linear trend and show it grapically. Estimate the sale for the month of August using the equation obtained.

Month	January	February	March	April	May	June
Sale (thousand ₹)	80	85	90	76	82	88

The data for $n = 6$ months are given here. Hence, we will take $t = 1, 2, \ldots, 6$ for the given months respectively.

Calculation for fitting linear Trend

Month	**Sales y (thousand ₹)**	t	t^2	ty	$\hat{y} = 81.79 + 0.49\ t$
January	80	1	1	80	82.28
February	85	2	4	170	82.77
March	90	3	9	270	83.26
April	76	4	16	304	83.75
May	82	5	25	410	84.24
June	88	6	36	528	84.73
Total	**501**	**21**	**91**	**1762**	

$$\bar{t} = \frac{\Sigma t}{n} = \frac{21}{6} = 3.5, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{501}{6} = 83.5$$

$$b = \frac{n\Sigma ty - (\Sigma t)(\Sigma y)}{n\Sigma t^2 - (\Sigma t)^2}$$

$$= \frac{6 \times 1762 - 21 \times 501}{6 \times 91 - (21)^2}$$

$$= \frac{10572 - 10521}{546 - 441}$$

$$= \frac{51}{105}$$

$$= 0.4857$$

$$\simeq 0.49$$

$$a = \bar{y} - b\bar{t}$$

$$= 83.5 - 0.49 \times 3.5$$

$$= 83.5 - 1.715$$

$$= 81.785$$

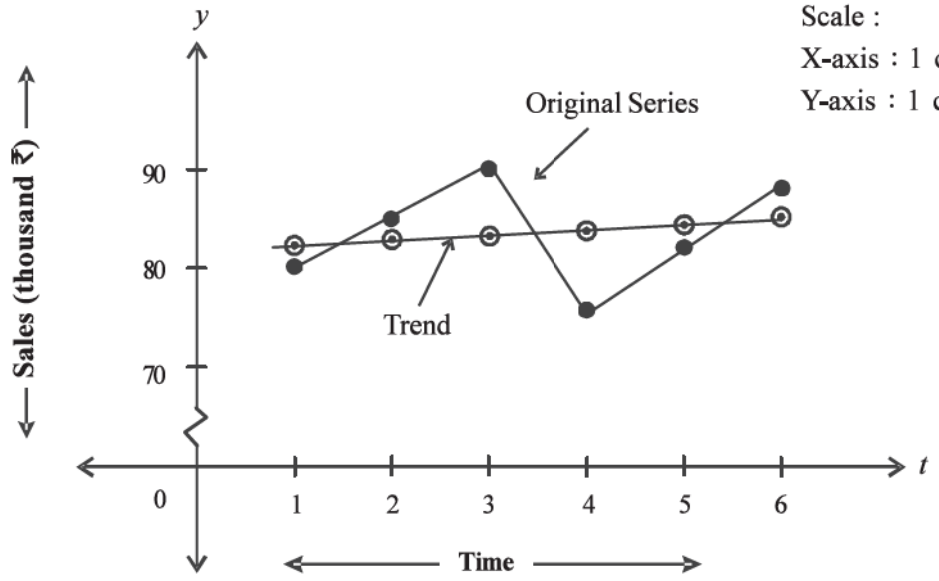
$$\simeq 81.79$$

Equation for trend $\hat{y} = a + bt$

$$\therefore \hat{y} = 81.79 + 0.49\,t$$

By substituting t = 1, 2, ... , 6 successively, we get the corresponding values of $\hat{y}$ which are shown in the table.

The trend values and the values in the given series can be shown in the graph as follows :



Now we take $t = 8$ for the month of August

$$\hat{y} = 81.79 + 0.49 \times 8$$

$$= 81.79 + 3.92$$

$$= 85.71$$

$$\therefore \hat{y} = ₹\ 85.71 \text{ thousand}$$

Thus, the trend estimate for sales of this company in the month of August is ₹ 85.71 thousand.

Note : It is not necessary to take all the values of $\hat{y}$ to draw the line representing the linear equation. The equation of trend can be shown in the graph by joining the values of $\hat{y}$ corresponding to any two values among $t = 1, 2, \ldots, 6$.

Illustration 7 : Obtain the linear equation for trend for a time series with $n = 8$, $\Sigma y = 344$, $\Sigma ty = 1342$

Since $n = 8$, we take $t = 1, 2, \ldots, 8$ Hence, we get $\Sigma t = 1 + 2 + \ldots + 8 = 36$ and

$\Sigma t^2 = 1^2 + 2^2 + \ldots + 8^2 = 1 + 4 + \ldots + 64 = 204$.

$$\bar{t} = \frac{\Sigma t}{n} = \frac{36}{8} = 4.5, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{344}{8} = 43$$

$$b = \frac{n\Sigma ty - (\Sigma t)(\Sigma y)}{n\Sigma t^2 - (\Sigma t)^2}$$

$$= \frac{8 \times 1342 - 36 \times 344}{8 \times 204 - (36)^2}$$

$$= \frac{10736 - 12384}{1632 - 1296}$$

$$= \frac{-1648}{336}$$

$$= -4.9048$$

$$\simeq -4.9$$

$$a = \bar{y} - b\bar{t}$$

$$= 43 - (-4.9) \times 4.5$$

$$= 43 + 22.05$$

$$= 65.05$$

Equation for trend $\hat{y} = a + bt$

$$= 65.05 + (-4.9)\, t$$

$$= 65.05 - 4.9\, t$$

EXERCISE 4.2

1. The information about death rate of a state in different years is given in the following table. Fit a linear equation to find trend and hence estimate the death rate for the year 2017.

Year	2009	2010	2011	2012	2013	2014	2015
Death rate	7.6	6.9	7.1	7.3	7.2	6.9	6.9

2. The data about Cost Inflation Index (CII) declared by the central government are as follows. The year 1981-82 is the base for this index. Find the estimate of this index for the year 2015-16 by fitting the linear equation to these data.

Year	2007 – 08	2008 – 09	2009 – 10	2010 – 11	2011 – 12	2012 – 13	2013 – 14	2014 – 15
CII	551	582	632	711	785	852	939	1024

3. The number of two wheelers registered (in thousand) in a city in different years is as follows. Use the method of fitting linear equation to these data to obtain the estimates for the number of vehicles registered in the year 2016 and 2017. Also find the trend values for each year.

Year	2010	2011	2012	2013	2014	2015
No. of vehicles (thousand)	69	75	82	91	101	115

4. The average age of women (in years) at the time of marriage obtained from the data of different census surveys in India are given in the following table. Fit an equation for a linear trend from the data and show it on a graph. Find the estimate for the value of the given variable for the year 2021 using the linear equation.

Year of census survey	1971	1981	1991	2001	2011
Average age of women at marriage (years)	17.7	18.7	19.3	20.2	22.2

*

4.3.3 Method of Moving Averages

The method of moving averages is very useful to find trend by eliminating the effect of short-term variations. The short-term variations are usually regular and have repetitions. The period of repetition of these variations can be found by past experience or other techniques and the average is found for the number of observations corresponding to this period. Since the average value lies in the center, we get the values that are free from the short-term fluctuations which show the trend.

Suppose, the values of the dependent variable in the given time series are $y_1, y_2, \ldots, y_n$ for time $t = 1, 2, \ldots, n$ respectively and the interval for short-term cyclical fluctuations is 3 years. The mean of first three observations y_1, y_2, y_3 is found as $\frac{y_1 + y_2 + y_3}{3}$ and it is written against the center of these three values which is y_2. Further, the mean of successive three values y_2, y_3, y_4 is found as $\frac{y_2 + y_3 + y_4}{3}$ and it is written against y_3 which is the center of these three values. Similarly, all the means are calculated till the last value from the given values of the variable is included. The averages thus calculated are called three yearly moving averages which indicate the trend.

It is not necessary that the unit of time in every time series is year and time interval for repetitions in the pattern of the values of the variable may not be necessarily three years. The moving averages will be denoted according to the unit of time. For example, 5 days moving averages, three monthly moving averages, 4 weekly moving averages, etc. The unit for time is taken as 'year' for the discussion here.

While calculating for the given data, we first find the total of values of the variable in accordance with the time interval for the average. After finding the first total $y_1 + y_2 + y_3$, the next total namely $y_2 + y_3 + y_4$ is found by subtracting y_1 from the above total and then adding y_4 to it. All the successive totals are found in this manner and each total is divided by 3 to obtain three yearly moving averages.

Note : The first three yearly moving average is written against y_2 and thus the moving average against y_1 i.e. the trend value at that time cannot be obtained. Similarly, the trend value corresponding to y_n cannot be obtained.

Illustration 8 : The number of accounts opened in different weeks in a branch of a certain bank are given below. Find the trend using three-weekly moving averages.

Week	1	2	3	4	5	6	7	8	9	10
No. of accounts opened	26	27	26	25	22	24	25	23	22	21

Calculation for three weekly moving averages

Week t	No. of accounts opened y	Three weekly moving total	Three weekly moving average
1	26	—	—
2	27	26 + 27 + 26 = 79	$\frac{79}{3}$ = 26.33
3	26	79 – 26 + 25 = 78	$\frac{78}{3}$ = 26
4	25	78 – 27 + 22 = 73	$\frac{73}{3}$ = 24.33
5	22	73 – 26 + 24 = 71	$\frac{71}{3}$ = 23.67
6	24	71 – 25 + 25 = 71	$\frac{71}{3}$ = 23.67
7	25	71 – 22 + 23 = 72	$\frac{72}{3}$ = 24
8	23	72 – 24 + 22 = 70	$\frac{70}{3}$ = 23.33
9	22	70 – 25 + 21 = 66	$\frac{66}{3}$ = 22
10	21	—	—

Illustration 9 : Find the trend using five yearly moving averages for the following data about yearly production (in tons) of a factory.

