BBA-201

BUSINESS STATISTICS

DIRECTORATE OF DISTANCE EDUCATION

SWAMI VIVEKANAND
SUBHARTI UNIVERSITY
Meerut (National Capital Region Delhi)

Contents

Business Statistics

Course Code: BBA- 201N		
Course Credit: 5	Lecture: 04	Tutorial: 1
Course Type:	Core Course	
Lectures delivered:	40 L + 10 T	

End Semester Examination System

Maximum Marks Allotted	Minimum Pass Marks	Time Allowed
70	28	3 Hours

Continuous Comprehensive Assessment (CCA) Pattern

Tests	Assignment/ Tutorial/ Presentation/class test	Attendance	Total
15	5	10	30

Course Objective: The objective of course is to provide basic knowledge of quantitative methods and their commercial application for decision making in business.

UNIT	Content	Hours
I	INTRODUCTION TO STATISTICS: Background and Basic concepts: Introduction, Definition of Statistics, Functions, Scope, Limitations, Classification and Tabulation of Data.	8
II	MEASURES OF CENTRAL TENDENCY: Introduction, Types of averages, Arithmetic Mean (Simple and Weighted), Median, Mode, Graphic location of Median and Mode through Ogive Curves and Histogram.	12
III	MEASURES OF DISPERSION AND SKEWNESS: Meaning,Calculation of Absolute and Relative measures of dispersion, Range ,Mean Deviation , Standard Deviation Quartile Deviation , and Coefficient of Variation. **Measures of Skewness:** Meaning of Skewness, Symmetrical &Skewed Distributions, Absolute and Relative Measures of Skewness, Karl Pearson's Coefficient of Skewness.	14
IV	CORRELATION AND REGRESSION ANALYSIS: Correlation, Meaning &Definition,Uses,Types, Probable error, Karl Pearson's & Spearman's Rank Correlation (Excluding Bivariate and Multiple correlation). Regression - Meaning and Definition, Regression Equations, Problems	12
V	INDEX NUMBERS: Meaning &Definition , Uses , Classification , Construction of Index Numbers , Methods of constructing Index Numbers , Simple Aggregate Method , Simple Average of Price Relative Method , Weighted Index numbers , Fisher's Ideal Index (including Time and Factor Reversal tests) , Consumer Price Index, Problems	14

Course Outcomes: After studyingthis course the student should be able to

1. Understand the basic concepts of the Statistics and use the idea about mean median, mode in real life.
2. Estimate the level of correlation, regression and the relationship for the given bivariate data and its various applications.
3. Explain the basic concepts of index number & time series and its uses in real life applications.

Text Books:

1. S P Gupta: Statistical Methods, Sultan Chand, Delhi
2. KK Sharma: Business Statistics, Krishna Educational Publishers

Reference Books:

1. C.R.Reddy : Quantitative Techniques for Management Decisions, HPH.
2. Dr. B N Gupta: Statistics (SahitytaBhavan), Agra

NOTES

1. INTRODUCTION OF STATISTICS

STRUCTURE

1.1. INTRODUCTION

In ancient times, the use of statistics was very much limited and is just confined to the collection of data regarding manpower, agricultural land and its production, taxable property of the people etc. But as the time passed, the utility of this subject increased manifold. Many researches were conducted in this field and with the result of this it started growing as a separate subject of study. Many experts in the field of mathematics and economics contributed toward the development of this subject. The word 'Statistics' which was once used in the sense of just collection of data is now considered as a full fledged subject. The knowledge of this subject is used for taking decisions in the midst of uncertainty.

1.2. MEANING OF STATISTICS

The word 'Statistics' has been defined differently by different statisticians from time to time. This word is understood in two different forms, namely 'numerical data' and as a 'science'. When statistics is defined as numerical data, it is said to be defined in the plural sense. And when it is defined as a science, it is said to have been defined in the singular sense. In 1935, *Dr. W.F. Willifox* listed over a hundred definitions of statistics and the list was not complete. The best considered definitions of the word 'Statistics' are given by *Horace Secrist* and *Croxton and Cowden*.

1.3. DEFINITION OF STATISTICS IN PLURAL SENSE (AS DATA)

NOTES

Horace Secrist defined statistics as aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.

1. Aggregate of facts. Statistics refers to set of numerical data. It does not recognise individual items. Suppose, the height of Mr. X is 70 inches. This statement does not constitute statistical data. If we are given the profit figures of a particular firm for the last twenty years, that would constitute statistical data.

2. Multiplicity of causes. This is an important characteristic of statistics. Economic and business phenomenon are very complex. These are influenced by a large number of forces. For example, the statistics of production of a crop is in general affected by a number of factors like soil conditions, quantity of fertilizers used, quantity of rainfall, method of cultivation, quality of land, quality of seed, state agriculture policy etc. Whatsoever may be the nature of statistical variable, its value is likely to be affected by causes like human psychology, state policies, market conditions, social relations etc.

3. Numerically expressed. According to professor *Horace Secrist*, "the statistical approach to a subject is numerical. Things, attributes and conditions are counted, totalled, divided and sub-divided and analysed". Data regarding qualitative variables like beauty, intelligence, honesty, poverty cannot be expressed directly.

4. Reasonable standard of accuracy. In statistical data, we do not require hundred per cent accuracy in measurement, which is neither possible nor indispensable. Suppose, we are considering the height of students of a class. It will be quite sufficient to measure upto 1/10 of a centimetre. Most of the times the data is expressed in round figures. A statistician would be contented to take 0.3 as the value of 1/3, but on the other hand, its exact value is 0.33333......

5. Systematic manner. It means that the data collected should be in a systematic manner. Suppose, we have data regarding the income and age of population of a certain city. It would be advisable to arrange the data in accordance with increasing values of the variables under consideration.

6. Pre-determined purpose. The data must be collected with a pre-determined purpose in view. An enumerator cannot collect data about any business and economic variable unless he is told about the purpose. Suppose the data is collected for preparing index number, then it must be very much clear to the enumerator, whether whole sale price-index or consumer price-index is to be calculated.

7. Placed in relation to each other. It means that all data, which is collected, should be comparable. The data may be comparable w.r.t. time, place etc. The data regarding the total strength of a particular college from 1975-86 is a statistical data. On the other hand, we may also collect the data regarding the strength of different colleges in Haryana in the year 1986.

From this definition, we see that all numerical data are not statistical data. On the other hand, all statistical data are numerical data.

1.4. DEFINITION OF STATISTICS IN THE SINGULAR SENSE (AS A SUBJECT)

NOTES

In the singular sense, the word 'statistics' is defined as a subject. This subject mainly deals with the analysis of classified data. The tools of mathematics are used extensively in this subject. *Croxton* and *Cowden* defined the subject statistics as the collection, presentation, analysis and interpretation of numerical data. In statistics, we learn the methods of classifying data, graphical presentation of data, averages of data, dispersion of data. We also study skewness and kurtosis of distribution. We also study correlation, regression; index numbers, analysis of time series, probability, theoretical distributions, tests of significance. In brief, we can say that the subject statistics is concerned with the scientific methods of collecting, summarizing, presenting and analysing data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis. In nutshell, we can remark that statistics studies statistics. The subject matter of the subject statistics is divided in two parts namely, *descriptive statistics* and *inferential statistics.*

1.5. DESCRIPTIVE STATISTICS

The part of the subject statistics which deals with the analysis of a given group without drawing conclusions about a larger group is called **descriptive statistics.** Descriptive statistics includes, collection of data, presentation of data, measures of averages, dispersion, skewness, kurtosis, correlation, regression, index numbers, components of time series. In our present course, we shall be mainly dealing with descriptive statistics. Descriptive statistics is also known as **deductive statistics.**

1.6. INFERENTIAL STATISTICS

The part of the subject statistics which deals with the analysis of a given group and drawing conclusions about a larger group is called **inferential statistics.** For studying data regarding a group of individuals or objects, such as heights, weights, income, expenditure of persons in a locality or number of defective and non-defective articles produced in a factory, it is generally impracticable to collect and study data regarding the entire group. Instead of examining the entire group, we concentrate on a small part of the group called a **sample.** If this sample happen to be a true representative of the entire group, called **population,** important conclusions can be drawn from the analysis of the sample. The conditions under which the conclusions for samples can be considered valid for the corresponding populations are studied in inferential statistics. Since such conclusions cannot be absolutely certain, the language of probability is often used in stating conclusions. Theoretical distributions are also needed in inferential statistics. In the present course, we shall be studying probability and theoretical distributions. Binomial, Poisson and Normal. Inferential statistics is also known as **inductive statistics.**

1.7. SCOPE

NOTES

The importance of statistics cannot be underestimated. It is considered to be both a science and an art. According to *Tippett*, "Statistics is both a science and an art. It is a science in that its methods are basically systematic and have general application and an art in that their successful application depends to a considerable degree on the skill and special experience of the statistician and on his knowledge of the field of application *e.g.*, economics". Statistical methods help simplifying complexities. Statistical tools like diagrams and graphs presents data in simple and attractive form. That is why, diagrams and graphs are used for presenting data in exhibitions etc. On the extreme end, statistical methods are used to prepare control charts and for analysis of variance etc. Statistical methods are helping in enlarging human experience and is becoming indispensable. Statistics is related to almost every other discipline of knowledge like planning, mathematics, economics, physical science, psychology etc.