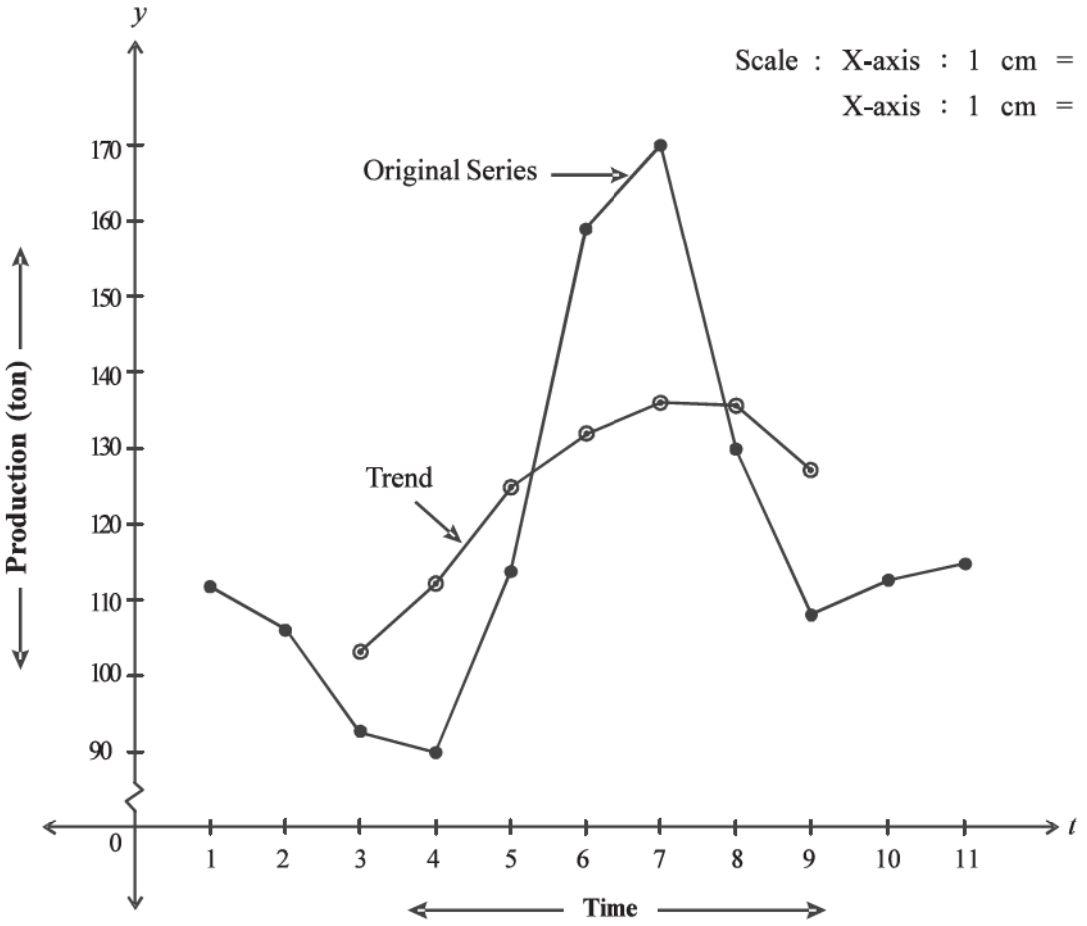
Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Production (tons)	112	106	93	90	114	159	170	130	108	113	115

Calculation for five yearly moving averages

Year	Production y	t	Five yearly moving total	Five yearly moving average (trend)
2006	112	1	—	—
2007	106	2	—	—
2008	93	3	112+106+93+90+114=515	$\frac{515}{5}$ =103
2009	90	4	515–112+159=562	$\frac{562}{5}$ =112.4
2010	114	5	562–106+170=626	$\frac{626}{5}$ =125.2
2011	159	6	626–93+130=663	$\frac{663}{5}$ =132.6
2012	170	7	663–90+108=681	$\frac{681}{5}$ =136.2
2013	130	8	681–114+113=680	$\frac{680}{5}$ =136
2014	108	9	680–159+115=636	$\frac{636}{5}$ =127.2
2015	113	10	—	—
2016	115	11	—	—

Additional information for understanding

We shall show the values of the variable and the trend obtained by five yearly moving averages to understand the trend found by this method.



The curve passing through all the averages shows the trend.

If the interval of time for the moving averages is odd number like 3, 5, 7, ... then the trend is found as shown earlier. But the calculation of moving averages becomes difficult if this interval is an even number.

Suppose four yearly moving averages are to be found. First four yearly average will be found as $\frac{y_1 + y_2 + y_3 + y_4}{4}$. As the center for these four values is between y_2 and y_3, this average will be written at that position. Similarly, the successive averages namely $\frac{y_2 + y_3 + y_4 + y_5}{4}$, $\frac{y_3 + y_4 + y_5 + y_6}{4}$, ... will be found and written between y_3 and y_4, between y_4 and y_5, ... respectively. Since these averages are in between two years, the average of each pair of averages is found and it is written between two moving averages. Thus, the average value of the first two averages shown above will be written against y_3. The averages thus obtained are called as four yearly moving averages. The processes of finding an average is to be done twice here. To simplify these calculations, four yearly totals are obtained first and then totals of pairs of years are found. As these totals involve 8 values, each total is divided by 8 which gives the four yearly averages mentioned above.

Whenever the time of the cycles of short-term variations is an even number, moving averages are obtained by first finding the moving totals and then the pairwise totals as shown in the above method.

Illustration 10 : Find the trend using four monthly moving averages for the following data showing monthly sales (in lakh ₹) of a shop.

Month	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
Sales (lakh ₹)	5	3	7	6	4	8	9	10	8	9

Calculation of four monthly moving averages

Month	**Sales (lakh ₹) *y***	***t***	**Four monthly moving total**	**Pairwise total**	**Four monthly moving average**
March	5	1		—	—
			—		
April	3	2		—	—
			$5+3+7+6=21$		
May	7	3		$21+20=41$	$\frac{41}{8}=5.13$
			$21-5+4=20$		
June	6	4		$20+25=45$	$\frac{45}{8}=5.63$
			$20-3+8=25$		
July	4	5		$25+27=52$	$\frac{52}{8}=6.5$
			$25-7+9=27$		
August	8	6		$27+31=58$	$\frac{58}{8}=7.25$
			$27-6+10=31$		
September	9	7		$31+35=66$	$\frac{66}{8}=8.25$
			$31-4+8=35$		
October	10	8		$35+36=71$	$\frac{71}{8}=8.88$
			$35-8+9=36$		
November	8	9		—	—
			—		
December	9	10		—	—

The trend of the time series is shown by the four monthly moving averages.

Merits and limitations of the method of moving averages are as follows :

Merits :

(1) The effect of short-term component is eliminated to a large extent using the averages and the trend of the series is obtained.

(2) The calculation is easy to understand as it is comparatively less and simple.

Limitations :

(1) The trend obtained by this method is not accurate if the interval for the moving averages is not chosen correctly.

(2) The estimates of trend for some initial and last time periods cannot be obtained.

(3) A specific mathematical formula is not obtained for future estimates.

EXERCISE 4.3

1. Find the trend by three yearly moving averages from the following data about the sales (in ten lakh ₹) of a company.

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Sales (ten lakh) ₹	3	4	8	6	7	11	9	10	14	12

2. The average monthly closing prices of shares of a company in the year 2016 are given in the following table. Find the trend using four monthly moving averages.

Month	January	February	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
Share price (₹)	253	231	350	261	262	266	263	261	281	278	278	272

3. Find the trend using five yearly moving averages from the following data of profit (in lakh ₹) of a trader in different years.

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Profit (lakh ₹)	15	14	18	20	17	24	27	25	23

4. The wholesale price index numbers for different quarters (Q) of a year are obtained as follows. Find the trend by four quarterly moving averages.