1. Statistics and Planning. Statistical methods are powerful tools in the hands of the Government. The modern era is the era of Planning. In almost every country of the world, governments prepare their future plans well before time. This is very essential for the economic development of the country. The success of a plan depends upon the proper use of suitable statistical methods.

2. Statistics and Mathematics. Statistics is very closely related with mathematics. Statistical formulae are developed on the sound knowledge of mathematics. Contributors to statistics were primarily talented mathematicians. *Connor* defined statistics as the branch of applied Mathematics which specialises in data. Advancement in this field is impossible without the use of mathematics. Role of mathematics in the development of statistics is ever increasing.

3. Statistics and Economics. Methods of statistics are very much important in the formulation of economic policies. Data regarding the consumption pattern of people helps in knowing the standard of their living and their taxable capacity. Government adjust the dearness allowance of its employees on the basis of cost of living index, which is prepared by using the statistics regarding the consumption pattern of the people concerned. Statistical methods are used very widely in solving economic problems relating to wages, prices, consumption function, analysis of time series, etc. The ever increasing use of statistics in the field of economics has led to the development of a new subject 'Econometrics'.

4. Statistics and Business. Statistics is quite indispensible in business. In big business concerns, it is very much difficult for the owner to act as salesmen, accountant, store-manager etc. It is quite difficult for him to maintain direct contact with his customers. The success of business depends upon accuracy of his statistical forecasting. Business houses avails the services of trained and skilled statisticians. Without making use of statistical tools, it is impossible to take decisions about the effect of different forces acting on the 'profit' variable.

5. Statistics and Physical Sciences. The methods of statistics are very much used in Physical sciences like Physics, Biology, Astronomy, Medical Sciences, Geology etc. *Professor Karl Pearson* studied that the 'theory of heredity' is based on Statistics. In Astronomy, the 'normal laws of errors' were developed by using the 'Principle of least squares'. In medical science, χ^2-test and other tests of significance are used widely to see whether a particular medicine is useful in curing a certain disease or not.

1.8. LIMITATIONS

NOTES

Statistics is considered to be a science as well as an art, which is used as an instrument of research in almost every sphere of our activities. There are limitations of statistics also. Care must be taken of these limitations while using statistical methods. *Newsholme* has well said, "It must be regarded as an instrument of research of great value but having several limitations which are not possible to overcome and as such they need our careful attention". Some of the limitations of statistics are as follows :

1. Statistics suits to the study of quantitative data only. Statistics deals with the study of quantitative data only. By using the methods of statistics, the problems regarding production, income, price, wage, height, weight etc. can be studied. Such characteristics are quantitative in nature. The characteristics like honesty, goodwill, duty, character, beauty, intelligence, efficiency, integrity etc. are not capable of quantitative measurement and hence cannot be directly dealt with statistical methods. These characteristics are qualitative in nature. In such type of characteristics, only comparison is possible. The statistical methods may be tried in studying qualitative characteristics only if they are expressed quantitatively. For example, the efficiency of workers in a hand made paper factory, may be studied by considering the number of paper sheets prepared daily by each worker. The use of statistical methods is limited to quantitative characteristics and those qualitative characteristics which are capable of being expressed numerically.

2. Statistical results are not exact. The task of statistical analysis is performed under certain conditions. It is not always possible, rather not advisable, to consider the entire population during statistical investigations. The use of samples is called for in statistical investigations. And the results obtained by using samples may not be universally true for the entire population. Data collected for a statistical enquiry may not be hundred percent true. Statistical results are true on an average. If we comment that the students of a particular class are intelligent, it does not necessarily imply that each and every student of the class is intelligent. The probability of getting a head in a single trial of an unbiased coin is 1/2, but we may not get exactly one head in two trials of the coin. That is why, statistics is not considered an exact science like Physics, Mathematics etc.

3. Statistics deals with aggregates only. Statistics does not recognise individual items. Consider the statement, "The weight of Mr. X in the college is 70 kg". This statement does not constitute statistical data. Statistical methods are not going to investigate anything about this statement. Whereas, if the weights of all the students of the college are given. the statistical methods may be applied to analyse that data. According to *Tippett*, "Statistics is essentially totalitarian because it is not concerned with individual values, but only with classes". Statistics is used to study group characteristics of aggregates. If we are given the profit figure of a firm manufacturing a particular item, it does not help in commenting on the performance of the company. On the other hand, if we are given the profit figures of the firm for the ten or fifteen consecutive years, we can make use of statistical methods to comment on the performance of the firm.

4. Statistics is useful for experts only. Statistics is both a science and an art. It is systematic and find applications in studying problems in Economics, Business, Astronomy, Physics, Medicines etc. Statistical methods are sophisticated in nature. Everyone is not expected to possess the intelligence required to understand and to apply these methods to practical problems. This is the job of an expert, who is well-versed

NOTES

with statistical methods. *W.I. King* says, "Statistics is a most useful servant but only of great value to those who understand its proper use".

A skilled statistician can never advise the public to walk in the middle of the road just on the plea that the number of padestrians who died in road accidents during 1975-86, on a particular road, was less for those who walked in the middle than those who walked on the road side. Such statements can only be given by those who cannot be considered as experts in using statistical methods. If such a type of decision is to be taken, then we must also consider the number of persons, who walk on the middle and on the sides of the road. In fact, if some body, accepts the advice of such an expert and start walking in the middle of the road, he is not hoped to survive much longer. *Yule* and *Kendall* has rightly said that "statistical methods are most dangerous tools in the hands of inexperts".

The methods of statistics must be tried by experts only. The results derived by using statistical methods would very much depends upon the skill of the user. According to *King*, "Statistics are like clay of which one can make a god or devil as one pleases".

5. Statistics does not provide solutions to the problems. The statistical methods are used to explore the essentials of problems. It does not find use in inventing solutions to problems. For example, the methods of statistics may reveal the fact that the average result of a particular class in a college is deteriorating for the last ten years, *i.e.*, the trend of the result is downward, but statistics cannot provide solution to this problem. It cannot help in taking remedial steps to improve the result of that class. Statistics should be taken as a means and not as an end. The methods of statistics are used to study the various aspects of the data.

EXERCISE 1.1

1. Define fully the importance of statistics as an aid to business and commerce.
2. Explain clearly the three meanings of the word statistics contained in the statement given below :

   "You compute statistics from statistics by statistics."
3. Statistics are aggregate of facts, affected to marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other". (*Horace Secrist*)

   Elucidate the above definition.
4. "The science of statistics is a most useful servant but only of great value of those who understand its proper use". (*King*)

   Comment on the above statement and discuss the limitations of statistics.
5. Define statistics and discuss its scope and limitations.
6. How a basic knowledge of statistics is essential to becoming an efficient citizen in a modern democracy ?
7. "Statistics only furnish a tool, necessary though imperfect, which is dangerous in the hands of those who do not know its uses and deficiencies". (*Bowley*).

   Discuss.
8. In what sense do we say that a basic knowledge of statistics is essential to become an efficient businessman in a modern world.
9. "Statistics are numerical statements of facts, but all facts numerically stated are not statistics". Clarify this statement and point out briefly which numerical statements of facts are statistics.

10. "Statistics are not mere mass of figures". Elucidate.
11. Describe with the help of suitable illustrations the functions of statistics.
12. What are the limitations of statistics ?
13. Statistics is said to be both science and art. Why ?
14. Indicate the usefulness of Statistics in modern times.
15. Comment on the following statements :
    (i) "Statistics is the science of counting".
    (ii) "Statistics is the science of averages".
    (iii) "Statistics is the science of measurement of social organism regarded as a whole in all manifestations".
    (iv) "Statistics is the science of estimates and probabilities".
16. What is the importance and functions of statistics ?

1.9. SUMMARY

- *Horace Secrist* defined statistics as aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other.
- The part of the subject statistics which deals with the analysis of a given group without drawing conclusions about a larger group is called **descriptive statistics.** Descriptive statistics includes, collection of data, presentation of data, measures of averages, dispersion, skewness, kurtosis, correlation, regression, index numbers, components of time series.
- The part of the subject statistics which deals with the analysis of a given group and drawing conclusions about a larger group is called **inferential statistics.**
- Instead of examining the entire group, we concentrate on a small part of the group called a **sample.** If this sample happen to be a true representative of the entire group, called **population,** important conclusions can be drawn from the analysis of the sample.
- "Statistics is both a science and an art. It is a science in that its methods are basically systematic and have general application and an art in that their successful application depends to a considerable degree on the skill and special experience of the statistician and on his knowledge of the field of application *e.g.,* economics". Statistical methods help simplifying complexities.

NOTES

2. COLLECTION OF DATA

STRUCTURE

2.1. INTRODUCTION

The first step in a statistical investigation is the planning of the proposed investigation. After planning, the next step is the collection of data, keeping in view the object and scope of investigation. There are number of methods of collecting data. The mode of collection of data also depends upon the availability of resources. The collected data would be edited, presented, analysed and interpreted. If the job of data collection is not done sincerely and seriously, the results of the investigation is bound to be inaccurate and misleading. And so the resources used in performing the other steps would be wasted and the very purpose of the investigation would be defeated.

Types of Data

I. Primary Data II. Secondary Data

I. PRIMARY DATA

NOTES

2.2. DEFINITION OF PRIMARY DATA

Data is called **primary,** if it is originally collected in the process of investigation. Primary data are original in nature. Primary data are generally used in case of some special purpose investigation. The process of collecting primary data is time consuming. For example suppose we want to compare the average income of employees of two companies. This can be done by collecting the data regarding the incomes of employees of both companies. The data collected would be edited, presented and analysed by taking averages of both groups of data. On the basis of the averages, we would be able to declare as to the average income for which company is more. The data used in this investigation is primary, because the data regarding the incomes of employees was collected during the process of investigation.