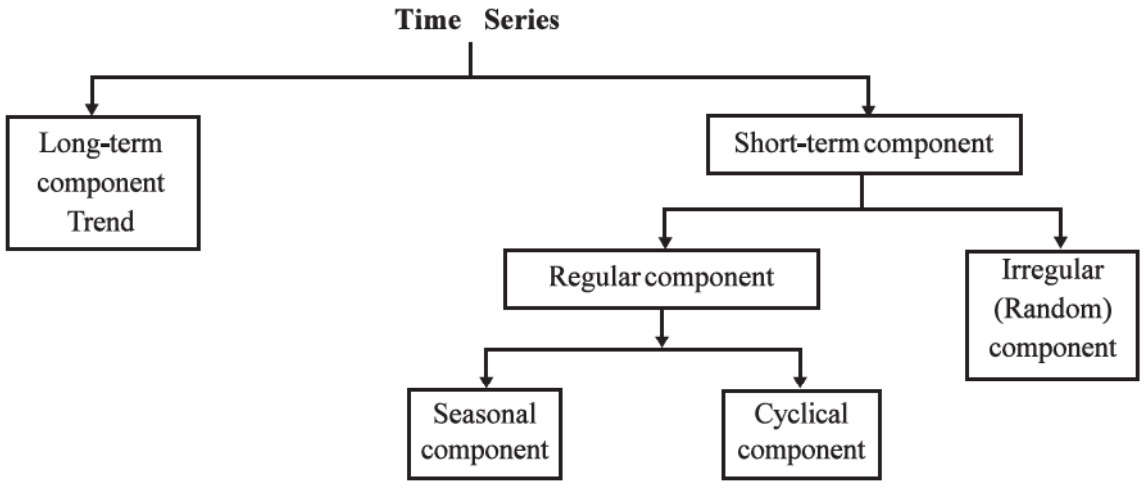
Year	2013				2014				2015			
Quarter	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
Index No.	110	110	125	135	145	152	155	168	131	124	132	153

*

Summary

- The data collected and arranged according to time is called Time Series.
- It is necessary to analyse the time series to find the future estimates of the given variable.
- There are four main components affecting the values of the variable in a time series :
  (1) Long-term Component (Trend) (2) Seasonal Component
  (3) Cyclical Component (4) Random (Irregular) Component
- Short-term fluctuations are found in the time series due to seasonal component, cyclical component and random component.
- Seasonal and cyclical fluctuations repeat almost regularly.
- Three methods of measuring trend :
  (1) Graphical Method (2) Method of least squares (3) Method of moving average

Chapter at a glance



List of formulae :

For fitting a linear equation $\hat{y} = a + bt$ to the given data

$$b = \frac{n\Sigma ty - (\Sigma t)(\Sigma y)}{n\Sigma t^2 - (\Sigma t)^2}, \qquad a = \bar{y} - b\bar{t}$$

Exercise 4

Section A

Find the correct option for the following multiple choice questions :

1. Which type of variations are produced in the time series variable due to seasonal component ?
(a) Long-term (b) Irregular (c) Regular (d) Zero

2. Which variation is shown in 'decrease in the production of a company' due to strike ?
(a) Random (b) Trend (c) Seasonal (d) Cyclical

3. Name the method for fitting the linear equation to find linear trend.
(a) Graphical Method (b) Method of least squares
(c) Method of moving average (d) Method of partial average

4. How do you show the additive model of the time series ?

(a) $y_t = T_t + S_t + C_t - R_t$ (b) $y_t = T_t + S_t + C_t + R_t$

(c) $y_t = T_t \times S_t + C_t \times R_t$ (d) $y_t = S_t + C_t + R_t$

5. State the independent variable of time series.

(a) y_t (b) S_t (c) t (d) x_t

6. Which component of the time series is impossible to predict ?

(a) Random component (b) Trend (c) Seasonal component (d) Cyclical component

7. Which of the following variations are due to cyclical component ?

(a) Rise in demand during winter

(b) Decrease in the share prices due to recession in share market

(c) Decrease in the agricultural produce due to excessive rains

(d) Continuously decreasing death rate

8. The trend equation obtained from a time series from January 2016 to December 2016 is $\hat{y} = 30.1 + 1.5\,t$. Find the value of trend for April 2016.

(a) 30.1 (b) 34.6 (c) 36.1 (d) 33.1

9. Which of the following fluctuations is the effect of seasonal component ?

(a) Increase in the migration to cities from rural areas

(b) Increasing number of vehicles on roads in a city

(c) Increase in the number of tourists during school vacation

(d) Increased death rate during a certain epidemic

10. Which method of finding trend is best to eliminate the effect of repetitive short-term variations ?

(a) Graphical Method (b) Method of least squares

(c) Karl Pearson's method (d) Method of moving average

Section B

Answer the following questions in one sentence :

1. Give an example of time series having decreasing trend.
2. What is a time series ?
3. Which of the components of time series produce short-term variations ?
4. What is meant by analysis of time series ?
5. What is the notation to show the cyclical component of the time series ?
6. State the names of methods of measuring trend.
7. The effect of which component indicates fluctuations repeating within one year ?
8. State the components of time series.
9. When is the method of moving average more useful to find trend ?
10. The linear equation fitted using the data of 7 weeks for a variable y is $\hat{y} = 25.1 - 1.3\,t$. Estimate the value of y for the eighth week.

Section C

Answer the following questions :

1. Describe the additive model of time series.
2. What is meant by cyclical component ?
3. How does seasonal component differ from the cyclical component ?
4. Explain the irregular component.
5. State the limitations of graphical method.
6. Explain the meaning of moving average.
7. Define time series.
8. State the merits of the method of moving average to measure trend.
9. Describe the graphical method to measure trend.

Section D

Answer the following questions :

1. Explain the importance of time series.
2. State the uses of analysis of time series.
3. What is meant by trend of a time series ? Explain with an illustration.
4. Write a short note on seasonal component.
5. Explain the method of fitting a linear equation to the given data using the method of least squares.
6. State the merits and limitations of the method of least squares.
7. Describe the method of moving average to find trend.
8. Discuss the limitations of the method of moving average.
9. The following time series shows the daily production of a factory. Find the trend using graphical method.

Day	1	2	3	4	5	6	7	8	9	10
Production (units)	21	22	23	25	24	22	25	26	27	26

10. Fit a linear equation from the following data for variable (y) of a time series.
    $n = 4, \quad \Sigma y = 270, \quad \Sigma ty = 734$
11. The data collected about the demand of a commodity from a store are as follows. Find the trend using three monthly moving averages.

Month	January	February	March	April	May	June	July
Demand (units)	15	16	18	18	23	23	20

Section E

Solve the following :

1. The data about exports (in crore ₹) of ready-made garments of a textile manufacturer are shown below :

Year	2010	2011	2012	2013	2014	2015
Export (crore ₹)	22	25	23	26	20	25

Fit a linear trend to these data and estimate the trend for the export in the year 2017.

2. The following data are available for the number of passengers who travelled in the last 5 years by the aircrafts of an airline company. Estimate the trend for the year 2016 by fitting linear trend.

Year	2011	2012	2013	2014	2015
No. of passengers (thousands)	45	47	44	40	38

3. The data about closing prices of shares of a company registered in a stock exchange for different months is given in the following table. Find the trend using three monthly moving averages.

Month	2015 April	May	June	July	August	Sept.	Oct.	Nov.	Dec.	2016 January
Share price (₹)	76	73	65	68	67	60	63	67	65	66

4. The following data show the sales (in thousand ₹) of a commodity. Find the trend by graphical method.

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Sales (thousand ₹)	200	216	228	235	230	232	236	235	230	233

5. The quantity index numbers of consumption of edible oil in a state are given in the following table. Find the trend using five yearly moving averages.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Index No.	115	121	119	120	117	119	120	118	116	124	125

Section F

Solve the following:

1. Find a linear equation using the method of least squares for the trend of production from the following data about sugar production of a country recorded for the last 6 years. Find the trend estimates for the production of the year 2016-17 and 2017-18.

Year	2009 – 10	2010 – 11	2011 – 12	2012 – 13	2013 – 14	2014 – 15
Sugar production (crore tons)	29.2	34.2	35.4	36.4	33.6	37.7

2. The number of students studying in a college are shown in the following table. Find the trend by four yearly moving averages.

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
No. of students	332	317	357	392	402	405	410	427	405	438

3. The birth rates of a state in different years are given in the following table. Fit a linear trend for these data. Also find the estimates for birth rates in the year 2016 and 2017.

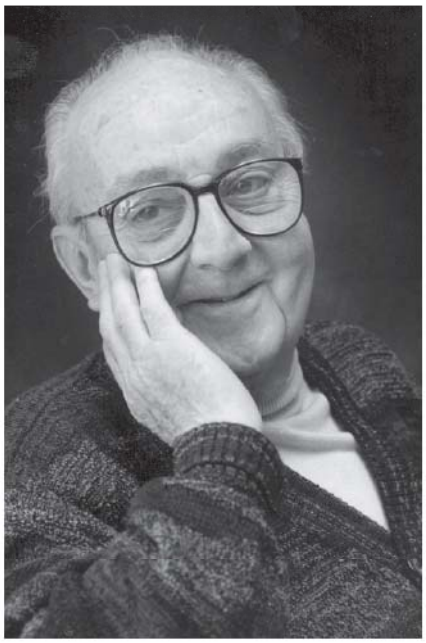
Year	2009	2010	2011	2012	2013	2014	2015
Birth rate	22.2	21.8	21.3	20.9	20.6	20.2	19.9

4. The data about goods transported in different years by a division of railways are given below. Find the estimates for each year by fitting a linear equation and represent it by a graph. Also find the estimate for the year 2016.

Year	2011	2012	2013	2014	2015
Goods transported (tons)	180	192	195	204	202

5. The data of weekly prices (in USD per barrel) of crude oil are given in the following table. Find the trend using four weekly moving averages.

Month	March 2016				April 2016				May 2016			
Week	1	2	3	4	1	2	3	4	1	2	3	4
Price of Crude oil	35.92	38.50	39.44	39.46	36.79	39.72	40.36	43.73	45.92	44.66	46.21	48.45



George Edward Pelham Box
(1919 -2013)

George E. P. Box worked in the areas of quality control, time series analysis, design of experiments and Bayesian inference. He has been called "one of the greatest statistical minds of the 20th century." He has been associated with University at Raleigh (now North Carolina State University), Princeton University, University of Wisconsin–Mandison. Box has published numerous articles and papers and he is an author of many books. He is a recipient of prestigious honours, medals and was the president of American Statistical Association in 1978 and of the Institute of Mathematical Statistics in 1979. His name is associated with results in statistics such as Box–Jenkins models, Box–Cox transformations, Box–Behnken designs, and others. Box was elected a member of the American Academy of Arts and Sciences in 1974 and a Fellow of the Royal Society (FRS) in 1985.