2.3. METHODS OF COLLECTING PRIMARY DATA

(*i*) Direct personal investigation

(*ii*) Indirect oral investigation

(*iii*) Through local correspondents

(*iv*) Through questionnaires mailed to informants

(*v*) Through schedules filled in by enumerators.

Now we shall discuss the process of collecting primary data by these methods. We shall also discuss the suitability, merits and demerits regarding the above mentioned methods of collecting primary data.

2.4. DIRECT PERSONAL INVESTIGATION

In this method of collecting data, the investigator directly comes in contact with the informants to collect data. The investigator himself visits the different informants, covered in the scope of the investigation and collect data as per the need of the investigation. Suppose an investigator wants to use this method to collect data regarding the wages of the employees of a factory then he would have to contact each and every employee of the factory in order to collect the required data. In the context of this method of collecting primary data, *Professor C.A. Moser* has remarked, "In the strict sense, observation implies the use of the eyes rather than of the ears and the voice". The suitability of this method depends upon the personality of the investigator. The investigator is expected to be tactful, skilled, honest, well behaved and industrious. It is suitable when the area to be covered is small. This is also suitable when the data is to be kept secret.

Merits

(*a*) The degree of accuracy in data collected is very high.

(*b*) Because of the personal visit of the investigator, the response of informants is expected to be very encouraging.

NOTES

(c) The data collected is reliable.

(d) This method is flexible in the sense that the investigator can modify the nature of data to be collected in accordance with the prevailing circumstances.

(e) Consistency and homogeneity is present in the data collected.

(f) Data regarding delicate and sensitive questions can also be gathered by twisting the questions, as per the need.

Demerits

(a) This method is very expensive and time consuming.

(b) This method is not useful when the informants are spread over a wide area.

(c) This method is not useful in case the investigator is expected to be biased.

2.5. INDIRECT ORAL INVESTIGATION

In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the data about the informants is collected from some selected persons who are expected to be acquinted with the informants as well as the object of the investigation. A person giving data about the informants is called 'witness.'

The suitability of this method depends upon the personalities of the investigator and the witnesses. The investigator is expected to be tactful, skilled, honest and well behaved. At the same time a person giving information about the informants is expected to be unbiased and well acquainted with the object of the investigation. This method is suitable in the cases where the informants are not expected to give data frankly, when contacted directly. Suppose an investigator wants to collect data regarding the level of intelligence of the students in different classes of a college. The investigator may not be able to get the required data by contacting the students and asking them about their intelligence. The required information in this case can be obtained by contacting the class teachers or the Head of the Institution. This method of collecting data is also useful when the area to be covered is very large.

Merits

(a) This method is economical in respect of money, labour and time.

(b) This method can be easily used even if the area to be covered is widely spread.

(c) In this method, the data is collected from the third person and so the data is not effected by the bias on the part either the informants or the investigator.

(d) In this method, the investigator can use the views and knowledge of experts by contacting them during the process of collecting data.

Demerits

(a) The data is expected to be highly affected by the bias of the witnesses. They can misguide the investigator by distorting the data.

(b) The data collecting is not expected to be accurate due to the carelessness of the witnesses.

2.6. THROUGH LOCAL CORRESPONDENTS

In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the data about the informants is collected and sent to the investigator by the local correspondents, appointed by the investigator. Newspaper agencies collect data by using this method. They appoint their correspondents area wise. The correspondents themselves send the desired data to the offices of their respective newspaper. The suitability of this method depends upon the personality of the correspondents. He is expected to be unbiased, skilled and honest. To eliminate the bias of the correspondents, it is advisable to appoint more than one correspondent in each area.

Merits

(*a*) This method of collecting data is most economical in respect of money, labour and time.

(*b*) This method can be easily used even if the area to be covered is widely spread.

(*c*) This method is specially recommended for investigations, in which data is to be collected on regular basis.

Demerits

(*a*) The data collected is not expected to be very accurate.

(*b*) The data collected is expected to be affected by the bias of the correspondents.

(*c*) The data collected is not expected to be very reliable.

2.7. THROUGH QUESTIONNAIRES MAILED TO INFORMANTS

In this method of collecting data, the informants are not directly contacted by the investigator, but instead the investigator send questionnaires by post to the informants with the request of sending them back after filling the same. The suitability of this method depends upon the quality of the 'questionnaire' and the response of the informants. This method is useful when area to be covered is widely spread. This method would not work, in case the informants are illiterate or semi-literate.

Merits

(*a*) This method is economical in respect of money, labour and time.

(*b*) This method can be very easily used if the area to be covered is widely spread.

Demerits

(*a*) This method is not useful in case the informants are illiterate or semi-literate.

(*b*) The collected data is not expected to be very accurate due to lack of seriousness in the informants.

(*c*) The response of the informants is expected to be poor because people generally avoid giving reply in written statement form.

(*d*) The reliability of the data collected cannot be judged by the investigator.

(*e*) The data may be unduly affected by the expected bias of the informants.

2.8. THROUGH SCHEDULES FILLED IN BY ENUMERATORS

NOTES

In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the enumerators are deputed to contact the informants and to fill in the schedules on the spot, after collecting data as per the need of the schedule. The basic difference between this method and the previous method is that, in this method the schedules are filled in by the enumerators after getting information from the informants, whereas in the previous method, the questionnaires were to be filled in by the informants themselves. The suitability of this method depends upon the enumerators. The enumerators are expected to be skilled honest, hard working, well-behaved and free from bias. This method of collecting data is suitable in case the informants are illiterate or semi-literate. In our country, census data about all the citizens is collected after every ten years by using this method.

Merits

(*a*) In this method of collecting data, the degree of accuracy is expected to be very high.

(*b*) The data collected is very reliable.

(*c*) Because of the direct contact between the enumerators and the informants, data can also be gathered about sensitive questions by twisting the questions accordingly.

(*d*) The data collected is least affected by the bias of the enumerators and the informants.

(*e*) This method can also be used in case the informants are illiterate or semi-literate.

Demerits

(*a*) This method of collecting data is very expensive.

(*b*) Much time is taken in collecting data.

(*c*) The enumerator have to be trained before they are deputed to start collecting data.

2.9. REQUISITES OF A GOOD 'QUESTIONNAIRE' AND 'SCHEDULE'

In the last two methods of collecting primary data, we discussed the method of questionnaires to be filled in by the informants and the method of filling schedules by the enumerators. In fact, there is no fundamental difference between a questionnaire and a schedule. Both questionnaire and schedule contain some questions. The only difference between the two is that the former is filled in by the informants themselves, whereas in the case of later, the data concerning the informants is filled in by the enumerators. The success of collecting data by using either questionnaire or schedule depends upon the quality of itself. Preparation of questionnaire and schedule is an art. Now we shall discuss in detail the requisites of a good questionnaire and schedule.

NOTES

1. Forwarding Letter. The investigator must include a forwarding letter in case of sending questionnaires to the informants. The investigator must request the informants to fill in the same and to return it back after filling it. The object of the investigation should also be mentioned in the latter. The informants should also be ensured that the filled questionnaires would be kept secretly, if desired. To encourage the response of informants, special concessions and free gifts may be offered to the informants.

2. Questions should be Minimum in Number. The number of questions in a questionnaire or a schedule should be as small as possible. Unnecessary questions should never be included. Inclusion of more than 20 or 25 questions would be undesirable.

3. Questions should be Simple to Understand. The questions included in a questionnaire or a schedule should be simple to understand. The questions should not be confusing in nature. The language used should also be simple and the use of highly technical terms should also be avoided.

4. Questions should be Logically Arranged. The questions in a questionnaire or a schedule also be logically arranged. The questions should be arranged so that there is natural and spontaneous reaction of the informants to the questions. It is not fair to ask the informant whether he is employed or unemployed after asking his monthly income. Such sequence of questions create bad impression on the mind of the informants.

5. Only well-defined terms should be used in questions. In drafting questions for a questionnaire or a schedule, only well defined terms should be used. For example, the term 'income' should be clearly defined in the sense whether it is to include allowances etc. along with the basic income or not. Similarly, in case of businessman, whether the informants are to inform about their gross profits or net profits etc.

6. Prohibited questions should not be included. No such question should be included in the questionnaire or schedule which may immediately agitate the mind of the informants. Question like, "Have you given up the habit of telling a lie" or "How many times in a month, do you quarrel with your wife", would immediately mar the spirit of the informants.

7. Irrelevant questions should be avoided. In questionnaire or schedule, only those questions should be included which bears direct link with the object of the investigation. If the object is to study the problem of unemployment, then it would be useless to collect data regarding the heights and weights of the informants.

8. Pilot Survey. Before the questionnaire is sent to all the informants for collecting data, it should be checked before hand for its workability. This is done by sending the questionnaire to a selected sample and the replies received are studied thoroughly. If the investigator finds that most of the informants in the sample have left some questions un-answered then those questions should be modified or deleted altogether, provided the object of the investigation permits to do so. This is called *Pilot Survey*. Pilot Survey must be carried out before the questionnaire is finally accepted.

NOTES

SPECIMEN OF QUESTIONNAIRE

Department of Industries,
....

Dear student,

We are conducting this investigation to know about the future planning of our students, studying at present in colleges. The Govt. has decided to allocate funds for helping deserving students in establishing their career. You are requested to fill in the enclosed questionnaire and send it back. A stamped self-addressed envelope has also been enclosed for your convenience.

Thanking you,

Yours sincerely,
....

QUESTIONNAIRE

Department of Industries

Object : Future Planning of Students (2021)

1. Name ....
2. Class ....
3. Occupation of your father ....
4. Address ....
5. Which of the following occupation would you like to choose after completing your education
   (*a*) Business ()
   (*b*) Service ()
   (*c*) Manufacturing concern ()
   (*d*) Doctor ()
   (*e*) Any other ()
6. Why do you want to choose this occupation ?
   (*a*) Less competition ()
   (*b*) Greater possibility of money making ()
   (*c*) Future is bright ()
   (*d*) Parental occupation ()
7. If you plan to establish a manufacturing concern
   (*a*) What amount of foreign exchange you can earn ....
   (*b*) Which of the countries can be the expected buyers of your product ....
8. Sources of raw material ....
9. Sources of machinery ....
10. Gestation period ....
11. What type of help would you seek from the Govt.
    (*a*) Financial ()
    (*b*) Technical knowledge ()
    (*c*) Imported machinery ()
    (*d*) Rent free land ()
12. State the amount of financial help which you would like to have
    ....
13. In what way, your product would be superior to those of producing the same product ....
14. In what way, you would be helping the nation to prosper
    ....

SPECIMEN OF SCHEDULE

Government of Haryana

Ration Card

Sr. No. ....

Name of City/Village ....

Name of Head of Family ....

Name of Father/Husband ....

House No. ....

Ward/Sector No. ....

Detail of Family

No. of units

Cereal ( ) Sugar ( )

....

Sig./Thumb Impression
of Head of Family

Signature and Designation of
authorised officer

Sr. No.	*Name*	*Relation with Head of Family*	*Age*
——	——	——	——
——	——	——	——

Total number of members :

Above 12 Years ( ) Between 2 Years and 12 Years ( )

Below 2 Years ( )

To be filled in by depot holder

S. No. and Address of the Depot ....

Registration No. ....

....

Sig. and stamp of the
Depot Holder

II. SECONDARY DATA

NOTES

2.10. DEFINITION OF SECNDARY DATA

Data is called **secondary** if it is not originally collected in the process of investigation, but instead, the data collected by some other agency is used for the purpose. If the investigation is not of very special nature, then the use of secondary data may be made provided that can serve the purpose. Suppose we want to investigate the extent of poverty in our country, then this investigation can be carried out by using the national census data which is obtained regularly after every 10 years. The use of secondary data economise the money spent. It also reduces the time period of investigation to a great extent. If in an investigation some secondary data could be made use of, then we must use the same. The secondary data are ought to be used very carefully. In this context, *Connor* has remarked, "Statistics, especially other peoples' statistics are full of pitfalls for the user."

2.11. METHODS OF COLLECTING SECONDARY DATA

(*i*) Collection from Published Data

(*ii*) Collection from Un-published Data.

2.11.1 Collection From Published Data

There are agencies which collect statistical data regularly and publish it. The published data is very important and is used frequently by investigators. The main sources of published data are as follows :

(*a*) *International Publications*. International Organisations and Govt. of foreign countries collect and publish statistical data relating to various characteristics. The data is collected regularly as well as on ad-hoc basis. Some of the publications are :

(*i*) U.N.O. Statistical Year Book

(*ii*) Annual Reports of I.L.O.

(*iii*) Annual Reports of the Economic and Social Commission for Asia and Pacific (ESCAP)

(*iv*) Demography Year Book

(*v*) Bulletins of World Bank.

(*b*) *Govt. Publications*. In India, the Central Govt. and State Govt. collects data regarding various aspects. This data is published and is found very useful for investigation purpose. Some of the publications are :

(*i*) Census Report of India

(*ii*) Five-Year Plans

(*iii*) Reserve Bank of India Bulletin

(*iv*) Annual Survey of Industries

(*v*) Statistical Abstracts of India.

(c) *Report of Commissions and Committees.* The Central Govt. and State Govt. appoints Commissions and Committees to study certain issues. The reports of such investigations are very useful. Some of these are :

(i) Reports of National Labour Commission

(ii) Reports of Finance Commission

(iii) Report of Hazari Committee etc.

(d) *Publications of Research Institutes:* There are number of research institutes in India which regularly collect data and analyse it. Some of the agencies are :

(i) Central Statistical Organisation (C.S.O.)

(ii) Institute of Economic Growth

(iii) Indian Statistical Institute

(iv) National Council of Applied Economic Research etc.

(e) *Newspapers and Magazines.* There are many newspapers and magazines which publish data relating to various aspects. Some of these are :

(i) Economic Times (ii) Financial Express

(iii) Commerce (iv) Transport

(v) Capital etc.

(f) *Reports of Trade Associations.* The trade associations also collect data and publish it. Some of the agencies are :

(i) Stock Exchanges (ii) Trade Unions

(iii) Federation of Indian Chamber of Commerce and Industry.

2.11.2 Collection From Un-published Data

The Central Government, State Government and Research Institutes also collect data which is not published due to some reasons. This type of data is called un-published data. Un-published data can also be made use of in Investigations. The data collected by research scholars of Universities is also generally not published.

2.12. PRECAUTIONS IN THE USE OF SECONDARY DATA

The secondary data must be used very carefully. The applicability of the secondary data should be judged keeping in view the object and scope of the Investigation. *Prof. Bowley* has remarked, "Secondary data should not be accepted at their face value." Following are the basis on which the applicability of secondary data is to be judged.

(i) **Reliability of Data.** Reliability of data is assessed by reliability of the agency which collected that data. The agency should not be biased in any way. The enumerators who collected the data should have been unbiased and well trained. Degree of accuracy achieved should also be judged.

(ii) **Suitability of Data.** The suitability of the data should be assessed keeping in view the object and scope of Investigation. If the data is not suitable for the investigation, then it is not to be used just for the sake of economy of time and money. The use of unsuitable data can lead to only misleading results.

(iii) **Adequacy of Data.** The adequacy of data should also be judged keeping in view the object and scope of the investigation. If the data is found to be inadequate, it should not be used. For example, if the object of investigation is to study the problem of unemployment in India, then the data regarding unemployment in one state say U.P. would not serve the purpose.

NOTES

NOTES

EXERCISE 2.1

1. Discuss the merits and limitation of any three methods of collecting Primary Data.
2. Explain the various methods that are used in the collection of Primary Data, pointing out their merits and demerits.
3. Distinguish between primary source and secondary source of statistical data. What precautions would you take before using data from a secondary source ?
4. What are the various methods of collection of primary data ? Briefly explain two such methods, pointing out their merits and demerits.
5. What are the various methods of collecting statistical data ?

   Which of these is not reliable and why ?
6. Describe the points that you would consider in drafting a questionnaire ?
7. Why is it necessary that secondary data must be scrutinised and edited before use ? What precautions would you take before making use of such statistical data ?
8. Distinguish clearly between primary and secondary data. Explain the various methods of collecting primary data and point out their respective merits and demerits.
9. What are the various methods of collecting statistical data ? Which of these are not reliable and why ?
10. What are the various methods of collecting primary data ? Why we prefer primary data than secondary data ? Explain.

2.13. SUMMARY

- The first step in a statistical investigation is the planning of the proposed investigation. After planning, the next step is the collection of data, keeping in view the object and scope of investigation.
- Data is called **primary,** if it is originally collected in the process of investigation.
- In this method of collecting data, the investigator directly comes in contact with the informats to collect data. The investigator himself visits the different informants covered in the scope of the investigation and collect data as per the need of the investigation.
- In this method of collecting data, the informants are not directly contacted by the investigator, but instead, the data about the informants is collected from some selected persons who are expected to be acquinted with the informants as well as the object of the investigation.
- The investigator is expected to be tactful, skilled, honest and well behaved.
- There are agencies which collect statistical data regularly and publish it. The published data is very important and is used frequently by investigators.
- The Central Government, State Government and Research Institutes also collect data which is not published due to some reasons. This type of data is called unpublished data. Un-published data can also be made use of in.Investigations. The data collected by research scholars of Universities is also generally not published.

NOTES

3. MEASURES OF CENTRAL TENDENCY

STRUCTURE

3.1. TYPES OF MEASURES OF CENTRAL TENDENCY (Averages)

I. Arithmetic Mean (A.M.)   II. Geometric Mean (G.M.)
III. Harmonic Mean (H.M.)   IV. Median
V. Mode.

I. ARITHMETIC MEAN (A.M.)

NOTES

3.2. DEFINITION OF ARITHMETIC MEAN

This is the most popular and widely used measure of central tendency. The popularity of this average can be judged from the fact that it is generally referred to as 'mean'. The **arithmetic mean** of a statistical data is defined as the quotient of the sum of all the values of the variable by the total number of items and is generally denoted by $\bar{x}$.

∴ (*a*) **For an individual series**, the A.M. is given by

$$\text{A.M.} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ or more briefly as } \frac{\Sigma x}{n}$$

i.e., $$\bar{\mathbf{x}} = \frac{\Sigma \mathbf{x}}{\mathbf{n}}$$

where $x_1, x_2, \ldots\ldots, x_n$ are the values of the variable, under consideration.