Answers

Exercise 1.1

1. (1) Fixed base index numbers : 100, 103.27, 105.09, 106.55, 108, 113.82, 119.27, 125.45

(2) Chain base index numbers : 100, 103.27, 101.76, 101.38, 101.37, 105.39, 104.79, 105.18

(3) Index numbers using average wage : 91.36, 94.35, 96.01, 97.34, 98.67, 103.99, 108.97, 114.62

2. (1) Fixed base index numbers : 100, 101.79, 105.36, 107.14, 110.71, 114.29,121.43, 128.57

(2) Chain base index numbers : 100, 101.79, 103.51, 101.69, 103.33, 103.23, 106.25, 105.88

(3) Index numbers using average price : 96.55, 98.28, 101.72, 103.45, 106.90, 110.34, 117.24, 124.14

3. (1) Fixed base index numbers : 100, 108.70, 112.78, 115.19, 119.44

(2) Chain base general price index numbers : 100, 108.70, 103.65,102.26, 103.71

4. General index number of n items : 126.45; Overall increase in the price of fuel items is 26.45 %

Exercise 1.2

1. Fixed base index numbers : 100, 110, 104.5, 112.86, 135.43, 143.56, 157.92

2. Chain base index numbers : 117.4, 100.51, 102.80, 103.13, 102.64, 102.49, 102.28

3. Chain base index numbers : 100, 99.63, 99.26, 100, 103.73, 101.80, 100, 103.53, 100, 102.05

4. Fixed base index numbers : 110, 123.2, 134.29, 145.03, 152.28, 169.03

Exercise 1.3

1. $I = 307$, prices have increased by 207 %.

2. $I = 123.80$, prices have increased by 23.80 %.

3. $I_L = 126.72$, $I_P = 126.85$, $I_F = 126.78$

4. $I_L = 141.13$, $I_P = 140.15$, $I_F = 140.64$ **5.** $I_F = 142.57$ **6.** $I_P = 115.2$, $I_F = 115.14$

Exercise 1.4

1. Index number by family budget method = 135.64 and total expenditure has increased by 35.64 %. Average monthly disposable income = ₹ 20,346.

2. Index number $I = 128.53$ and rise in total expenditure is 28.53 %.

3. Index number $I = 132.51$ and rise in total expenditure is 32.51 %.

4. Index number $I = 213.20$ and rise in total expenditure is 113.20 %.

5. Index number by family budget method = 129.64 and by total expenditure method $I = 129.64$ Thus, both index numbers are same.

Exercise 1

Section A

1. (c) **2.** (a) **3.** (d) **4.** (c) **5.** (d)

6. (d) **7.** (c) **8.** (c) **9.** (c) **10.** (c)

11. (a) **12.** (c)

Section B

12. The statement is false. Price index number of oil is 500.

Section C

7. Real wage ₹ 16,392.85 and loss to worker ₹ 1642.85 (Decrease in purchasing power)

8. Real wages ₹ 29166.67, 26666.67, 32307.69, 31250

9. Rate of inflation for year 2015 : 2.03 %

10. 449.55

11. Average monthly disposable income = ₹ 30,000

12. Index number of income = 125 **13.** Index number of production = 280 **14.** $I_p = 222.5$

Section D

7. 161.87

8. Fixed base index numbers = 100, 111.11, 133.33, 144.44, 166.67, 222.22, 263.89

9. Chain base index numbers = 100, 104, 100.96, 102.86, 100.93, 116.51

10. Fixed base index numbers = 120, 108, 151.20, 189

11. Chain base index numbers = 100, 112.5, 106.67, 114.58, 109.09, 116.67

12. Index number = 226.6 **13.** $I_L = 166.67$, $I_p = 150$, $I_F = 158.12$ **14.** $I_p = 167.71$

Section E

1. General index number = 122.32 **2.** Index number by total expenditure method = 149.41

3. Index number by total expenditure method = 115.69

4. Fixed base index numbers = 100, 118.75, 125, 131.25, 140.63, 187.5, 203.13;

Index numbers using average price = 91.43, 108.57, 114.29, 120, 128.57, 171.43, 185.71

5. Index number of industrial production I = 379.19

6. Index number I = 126.79 and rise in price is 26.79 %.

7. Real wages = 12,500, 10,000, 9268.29, 9090.91, 9361.7, 9615.38 Purchasing power of money = ₹ 0.38

Section F

1. $I_L = 113.65$, $I_P = 113.94$, $I_F = 113.79$ and rise in price is 13.79 %.

2. $I_P = 191.53$, $I_F = 211.52$

3. $I_F = 84.84$

4. $I_L = 109.52$, $I_P = 110.29$, $I_F = 109.90$

5. Index number by family budget method = 118.58 and index number by total expenditure method = 118.58. Thus, both index numbers are same.

6. Index number for year 2014 $I_1 = 239.41$ and index number for year 2015 $I_2 = 253.44$. The rise in cost of living in the current year is 14.03 %. The percentage rise in the price index number is 5.86 % and rise in wage is 5 %. Hence, wage rise is 0.86 % less.

7. Index number $I = 231.44$ Income should be ₹ 13,886.40 to maintain earlier standard of living.

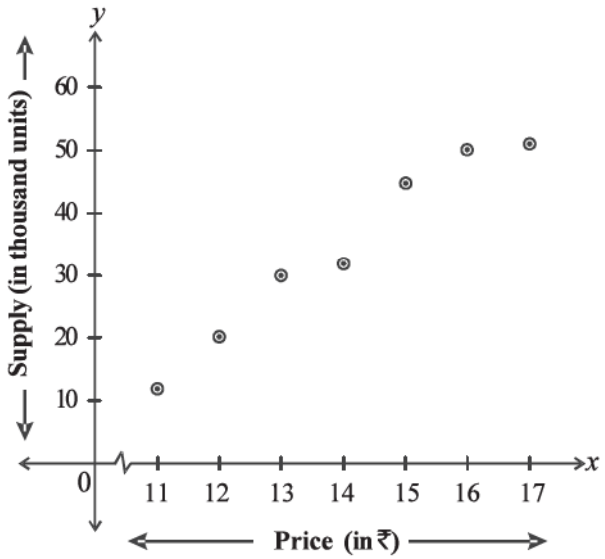
8. Index number of industrial production $= 100.10$, which indicates a rise of 0.10 % with respect to the base year.

9. Index number $I = 128.75$.

10. Cost of living index number $= 196.35$ and the rise is $(196.35 - 100) = 96.35$ % as compared to the base year.

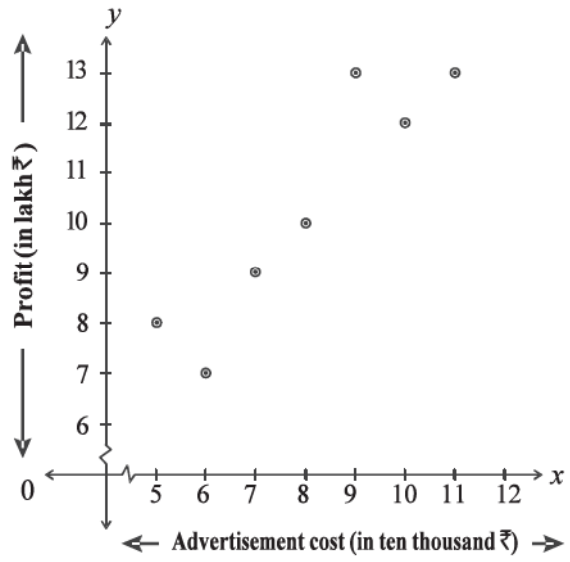
Exercise 2.1

1.



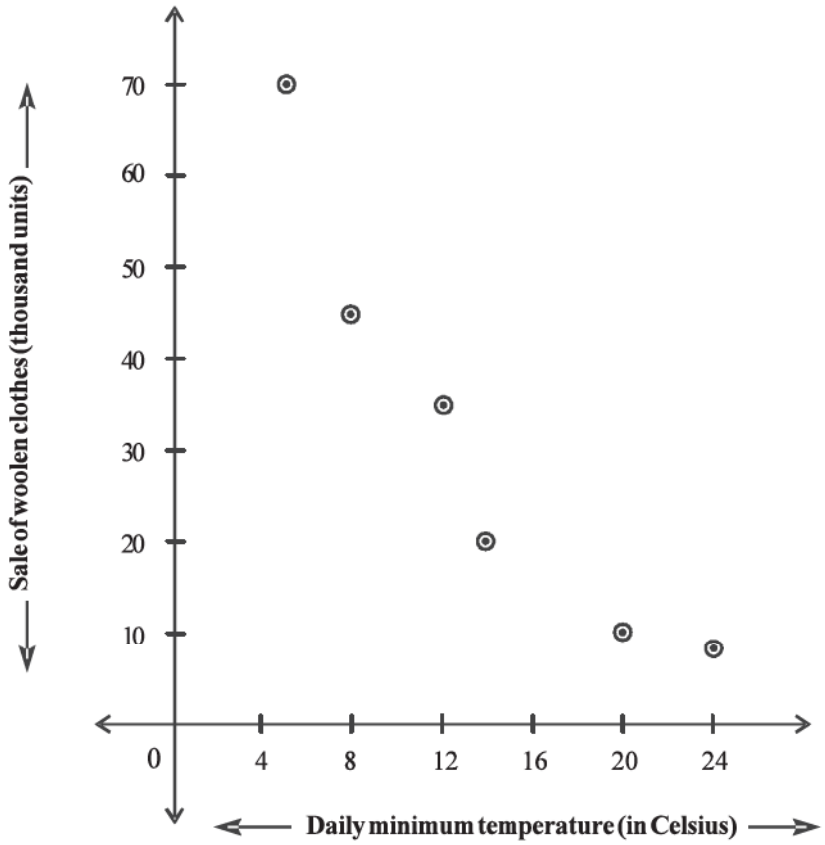
There is partial positive corelation between price and supply

2.