(*b*) **For a frequency distribution,**

$$\text{A.M.} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots + f_nx_n}{f_1 + f_2 + \ldots\ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\Sigma fx}{\Sigma f} = \frac{\Sigma fx}{\text{N}},$$

i.e., $$\bar{\mathbf{x}} = \frac{\Sigma \mathbf{fx}}{\mathbf{N}}$$

where f_i is the frequency of x_i $(1 \le i \le n)$. For simplicity, Σf, *i.e.*, the total number of items is denoted by N.

When the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable (x).

WORKING RULES TO FIND A.M.

Rule I. *In case of an individual series, first find the sum of all the items. In the second step, divide this sum by n, total number of items. This gives the value of* $\bar{x}$.

Rule II. *In case of a frequency distribution, find the products (fx) of frequencies and value of items. In the second step, find the sum (Σfx) of these products. Divide this sum by the sum (N) of all frequencies. This gives the value of* $\bar{x}$.

Rule III. *If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.*

Example 1. *Find the A.M. of the following data :*

Roll No.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Marks in Maths	*12*	*8*	*6*	*9*	*7*	*8*	*7*	*14*

Solution. Let the variable 'marks in maths' be denoted by x.

$$\therefore \quad \bar{x} = \frac{\text{Sum of values of } x}{\text{Number of items}} = \frac{12+8+6+9+7+8+7+14}{8} = \frac{71}{8} = \textbf{8.875 marks.}$$

Example 2. *The average weight of a group of 24 boys was calculated to be 78.4 kg. It was later discovered that one weight was misread as 69 kg instead of 96 kg. Calculate the correct average (Average used is A.M.).*

Solution. No. of items = 25

Incorrect average = 78.4 kg

Incorrect item = 69 kg

Correct item = 96 kg

Let the variable 'weight' be denoted by 'x'.

Now $$\bar{x} = \frac{\Sigma x}{n}$$

$$\therefore \quad \text{Incorrect } \bar{x} = \frac{\text{incorrect } \Sigma x}{25}$$

$$\therefore \quad 78.4 = \frac{\text{incorrect } \Sigma x}{25}$$

$$\therefore \quad \text{Incorrect } \Sigma x = 78.4 \times 25 = 1960 \text{ kg}$$

The correct $\bar{x}$ is obtained by using correct Σx in the formula.

$$\text{Correct } \Sigma x = \text{incorrect } \Sigma x - \text{incorrect item} + \text{correct item}$$
$$= 1960 - 69 + 96 = 1987 \text{ kg}$$

$$\therefore \quad \text{Correct } \bar{x} = \frac{\text{correct } \Sigma x}{25} = \frac{1987}{25} = \mathbf{79.48 \text{ kg.}}$$

Example 3. *Calculate A.M. for the following data :*

Income (in ₹)	*500*	*520*	*550*	*600*	*800*	*1000*
No. of employees	*4*	*10*	*6*	*5*	*3*	*2*

Solution. **Calculation of A.M.**

S. No.	*Income (in ₹)* x	*No. of employees* f	fx
1	500	4	2000
2	520	10	5200
3	550	6	3300
4	600	5	3000
5	800	3	2400
6	1000	2	2000
		N = 30	$\Sigma fx = 17900$

Now $$\bar{x} = \frac{\Sigma fx}{N} = \frac{17900}{30} = \text{₹ } \mathbf{596.67.}$$

Example 4. *Calculate the A.M. for the following data :*

Marks	*0–10*	*10–30*	*30–40*	*40–50*	*50–80*	*80–100*
No. of students	*5*	*7*	*15*	*8*	*3*	*2*

NOTES

Solution. **Calculation of A.M.**

Marks	*No. of students* *f*	*Mid-points of classes* *x*	*fx*
0–10	5	5	25
10–30	7	20	140
30–40	15	35	525
40–50	8	45	360
50–80	3	65	195
80–100	2	90	180
	N = 40		$\Sigma fx = 1425$

$$\therefore \quad \bar{x} = \frac{\Sigma fx}{N} = \frac{1425}{40} = \textbf{35.625 marks.}$$

3.2.1 Step Deviation Method

When the values of the variable (x) and their frequencies (f) are large, the calculation of A.M. may become quite tedious. The calculation work can be reduced considerably by taking *step deviations* of the values of the variable.

Let A be any number, called **assumed mean**, then $d = x - \text{A}$ are called the **deviations** of the values of x, from A.

If the values of x are $x_1, x_2, ..., x_n$, then the values of deviations are $d_1 = x_1 - \text{A}, d_2 = x_2 - \text{A}, ..., d_n = x_n - \text{A}$. We define $u = \frac{x - \text{A}}{h}$, where h is some suitable common factor in the deviations of values of x from A. The definition of 'u' is meaningful, because at least $h = 1$ is a common factor for all the values of the deviations. The different values of $u = \frac{x - \text{A}}{h}$ are called the **step deviations** of the corresponding values of x. In this case, the values of the step deviations are $u_1 = \frac{x_1 - \text{A}}{h}$, $u_2 = \frac{x_2 - \text{A}}{h}$, ..., $u_n = \frac{x_n - \text{A}}{h}$.

$$\therefore \quad \text{For} \quad 1 \le i \le n, \quad u_i = \frac{x_i - \text{A}}{h} \quad i.e., \; x_i = \text{A} + u_i h$$

$$\therefore \quad \bar{x} = \frac{1}{\text{N}}\Sigma f_i x_i = \frac{1}{\text{N}}\Sigma f_i(\text{A} + u_i h) = \frac{1}{\text{N}}\Sigma f_i \text{A} + \frac{1}{\text{N}}\Sigma f_i u_i h$$

$$= \text{A} . \frac{\Sigma f_i}{\text{N}} + \frac{1}{\text{N}}(\Sigma f_i u_i)h = \text{A} + \frac{\Sigma f_i u_i}{\text{N}} h \qquad (\because \quad \Sigma f_i = \text{N})$$

NOTES

$$\therefore \qquad \bar{\mathbf{x}} = \mathbf{A} + \left(\frac{\Sigma \mathbf{f_i u_i}}{\mathbf{N}}\right)\mathbf{h}.$$

In brief, the above formula is written as $\bar{\mathbf{x}} = \mathbf{A} + \left(\frac{\Sigma \mathbf{fu}}{\mathbf{N}}\right)\mathbf{h}.$

In case of individual series, this formula takes the form $\bar{\mathbf{x}} = \mathbf{A} + \left(\frac{\Sigma \mathbf{u}}{\mathbf{n}}\right)\mathbf{h}.$

In dealing with practical problems, it is advisable to first take deviations (d) of the values of the variable (x) from some suitable number (A). Then we see, if there is any common factor, greater than one in the values of the deviations. If there is a common factor $h(> 1)$, then we calculate $u = \frac{d}{h} = \frac{x - A}{h}$ in the next column. In case, there is no common factor other than one, then we take $h = 1$ and u becomes $\frac{d}{1} = d = x - A$. In this case, the formulae reduces as given below :

$$\bar{\mathbf{x}} = \mathbf{A} + \frac{\Sigma \mathbf{d}}{\mathbf{n}} \qquad \textbf{(For Individual Series)}$$

$$\bar{\mathbf{x}} = \mathbf{A} + \frac{\Sigma \mathbf{fd}}{\mathbf{N}} \qquad \textbf{(For Frequency Distribution)}$$

where d = x – A and A is any constant ; to be chosen suitably.

WORKING RULES TO FIND A.M.

Rule I. *In case of an individual series, choose a number A. Find deviations d(= x – A) of items from A. Find the sum 'Σd' of the deviations. Divide this sum by n, the total number of items. This quotient is added to A to get the value of* $\bar{x}$.

If some common factor h (> 1) is available in the values of d, then we calculate 'u' by dividing the values of d by h and find $\bar{x}$ *by using the formula :*

$$\bar{x} = A + \left(\frac{\Sigma x}{n}\right)h.$$

Rule II. *In case of a frequency distribution, choose a number A. Find deviations d(= x – A) of items from A. Find the products fd of f and d. Find the sum 'Σfd' of these products. Divide this sum by N, the total number of items. This quotient is added to A to get the value of* $\bar{x}$.

If some common factor h(> 1) is available in the values of d, then we calculate 'u' dividing d by h and find $\bar{x}$ *by using the formula :*

$$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h.$$

Rule III. *If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.*

NOTES

Example 5. *Find the A.M. for the following individual series :*

12.36, 14.36, 16.36, 18.36, 20.36, 24.36.