There is partial positive correlation between advertisement cost and profit

3.



There is partial negative correlation between daily minimum temperature and sale of woolen clothes

Exercise 2.2

1. $r = 0.81$ **2.** $r = -0.90$ **3.** $r = 0.90$ **4.** $r = 0.24$ **5.** $r = 0.82$

6. $r = -0.96$ **7.** $r = 0.67$ **8.** $r = -0.92$ **9.** $r = 0.99$ **10.** $r = 0.80$

11. $r = 0.84$ **12.** $r = 0.5$ **13.** $r = 0.8$ **14.** (1) $r = 0.94$ (2) $r = 0.96$

15. $r = -0.55$

Exercise 2.3

1. $r = 0.49$ **2.** $r = 0.78$ **3.** $r = 0.7$ **4.** $r = 0.82$ **5.** $r = 0.91$

6. $r = 0.90$ **7.** $r = -0.30$ **8.** Corrected $\Sigma d^2 = 122.5, r = 0.26$

Exercise 2

Section A

1. (c) **2.** (a) **3.** (d) **4.** (b) **5.** (c)

6. (d) **7.** (b) **8.** (b) **9.** (b) **10.** (c)

11. (a) **12.** (b) **13.** (c) **14.** (c) **15.** (a)

16. (b) **17.** (b) **18.** (a)

Section B

3. Positive **4.** Positive **5.** Negative **6.** Negative **7.** Nonsense correlation

8. r remain unchanged due to change of origion, so $r = 0.4$ **10.** $r = 0$ **11.** Negative

Section C

11. $r = 0.67$ **12.** $r = -0.54$ **13.** $r = 0.27$

Section D

10. $r = 0.75$ **11.** $r = 0$ **12.** $r = -0.5$ **13.** $r = 0.2$

Section E

1. $r = -0.81$ **2.** $r = 0.43$ **3.** $r = 0.79$ **4.** $r = 0.77$ **5.** $r = 0.54$

6. $r = 0.13$

Section F

1. $r = 0.99$ **2.** $r = -0.96$ **3.** $r = 0.88$ **4.** $r = 0.81$ **5.** $r = 0.38$

6. $r = 0.79$ **7.** $r = 0$ **8.** $r = 0.6$ **9.** $r = 0.3$ **10.** $r = 0.79$

11. Corrected $\Sigma d^2 = 78$; $r = 0.53$ **12.** $r = 0.73$

●

Exercise 3.1

1. $\hat{y} = 31.44 - 1.34\,x$ and for price x = 20 ₹, estimate of demand $\hat{y} = 4.64$ (hundred units)

2. $\hat{y} = 3.35 + 1.93\,x$ and for usage time of car x = 5 year, Estimate of annual maintenance cost $\hat{y} = 13$ (thousand ₹)

$\therefore$ Error $e = y - \hat{y} = 13 - 13 = 0$ (Here for x = 5, the observed value of y given in the table is 13)

3. $\hat{y} = 64.27 + 0.83\,x$ and for average rain x = 35 cm, estimate of yield of crop $\hat{y} = 93.32$ (ton)

4. $\hat{y} = 69.7 + 1.13\,x$ and for experience of worker, x = 7 year, estimate of performance index $\hat{y} = 77.61$

Exercise 3.2

1. $\hat{y} = 54.84 + 2.52\,x$ and for 300 kg usage of fertilizer [$\therefore$ x = 30 (ten kg.)], estimate of crop of cotton $\hat{y} = 130.44$ (Quintal per Hectare)

2. $\hat{y} = 52.84 + 0.68\,x$ and for a father's height x = 170 cm, estimate of height of the son $\hat{y} = 168.44$ cm

3. $\hat{y} = 20.72 - 0.71\,x$ and for altitude x = 7 thousand feat, estimate of effective Oxygen $\hat{y} = 15.75$ %

4. $\hat{y} = -3495.7 + 327.73\,x$ and for carpet area x = 110 sq. meter estimated monthly rent $\hat{y} = 32554.6$ ₹

5. $\hat{y} = 0.53 + 0.02\,x$ and for x = 80 customers, estimated sales $\hat{y} = 2.13$ (thousand ₹)

6. $\hat{y} = 7.6 + 0.29\,x$; x = Profit (lakh ₹) and y = Administrative cost (lakh ₹)

7. $\hat{y} = 53.72 + 1.54\,x$ and for rainfall x = 60 cm, estimate of yield of corn $\hat{y} = 146.12$ Quintal

8. $\hat{y} = 8.74 + 1.02\,x$ and for price x = 16 ₹, estimated supply $\hat{y} = 25.06$ (hundred units)

9. $\hat{y} = -4.8 + 0.15\,x$ and for maximum daily temperature x = 42 celcius, estimate of sale of icecream $\hat{y} = 1.5$ (lakh ₹)

Exercise 3

Section A

1. (b) **2.** (a) **3.** (c) **4.** (d) **5.** (a)

6. (a) **7.** (c) **8.** (c) **9.** (d) **10.** (b)

11. (c) **12.** (c) **13.** (c) **14.** (c) **15.** (b)

Section B

8. Error = 0

9. Both variables are multiplied by 2 there fore $c_x = \frac{1}{2}$ and $c_y = \frac{1}{2}$. $\therefore$ Regression coefficient will not change

10. $b_{yx} = 0.5 \times \frac{4}{2} = 1$ **11.** $\hat{y} = 50$ **12.** $r = 1$ **13.** $r = -1$

Section C

2. Error $e = 1$ **3.** $a = 2$ and $\hat{y} = 2 + 0.6\,x$

4. $b_{yx} = 5$ So, it can be said that because of increase of 1 unit in x, there is appropriate 5 units of increase in y.

5. $s_y = 3$ **6.** $R^2 = 1$ **7.** $s_x = 5$ **8.** 5 Units

9. $b_{yx} = 1.2$ and $a = 13$ **10.** $b_{vu} = b_{yx} \times \frac{c_x}{c_y} = 0.75 \times \frac{\frac{1}{6}}{\frac{1}{2}} = 0.25$

Section D

8. $\hat{y} = 4 + 0.75\,x$ **9.** $\hat{y} = -10 + 2\,x$

10. $R^2 = 0.81$; 81 % variation of the total variation in y, can be explained by the regression model.

11. $b_{yx} = 2.52$ so it can be said that because of increase of 1 unit in x, there is approximate 2.52 units of increase in y.

12. (i) $b_{vu} = 0.8$ (ii) $b_{vu} = 1.6$ (iii) $b_{vu} = 0.08$ **13.** $\hat{y} = 12 + 0.88\,x$

Section E

1. $\hat{y} = 2 + 0.75\,x$ **2.** $\hat{y} = 38.8 + 0.67\,x$ **3.** $\hat{y} = 58 + 3.2\,x$

4. $\hat{y} = 764.8 + 11.4\,x$ and for $x = 20$ cm, estimate of yield of crop is $\hat{y} = 992.8$ kg.

5. $\hat{y} = 18 + 0.8\,x$ and for $x =$ ₹ 45 lakh, estimate of market price is $\hat{y} = 54$ (lakh ₹)

Section F

1. $\hat{y} = 73.29 - 1.59\,x$ and for price x = 40 ₹, estimate of demand $\hat{y} = 9.69$ (hundred units)

2. $\hat{y} = 73.43 + 0.9\,x$ and for Experience x = 17 years, the estimate of performance rating $\hat{y} = 88.73$

3. $\hat{y} = 34.8 + 0.74\,x$ and for daily income x = 500 ₹, estimate of expenditure $\hat{y} = 404.8$ ₹

4. $\hat{y} = 3.73 + 0.13\,x$ and for advertisement cost $x = 50$ (ten thousand ₹), estimate of sales $\hat{y} = 10.23$ crore ₹

5. $\hat{y} = -122.94 + 91.67\,x$ and $R^2 = 0.97$ $\therefore$ Regression model is reliable.

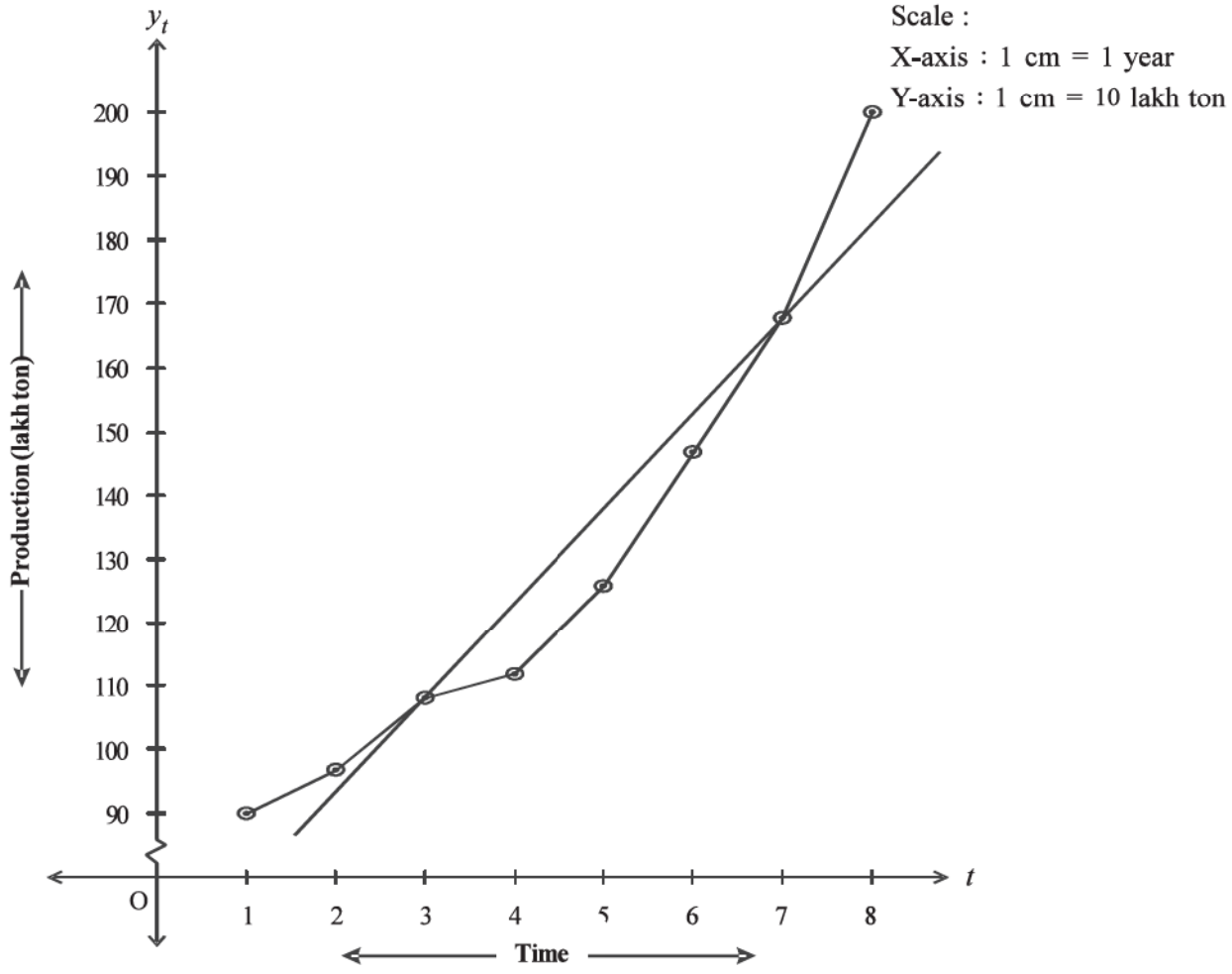
6. $\hat{y} = -10 + 1.6\,x$ and for x = 30 $\hat{y} = 38$

7. $\hat{y} = -0.44 + 0.7\,x$ and for x = 5, $\hat{y} = 3.06$

●

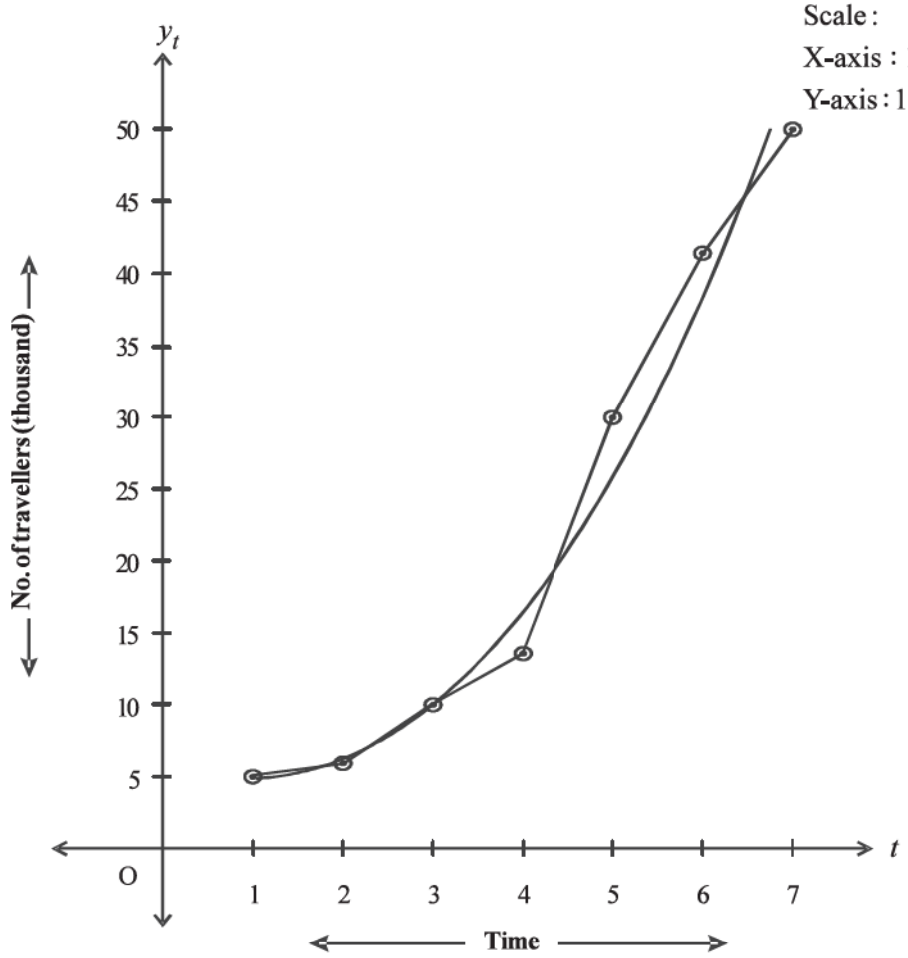
Exercise 4.1

1.



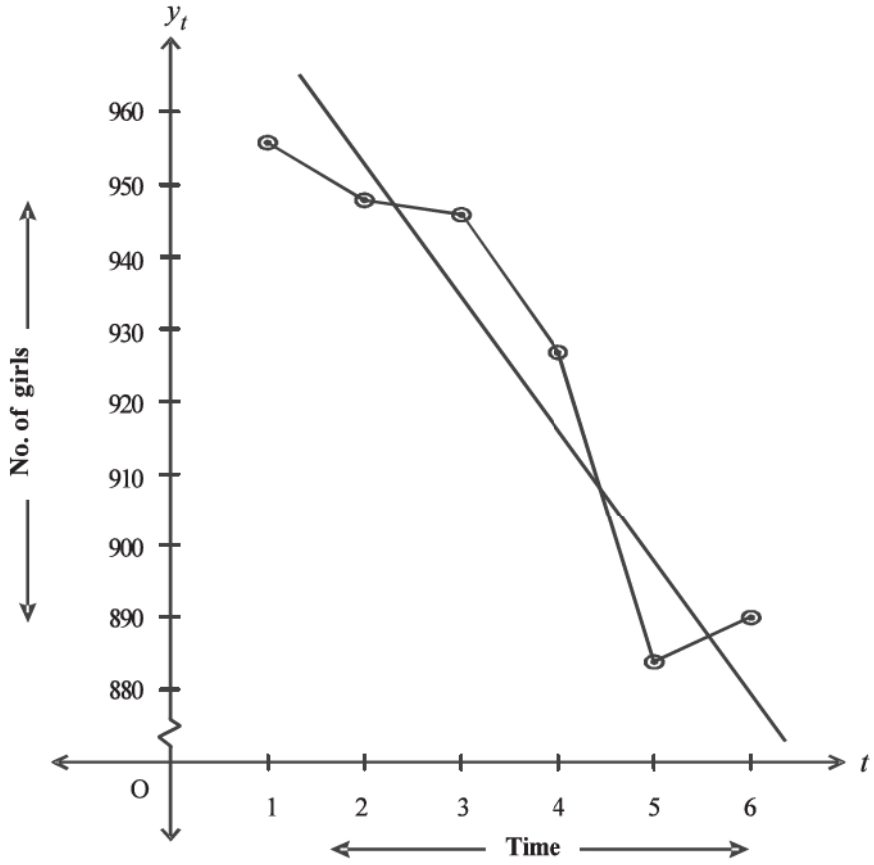
2.