Solution. **Calculation of A.M.**

Variable *x*	$d = x - A$ $A = 16.36$	$u = d/h$ $h = 2$
12.36	– 4	– 2
14.36	– 2	– 1
16.36	0	0
18.36	2	1
20.36	4	2
24.36	8	4
		$\Sigma u = 4$

Now $$\bar{x} = A + \left(\frac{\Sigma u}{n}\right)h = 16.36 + \left(\frac{4}{6}\right)2 = 16.36 + 1.33 = \mathbf{17.69}.$$

Example 6. *Find the A.M. for the following distribution :*

Marks	*No. of students*	*Marks*	*No. of students*
Above 0	*80*	*Above 60*	*28*
Above 10	*77*	*Above 70*	*16*
Above 20	*72*	*Above 80*	*10*
Above 30	*65*	*Above 90*	*8*
Above 40	*55*	*Above 100*	*0*
Above 50	*43*		

Solution. **Calculation of A.M.**

Marks	*Mid-points* *x*	*No. of students* *f*	$d = x - A$ $A = 55$	$u = d/h$ $h = 10$	*fu*
0—10	5	3	– 50	– 5	– 15
10—20	15	5	– 40	– 4	– 20
20—30	25	7	– 30	– 3	– 21
30—40	35	10	– 20	– 2	– 20
40—50	45	12	– 10	– 1	– 12
50—60	55	15	0	0	0
60—70	65	12	10	1	12
70—80	75	6	20	2	12
80—90	85	2	30	3	6
90—100	95	8	40	4	32
		N = 80			$\Sigma fu = -26$

Now $$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = 55 + \left(\frac{-26}{80}\right)10 = 55 - 3.25 = \mathbf{51.75\ marks.}$$

3.2.2 A.M. of Combined Group

Theorem. If $\bar{x}_1$ and $\bar{x}_2$ are the A.M. of two groups having n_1 and n_2 items, then the A.M. ($\bar{x}$) of the combined group is given by

$$\mathbf{\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}}.$$

Proof. Let $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ be the items in the two groups respectively.

$$\therefore \quad \bar{x}_1 = \frac{x_1 + x_2 + ... + x_{n_1}}{n_1}$$

$$\bar{x}_2 = \frac{y_1 + y_2 + ... + y_{n_2}}{n_2}$$

$$\therefore \quad x_1 + x_2 + ... + x_{n_1} = n_1\bar{x}_1$$

$$y_1 + y_2 + ... + y_{n_2} = n_2\bar{x}_2$$

Now
$$\bar{x} = \frac{\text{sum of items in both groups}}{n_1 + n_2}$$

$$= \frac{x_1 + x_2 + ... + x_{n_1} + y_1 + y_2 + ... + y_{n_2}}{n_1 + n_2} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$\therefore \quad \mathbf{\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}}.$$

This formula can also be extended to more than two groups.

Example 7. *The mean wage of 1000 workers in a factory running two shifts of 700 and 300 workers is ₹ 500. The mean wage of 700 workers, working in the day shift, is ₹ 450. Find the mean wage of workers, working in the night shift.*

Solution. No. of workers in the day shift (n_1) = 700

No. of workers in the night shift (n_2) = 300

Mean wage of workers in the day shift ($\bar{x}_1$) = ₹ 450

Mean wage of all workers ($\bar{x}$) = ₹ 500

Let mean wage of workers in the night shift = $\bar{x}_2$

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

$$\therefore \quad 500 = \frac{700\,(450) + 300\,(\bar{x}_2)}{700 + 300} \quad \text{or} \quad 500000 = 315000 + 300\bar{x}_2$$

$$\therefore \quad 300\bar{x}_2 = 185000$$

$$\therefore \quad \bar{x}_2 = \frac{185000}{300} = \textbf{₹ 616.67.}$$

3.3. WEIGHTED A.M.

NOTES

If all the values of the variable are not of equal importance, or in other words, these are of varying significance, then we calculate **weighted A.M.**

$$\text{Weighted A.M.} = \bar{x}_w = \frac{\Sigma wx}{\Sigma w}$$

where $w_1, w_2, \ldots\ldots, w_n$ are the weights of the values $x_1, x_2, \ldots\ldots, x_n$ of the variable, under consideration.

Example 8. *An examination was held to decide the award of a scholarship. The weights given to different subjects were different. The marks were as follows :*

Subjects	*Weight*	*Marks of A*	*Marks of B*	*Marks of C*
Statistics	*4*	*63*	*60*	*65*
Accountancy	*3*	*65*	*64*	*70*
Economics	*2*	*58*	*56*	*63*
Mercantile Law	*1*	*70*	*80*	*52*

The candidate getting the highest marks is to be awarded the scholarship. Who should get it ?

Solution. **Calculation of weighted A.M.**

Subject	*Weight* w	*Marks of A* x_1	wx_1	*Marks of B* x_2	wx_2	*Marks of C* x_3	wx_3
Statistics	4	63	252	60	240	65	260
Accountancy	3	65	195	64	192	70	210
Economics	2	58	116	56	112	63	126
Mercantile Law	1	70	70	80	80	52	52
	$\Sigma w = 10$		$\Sigma wx_1 = 633$		$\Sigma wx_2 = 624$		$\Sigma wx_3 = 648$

Weighted A.M. of $\text{A} = \dfrac{\Sigma wx_1}{\Sigma w} = \dfrac{633}{10} = \mathbf{63.3}$

Weighted A.M. of $\text{B} = \dfrac{\Sigma wx_2}{\Sigma w} = \dfrac{624}{10} = \mathbf{62.4}$

Weighted A.M. of $\text{C} = \dfrac{\Sigma wx_3}{\Sigma w} = \dfrac{648}{10} = \mathbf{64.8}$

∴ The student 'C' is to get the scholarship.

3.3.1 Merits of A.M.

1. It is the simplest average to understand.
2. It is easy to compute.
3. It is well-defined.
4. It is based on all the items.
5. It is capable of further algebraic treatment.
6. It has sampling stability.
7. It is specially used in finding the average speed, when time taken at different speeds are varying, or are equal.

3.3.2 Demerits of A.M.

1. It may not be present in the given series itself. For example, the A.M. of 4, 5, 6, 6 is $\frac{4+5+6+6}{4}$ = 5.25, which is not present in the series. So, sometimes it becomes theoretical.
2. It cannot be calculated for qualitative data.
3. It may be badly affected by the extreme item.

EXERCISE 3.1

1. Find the A.M. of the series 4, 6, 8, 10, 12.
2. The mean marks of 100 students was found to be 40. Later on, it was discovered that a score of 53 was misread as 83. Find the correct mean.
3. The A.M. of 25 items is found to be 78.4. If at the time of calculation, two items were wrongly taken as 96 and 43 instead of 69 and 34, find the value of the correct mean.
4. The marks obtained by 5 students are 11, 17, 9, 16, 22. Later on, 3 grace marks were awarded to each student. Find the mean marks of the increased marks of the students.
5. Find the arithmetic mean for the following data :

x	6	7	8	9	10
f	7	10	12	6	5

6. The postal expenses on the letters despatched from an office on a given day resulted in the following frequency distribution :

Postage (in paise)	15	30	35	60	70
No. of letters	47	33	56	41	25

Find the mean postage per letter. Convert the postal charges in rupees and then calculate the mean postage per letter.

7. Find the A.M. for the following frequency distribution :

Marks obtained	0—7	7—14	14—21	21—28
No. of Students	19	25	36	72
Marks obtained	28—35	35—42	42—49	
No. of Students	51	43	28	

NOTES

8. Calculate the arithmetic mean for the following data :

Wages (in ₹)	*No. of persons*	*Wages (in ₹)*	*No. of workers*
Less than 10	30	40 and above	332
Less than 20	70	50 and above	308
20—30	50	60—70	132
20—40	98	70 and above	14

9. From the following information, find out :

(*i*) Which of the factor pays larger amount as daily wages.

(*ii*) What is the average daily wage of the workers of two factories taken together.

	Factory A	*Factory B*
No. of wage earners	250	200
Average daily wages	₹ 20	₹ 25

10. The mean wage of 100 workers in a factory running two shifts of 60 and 40 workers is ₹ 38. The mean wage of 60 workers working in the day shift is ₹ 40. Find the mean wage of workers, working in the night shift.

11. The mean weight of 15 students is 110 lbs. The mean weight of 5 of them is 100 lbs and of another 5 is 125 lbs. What is the mean weight of the remaining students ?

12. Fifty students took up a test. The result of those who passed the test is given below :

Marks	4	5	6	7	8	9
No. of students	8	10	9	6	4	3

If the average of all the 50 students was 5.16 marks, find the average of those who failed.

13. The average weight of 150 students in a class is 80 kg. The average weight of boys in the class is 85 kg and that of girls is 70 kg. Tell the number of boys and girls in the class separately.

14. From the following results of two colleges A and B, find out which of the two is better.

Examination	*College A*		*College B*	
	Appeared	*Passed*	*Appeared*	*Passed*
B.Sc.	100	90	240	200
M. Com.	60	45	200	160
B. Com.	120	75	160	60
B.A.	200	150	200	140
Total	480	360	800	560

Answers

1. 8 **2.** 39.7 marks **3.** 76.96 **4.** 18 marks

5. 7.8 **6.** Paise 38.94, ₹ 0.39 **7.** 26.4927 marks **8.** ₹ 46.60

9. (*i*) Both factories are paying equal amount (*ii*) ₹ 22.22 **10.** ₹ 35

11. 105 lbs **12.** 2.1 marks **13.** Boys = 100, Girls = 50

14. College A is better.

II. GEOMETRIC MEAN (G.M.)

NOTES

3.4. DEFINITION OF GEOMETRIC MEAN

The **geometric mean** of a statistical data is defined as the nth root of the product of all the n values of the variable.

For an individual series, the G.M. is given by

$$\text{G.M.} = (x_1\, x_2 \,......\, x_n)^{1/n}$$

where x_1, x_2, ..., x_n are the values of the variable, under consideration. From the definition of G.M. we see that it involves the nth root of a product, which is not possible to evaluate by using simple arithmetical tools. To solve this problem, we take the help of logarithms.