Scale :
X-axis : 1 cm = 1 year
Y-axis : 1 cm = 5 thousand travellers

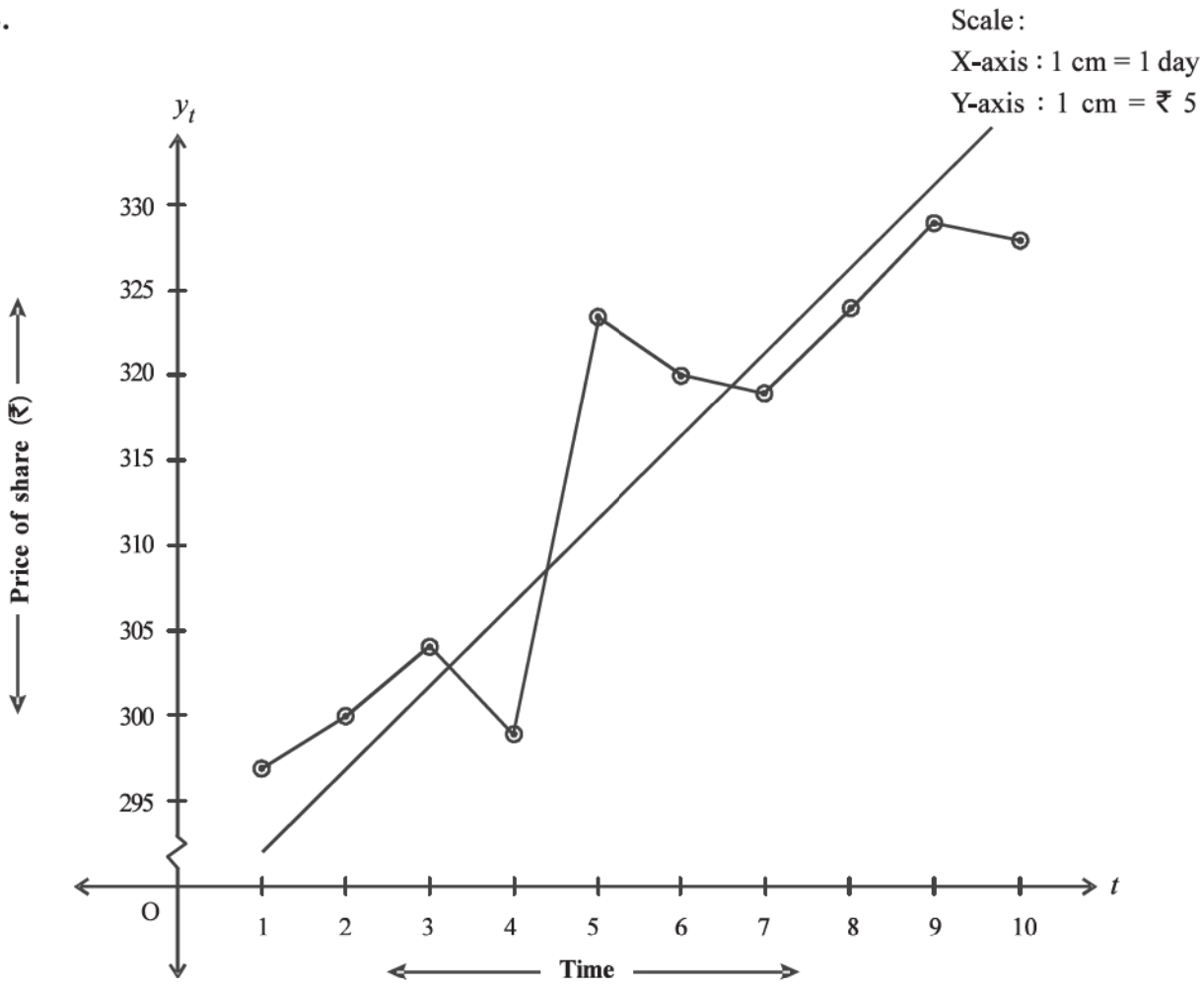


3.

Scale :
X-axis : 1 cm = 10 year
Y-axis : 1 cm = 10 girls



4.



Exercise 4.2

1. $\hat{y} = 7.41 - 0.07\,t$, for year 2017 $\hat{y} = 6.78$

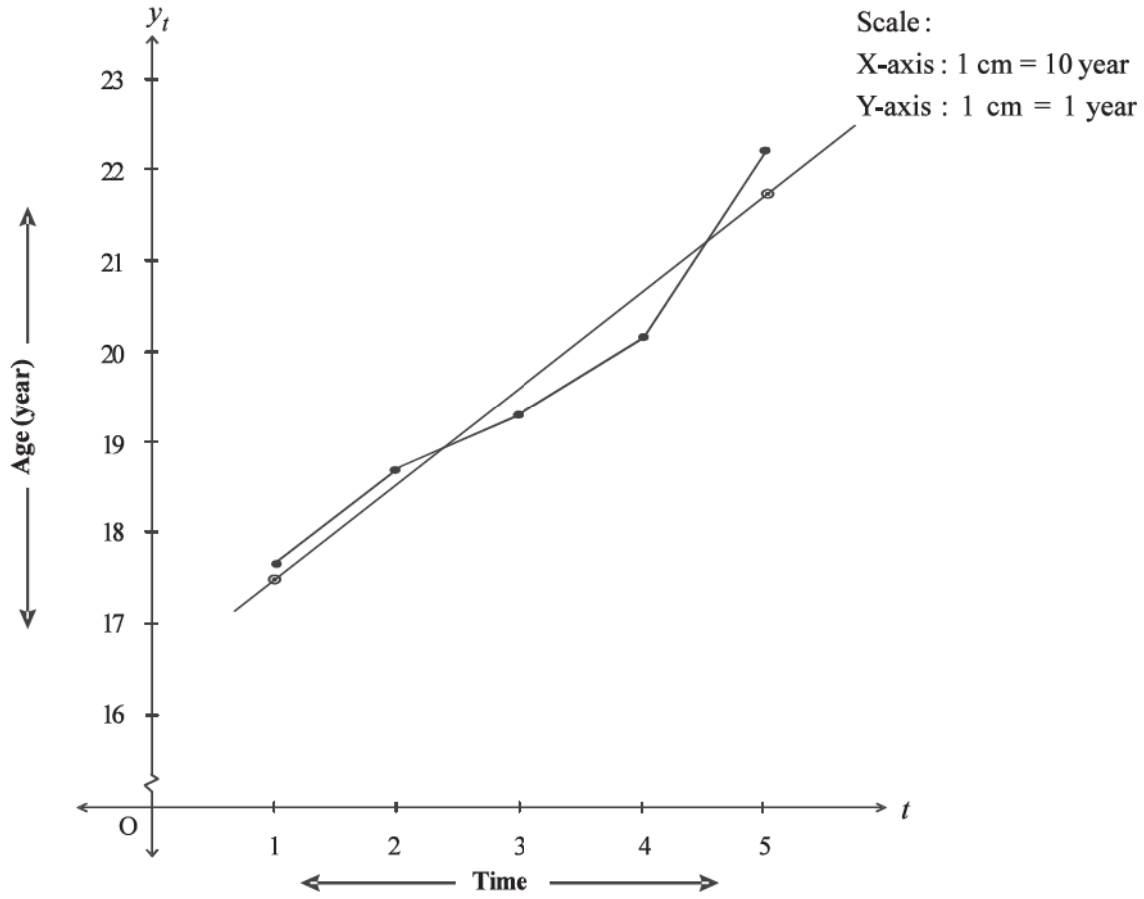
2. $\hat{y} = 447.2 + 69.4\,t$, for year 2015-16 $\hat{y} = 1071.8$

3. $\hat{y} = 57.12 + 9.06\,t$, $\hat{y} = 120.54$ thousand for year 2016

$\hat{y} = 129.6$ thousand for year 2017

Year	2010	2011	2012	2013	2014	2015
Estimated values of trend (thousand vehicles)	66.18	75.24	84.3	93.36	102.42	111.48

4. $\hat{y} = 16.47 + 1.05\,t$, $\hat{y} = 22.77$ years for year 2021



Exercise 4.3

1.

Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Three yearly moving average	–	5	6	7	8	9	10	11	12	–

2.

Month	Jan.	Feb.	March	April	May	June	July	August	Sept.	Oct.	Nov.	Dec.
Four yearly moving average	–	–	274.88	280.38	273.88	263	265.38	269.25	272.63	275.88	–	–

3.

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015
Four yearly moving average	–	–	16.8	18.6	21.2	22.6	23.2	–	–

4.

Year	2013				2014				2015			
Quarter	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4	Q_1	Q_2	Q_3	Q_4
Four Quarterly moving average	–	–	124.38	134	143	150.88	153.25	148	141.63	136.88	–	–

Exercise 4

Section A

1. (c) **2.** (a) **3.** (b) **4.** (b) **5.** (c)