We have $\quad \textbf{G.M.} = \mathbf{(x_1 x_2 \,......\, x_n)^{1/n}}$

$$= \text{Antilog } [\log (x_1 x_2 \,......\, x_n)^{1/n}] \;= \text{Antilog}\left[\frac{1}{n}\log(x_1 x_2 \,.....x_n)\right]$$

$$= \text{Antilog}\left[\frac{1}{n}(\log x_1 + \log x_2 + ... + \log x_n)\right]$$

$$\therefore \qquad \textbf{G.M.} = \textbf{Antilog}\left(\frac{\Sigma \log \mathbf{x}}{\mathbf{n}}\right)$$

For a frequency distribution,

$$\textbf{G.M.} = \mathbf{(x_1{}^{f_1}\; x_2{}^{f_2} \,......\, x_n{}^{f_n})^{1/N}}$$

where f_i is the frequency of x_i $(1 \le i \le n)$.

Proceeding on the same lines, we get

$$\textbf{G.M.} = \textbf{Antilog}\left(\frac{\Sigma\, \mathbf{f} \log \mathbf{x}}{\mathbf{N}}\right)$$

When the values of the variable are given in the form of classes, the mid-points are taken as the values of the variable (x).

WORKING RULES TO FIND G.M.

Rule I. *In case of an individual series, first find the sum of logarithms of all the items. In the second step, divide this sum by n, the total number of items. Next, take the 'antilogarithm' of this quotient. This gives the value of the G.M.*

Rule II. *In case of a frequency distribution, find the product (f log x) of frequencies and logarithm of value of items. In the second step, find the sum (Σ f log x) of these products. Divide this sum by the sum (N) of all the frequencies. Next, take the 'antilogarithm' of this quotient. This gives the value of the G.M.*

Rule III. *If the values of the variables are given in the form of classes, then their respective mid-points are taken as the values of the variable.*

NOTES

Example 9. *Calculate the G.M. for the following individual series 85, 70, 15, 75, 500, 8, 45, 250, 40, 36.*

Solution. **Calculation of G.M.**

S. No.	*x*	*log x*
1	85	1.9294
2	70	1.8451
3	15	1.1761
4	75	1.8751
5	500	2.6990
6	8	0.9031
7	45	1.6532
8	250	2.3979
9	40	1.6021
10	36	1.5563
n = 10		Σ log *x* = 17.6373

$$\therefore \qquad \text{G.M.} = \text{Antilog}\left(\frac{\Sigma \log x}{n}\right) = \text{Antilog}\left(\frac{17.6373}{10}\right)$$

$$= \text{Antilog } (1.76373) = ₹\ 58.03.$$

Example 10. *Find the G.M. for the data given below :*

Yield of wheat (in quintals)	*7.5—10.5*	*10.5—13.5*	*13.5—16.5*	*16.5—19.5*
No. of farms	*5*	*9*	*19*	*23*
Yield of wheat (in quintals)	*19.5—22.5*	*22.5—25.5*	*25.5—28.5*	
No. of farms	*7*	*4*	*1*	

Solution. **Calculation of G.M.**

Class	*Mid-point x*	*f*	*log x*	*f log x*
7.5—10.5	9	5	0.9542	4.7710
10.5—13.5	12	9	1.0792	9.7128
13.5—16.5	15	19	1.1761	22.3459
16.5—19.5	18	23	1.2553	28.8719
19.5—22.5	21	7	1.3222	9.2554
22.5—25.5	24	4	1.3802	5.5208
25.5—28.5	27	1	1.4314	1.4314
		N = 68		Σ *f* log *x* = 81.9092

Now
$$G = \text{Antilog}\left(\frac{\Sigma f \log x}{\text{N}}\right) = \text{Antilog}\left(\frac{81.9092}{68}\right)$$

$$= \text{Antilog } (1.2045) = \textbf{16.02 quintals.}$$

3.4.1 G.M. of Combined Group

Theorem. If G_1 and G_2 are the GMs of two groups having n_1 and n_2 items, then the G.M. (G) of the combined group is given by

$$\mathbf{G = Antilog\left(\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right).}$$

Proof. Let $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ be the items in the two groups respectively.

$$\therefore \quad G_1 = \text{Antilog}\left(\frac{\Sigma \log x}{n_1}\right)$$

$$\therefore \quad \log G_1 = \frac{\Sigma \log x}{n_1}$$

$$\therefore \quad n_1 \log G_1 = \Sigma \log x$$

Similarly, $n_2 \log G_2 = \Sigma \log y$

$$\text{Now} \quad G = \text{Antilog}\left(\frac{\text{sum of logarithms of all items}}{\text{no. of items in both groups}}\right)$$

$$= \text{Antilog}\left(\frac{\Sigma \log x + \Sigma \log y}{n_1 + n_2}\right)$$

$$\therefore \quad \mathbf{G = Antilog\left(\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right).}$$

This formula can also be extended to more than two groups.

Example 11. *The G.M. of wages of 200 workers working in a factory is ₹ 700. The G.M. of wages of 300 workers, working in another factory is ₹ 1000. Find the G.M. of wages of all the workers taken together.*

Solution. No. of workers in I factory (n_1) = 200

No. of workers in II factory (n_2) = 300

G.M. of wages of workers of I factory (G_1) = ₹ 700

G.M. of wages of workers of II factory (G_2) = ₹ 1000

Let G be the G.M. of wages of all the workers taken together.

$$\therefore \quad G = \text{Antilog}\left(\frac{n_1 \log G_1 + n_2 \log G_2}{n_1 + n_2}\right) = \text{Antilog}\left(\frac{200 \log 700 + 300 \log 1000}{200 + 300}\right)$$

$$= \text{Antilog}\left(\frac{200\ (2.8451) + 300\ (3.0000)}{500}\right) = \text{Antilog}\left(\frac{569.0200 + 900}{500}\right)$$

$$= \text{Antilog}\ (2.9380) = ₹\ \mathbf{867.}$$

3.5. AVERAGING OF PERCENTAGES

Geometric mean is specially used to find the average rate of increase or decrease in sale, production, population etc.

NOTES

If V_0 and V_n are the values of a variable at the beginning of the first and at the end of the *n*th period, then

$$\mathbf{V_n = V_0 \ (1 + r)^n}, \text{ where } \mathbf{r} \text{ is the average rate of growth per unit.}$$

Example 12. *A gave ₹ 10,000 to B on the terms that after expiry of 5 years, B will return him ₹ 12,294. What is the rate of interest ?*

Solution. Here $V_0 = 10{,}000$ and $V_5 = 12{,}294$.

Let *r* be the average rate of interest per rupee.

$$\therefore \quad V_5 = V_0 \ (1 + r)^5$$

$$\text{or} \quad 12{,}294 = 10{,}000 \ (1 + r)^5$$

$$\text{or} \quad (1 + r)^5 = \frac{12{,}294}{10{,}000} = 1.2294$$

$$\therefore \quad 5 \log (1 + r) = \log 1.2294 = 0.1120$$

$$\therefore \quad \log (1 + r) = 0.0224$$

$$\therefore \quad 1 + r = \text{Antilog } 0.0224 = 1.053$$

$$\therefore \quad r = 1.053 - 1 = 0.053$$

∴ Average percentage rate of interest = 0.053 × 100 = **5.3%.**

Example 13. *The annual rate of growth of output of a factory in 5 years are 5.0, 7.5, 2.5, 5.0 and 10.0 percent respectively. What is compound rate of growth per annum for the period ?*

Solution.

Year	*Rate of growth*	*Production at the end of the year, taking 100 in the beginning* *x*	*log x*
I	10%	105	2.0212
II	7.5%	107.5	2.0314
III	2.5%	102.5	2.0107
IV	5%	105	2.0212
V	10%	110	2.0414
			$\Sigma \log x = 10.1259$

$$\therefore \quad \text{G.M.} = \text{antilog} \left(\frac{\Sigma \log x}{n} \right) = \text{Antilog} \left(\frac{10.1259}{5} \right)$$

$$= \text{Antilog } 2.02518 = 105.9$$

∴ Average rate of growth = 105.9 – 100 = **5.9%.**

3.6. WEIGHTED G.M.

If all the values of the variable are not of equal importance, or in other words, these are of varying significance, then we calculate **weighted G.M.**

$$\text{Weighted G.M.} = \text{Antilog} \left(\frac{\Sigma w \log x}{\Sigma w} \right),$$

where $w_1, w_2, ..., w_n$ are the weights of the values $x_1, x_2, ..., x_n$ of the variable, under consideration.

Example 14. *The weighted G.M. of four numbers 8, 25, 17 and 30 is 15.3. If the weights of the first three numbers are 5, 3 and 4 respectively, find the weight of the fourth number.*

NOTES

Solution. Let K be the weight of the fourth number.

x	w	$\log x$	$w \log x$
8	5	0.9031	4.5155
25	3	1.3979	4.1937
17	4	1.2304	4.9216
30	K	1.4771	1.4771 K
	$\Sigma w = 12 + \text{K}$		$\Sigma w \log x$ $= 13.6308 + 1.4771\text{ K}$

$$\text{Now weighted G.M.} = \text{Antilog}\left(\frac{\Sigma w \ \log x}{\Sigma w}\right)$$

$$\therefore \quad 15.3 = \text{Antilog}\left(\frac{13.6308 + 1.4771\text{K}}{12 + \text{K}}\right).$$

$$\therefore \quad (12 + \text{K}) \log 15.3 = 13.6308 + 1.4771\text{ K}$$

$$\therefore \quad (1.1847)(12 + \text{K}) = 13.6308 + 1.4771\text{ K}$$

$$\text{or} \quad 14.2164 + 1.1847\text{ K} = 13.6308 + 1.4771\text{ K}$$

$$\text{or} \quad 14.2164 - 13.6308 = (1.4771 - 1.1847)\text{ K}$$

$$\text{or} \quad \text{K} = \frac{0.5856}{0.2924} = 2.0027 = \mathbf{2\ (Approx.)}.$$

3.6.1 Merits of G.M.

1. It is well defined.
2. It is based on all the items.
3. It is capable of further algebraic treatment.
4. It is used to find the average rate of increase or decrease in the variables like sale, production, population etc.
5. It is specially used in the construction of index numbers.
6. It is used when larger weights are to be given to smaller items and smaller weights to larger items.
7. It has sampling stability.