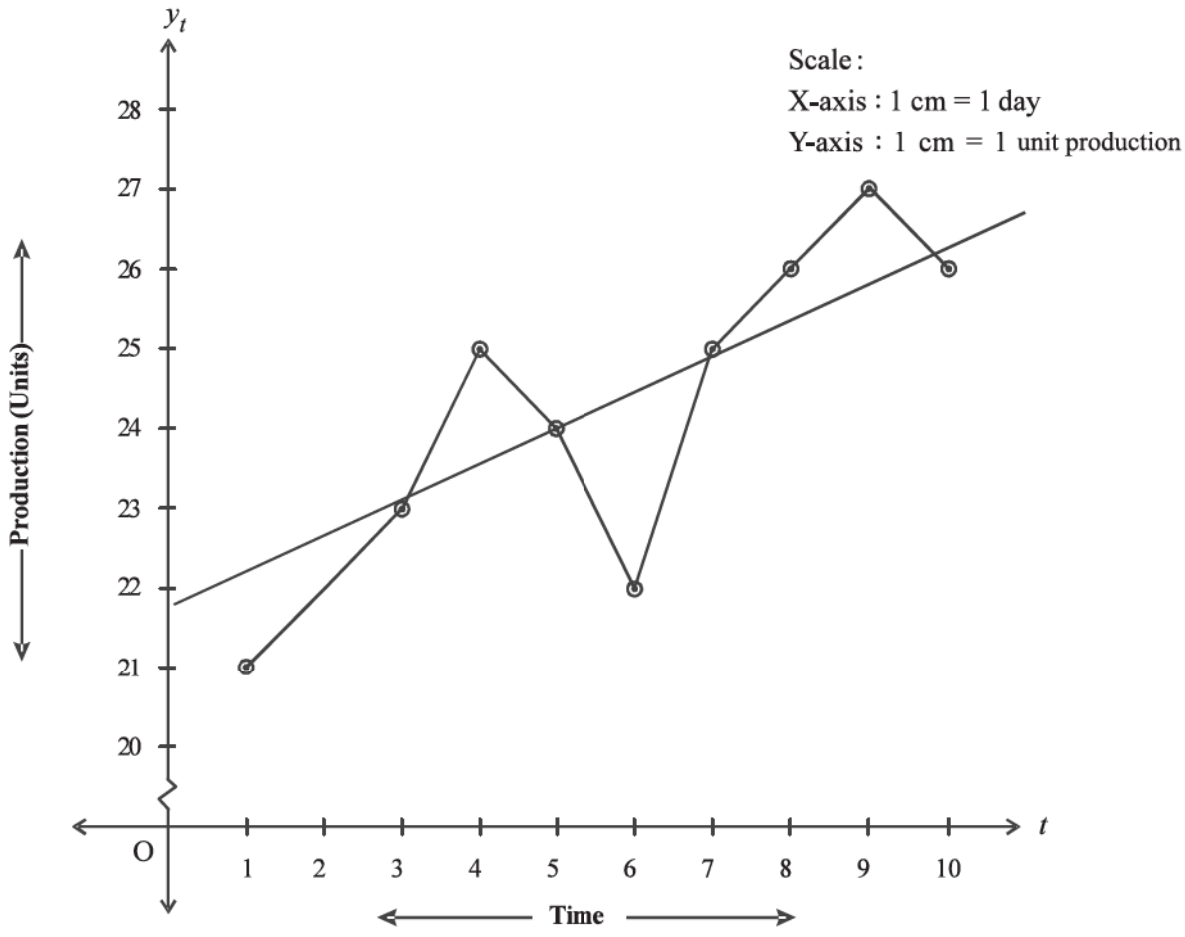
6. (a) **7.** (b) **8.** (c) **9.** (c) **10.** (d)

Section B

10. $\hat{y} = 14.7$ for eighth week

Section D

9.



10. $\hat{y} = 38 + 11.8\,t$

11.

Month	January	February	March	April	May	June	July
Three monthly moving average	–	16.33	17.33	19.67	21.33	22	–

Section E

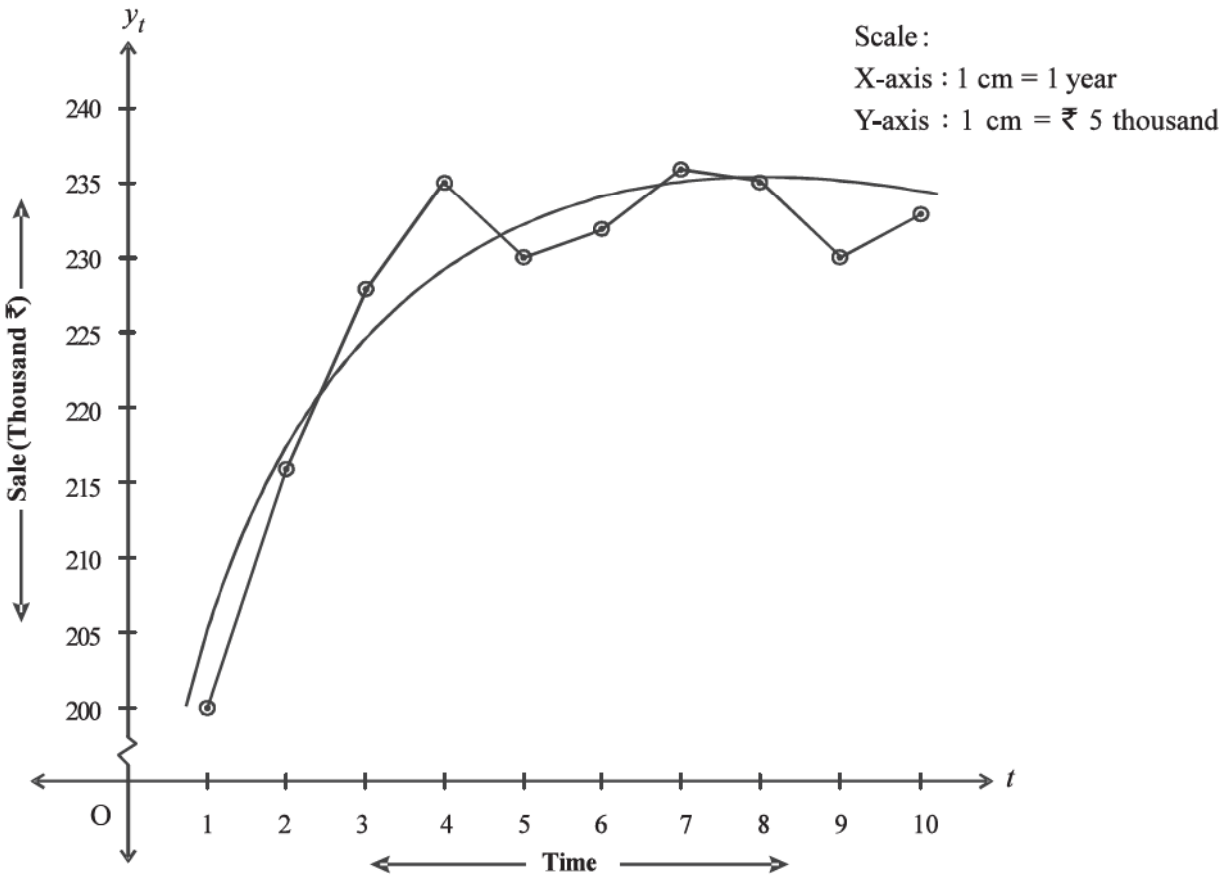
1. $\hat{y} = 23.19 + 0.09\,t$, $\hat{y} =$ 23.91 crore for year 2017

2. $\hat{y} = 49.1 - 2.1\,t$, $\hat{y} =$ 36.5 thousand for year 2016

3.

Month	April 2015	May	June	July	August	Sept.	Oct.	Nov.	Dec.	Jan. 2016
Three monthly moving average	–	71.33	68.67	66.67	65	63.33	63.33	65	66	–

4.



5.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Five yearly moving average	–	–	118.4	119.2	119	118.8	118	119.4	120.6	–	–

Section F

1. $\hat{y} = 30.26 + 1.19\,t$, $\hat{y} =$ 39.78 crore tons for year 2016-17

$\hat{y} =$ 40.97 crore tons for year 2017-18

2.

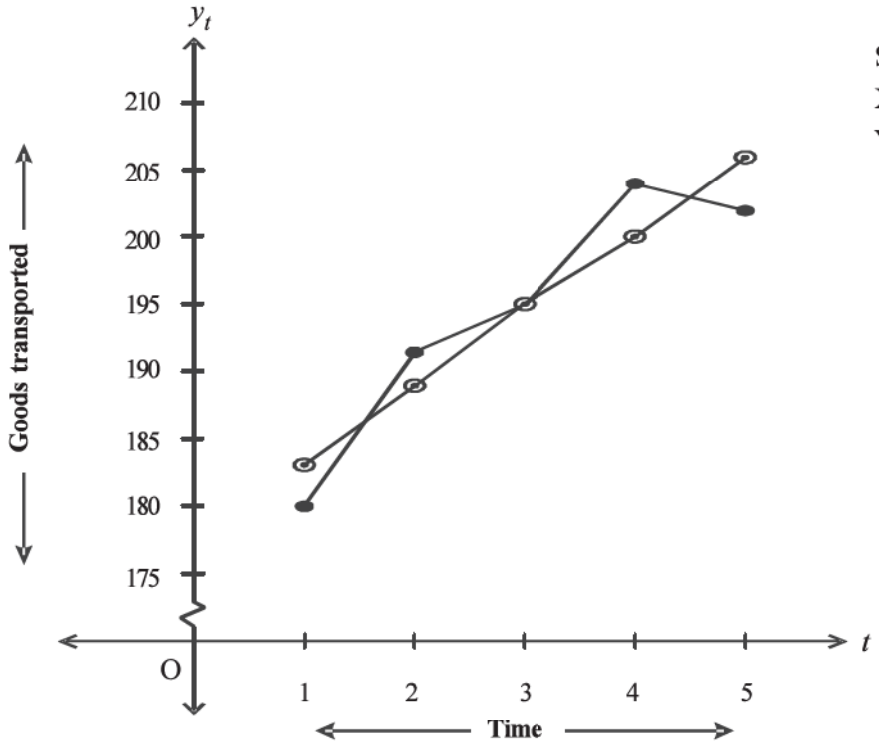
Year	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Four yearly moving average	–	–	358.25	378	395.63	406.63	411.38	415.88	–	–

3. $\hat{y} = 22.55 - 0.39\,t$, $\hat{y} =$ 19.43 for year 2016 $\hat{y} =$ 19.04 for year 2017

4. $\hat{y} = 177.8 + 5.6\,t$, $\hat{y} =$ 211.4 ton for year 2016

Year	2011	2012	2013	2014	2015
Estimated value of trend (tons)	183.4	189	194.6	200.2	205.8

5.



Scale :
X-axis : 1 cm = 1 year
Y-axis : 1 cm = 5 ton

Month	March				April				May			
Week	1	2	3	4	1	2	3	4	1	2	3	4
Four weekly moving average	–	–	38.44	38.7	38.97	39.62	41.29	43.05	44.4	45.72	–	–

● ● ●