3.6.2 Demerits of G.M.

1. It is not simple to understand.
2. It is not easy to compute.
3. It may become imaginary in the presence of negative items.
4. If any one item is zero, then its value would be zero, irrespective of magnitude of other items.

NOTES

EXERCISE 3.2

1. (*i*) Find the G.M. of the following individual series :

   15, 1.5, 1500, 0.0015.

   (*ii*) Find the G.M. of the following series :

   10, 110, 120, 50, 52, 80, 37, 60.

2. Find the G.M. for the following data relating to the profit of 30 firms :

Profit ('000 ₹)	25	26	27	28	29	30
No. of firms	4	7	12	2	4	1

3. Find the G.M. for the following frequency distribution :

Marks	0—10	10—20	20—30	30—40	40—50
No. of students	4	8	10	6	7

4. Find G.M. from the following :

Marks obtained (below)	10	20	30	40	50
No. of candidates	12	27	72	92	100

5. The G.M. of salaries paid to all employees of a company is ₹ 1700. The G.M. of salaries of male and female employees are ₹ 1800 and ₹ 1600 respectively. Determine the percentage of males and females employed in the company.
6. If the price of a commodity is doubled in five years, find out the annual average rate of increase.
7. The population of a town increased from 10000 to 20000 in 20 years. Find the annual average rate of growth.
8. Find the weighted G.M. of the items 15, 17, 19, 23, 29 with weights 1, 2, 1, 3, 1 respectively.

Answers

1. (*i*) 2.668	(*ii*) 52.84	**2.** ₹ 26.9 thousand
3. 22.06 marks	**4.** 21.4 marks	**5.** Males = 51.37%, Females = 48.63%
6. 14.9%	**7.** 3.5%	**8.** 20.32.

III. HARMONIC MEAN (H.M.)

3.7. DEFINITION OF HARMONIC MEAN

The **harmonic mean** of a statistical data is defined as the quotient of the number of items by the sum of the reciprocals of all the values of the variable.

(*a*) **For an individual series**, the H.M. is given by

$$\text{H.M.} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{\mathbf{n}}{\sum \frac{\mathbf{1}}{\mathbf{x}}},$$

where $x_1, x_2, ..., x_n$ are the values of the variable, under consideration.

(*b*) **For a frequency distribution,**

$$\text{H.M.} = \frac{f_1 + f_2 + \ldots.. + f_n}{f_1\left(\frac{1}{x_1}\right) + f_2\left(\frac{1}{x_2}\right) + \ldots.. + f_n\left(\frac{1}{x_n}\right)} = \frac{\Sigma f}{\sum f\left(\frac{1}{x}\right)} = \frac{\mathbf{N}}{\sum\left(\frac{\mathbf{f}}{\mathbf{x}}\right)},$$

where f_i is the frequency of x_i $(1 \le i \le n)$.

When the values of the variable are given in the form of classes, then the mid-points of classes are taken as the values of the variable (x).

WORKING RULES TO FIND H.M.

Rule I. *In case of an individual series, first find the sum of the reciprocals of all the items. In the second step, divide n, the total number of items by this sum of reciprocals. This gives the value of the H.M.*

Rule II. *In case of a frequency distribution, find the quotients (f/x) of frequencies by the value of items. In the second step, find the sum (Σ(f/x)) of these quotients. Divide N, the total of all frequencies by this sum of quotients. This gives the value of the H.M.*

Rule III. *If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.*

Example 15. *Calculate the H.M. for the following individual series :*

x	*4*	*7*	*10*	*12*	*19*

Solution. **Calculation of H.M.**

S. No.	*x*	*1/x*
1	4	0.2500
2	7	0.1429
3	10	0.1000
4	12	0:0833
5	19	0.0526
$n = 5$		$\sum\left(\frac{1}{x}\right) = 0.6288$

Now $$\text{H.M.} = \frac{n}{\sum\left(\frac{1}{x}\right)} = \frac{5}{0.6288} = \mathbf{7.9516.}$$

Example 16. *Find the H.M for the following frequency distribution :*

Profit ('000 ₹)	*12*	*13*	*14*	*15*	*16*	*17*
No. of firms	*4*	*8*	*6*	*5*	*9*	*2*

NOTES

Solution. Calculation of H.M.

Profit ('000 ₹) x	No. of firms f	$\frac{f}{x}$
12	4	0.3333
13	8	0.6154
14	6	0.4286
15	5	0.3333
16	9	0.5625
17	2	0.1176
	N = 34	$\sum\left(\frac{f}{x}\right) = 2.3907$

Now $\text{H.M.} = \dfrac{N}{\sum\left(\frac{f}{x}\right)} = \dfrac{34}{2.3907}$ = ₹ 14.22178 thousand = ₹ **14221.78.**

3.8. H.M. OF COMBINED GROUP

Theorem: If H_1 and H_2 are the H.M. of two groups having n_1 and n_2 items, then the H.M. of the combined group is given by

$$\mathbf{H = \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}}}.$$

Proof. Let $x_1, x_2, ..., x_{n_1}$ and $y_1, y_2, ..., y_{n_2}$ be the items in the two groups respectively.

$$\therefore \qquad H_1 = \frac{n_1}{\sum \frac{1}{x}}, \qquad H_2 = \frac{n_2}{\sum \frac{1}{y}}$$

$$\therefore \qquad \sum \frac{1}{x} = \frac{n_1}{H_1}, \qquad \sum \frac{1}{y} = \frac{n_2}{H_2},$$

Now
$$H = \frac{\text{no. of items in both groups}}{\text{sum of reciprocals of all the items in both groups}}$$

$$= \frac{n_1 + n_2}{\sum \frac{1}{x} + \sum \frac{1}{y}} \qquad \therefore \ \mathbf{H = \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}}}.$$

This formula can also be extended to more than two groups.

Example 17. *The H.M. of two groups containing 10 and 12 items are found to be 29 and 35. Find the H.M. of the combined group.*

Solution. Here $n_1 = 10, \quad n_2 = 12$

$H_1 = 29, \quad H_2 = 35$

Let H be the H.M. of the combined group

$$\therefore \qquad H = \frac{n_1 + n_2}{\frac{n_1}{H_1} + \frac{n_2}{H_2}} = \frac{10 + 12}{\frac{10}{29} + \frac{12}{35}} = \frac{22}{0.3448 + 0.3429} = \frac{22}{0.6877} = \mathbf{31.9907.}$$

3.9. WEIGHTED H.M.

If all the values of the variable are not of equal importance or in other words, these are of varying importance, then we calculate **weighted H.M.**

$$\text{Weighted H.M.} = \frac{\sum w}{\sum \left(\frac{w}{x}\right)}$$

where $w_1, w_2, \ldots\ldots, w_n$ are the weights of the values $x_1, x_2, \ldots\ldots, x_n$ of the variable, under consideration.

Example 18. *Find the weighted H.M. of the items 4, 7, 12, 19, 25 with weights 1, 2, 1, 1, 1 respectively.*

Solution. **Calculation of weighted H.M.**

x	w	w/x
4	1	0.2500
7	2	0.2857
12	1	0.0833
19	1	0.0526
25	1	0.0400
	$\sum w = 6$	$\sum \left(\frac{w}{x}\right) = 0.7116$

$$\text{Now weighted H.M.} = \frac{\sum w}{\sum \left(\frac{w}{x}\right)} = \frac{6}{0.7116} = \mathbf{8.4317.}$$

3.9.1 Merits of H.M.

1. It is well-defined.
2. It is based on all the items.
3. It is capable of further algebraic treatment.
4. It has sampling stability.
5. It is specially used in finding the average speed, when the distances covered at different speeds are equal or unequal.

3.9.2 Demerits of H.M.

1. It is not simple to understand.
2. It is not easy to compute.
3. It gives higher weightage to smaller items, which may not be desirable in some problems.

NOTES

EXERCISE 3.3

1. Find the H.M. for the following series :

   2, 4, 7, 12, 19.

2. Find the H.M. for the following series :

   15, 20, 21, 22, 26, 29.

3. Calculate (*a*) the Arithmetic Mean, (*b*) the Geometric Mean and (*c*) the Harmonic Mean for the following incomes :

   10, 17, 29, 95, 95, 100, 175, 250, 750.

4. Find the H.M. for the following frequency distribution :

x	10	11	12	13	14	15
f	4	7	6	2	2	1

5. The following table gives the marks (out of 30) obtained by a group of students in a test. Calculate the harmonic mean of this series :

Marks	20	21	22	23	24	25
No. of students	4	2	7	1	3	1

6. Find the H.M. for the following frequency distribution :

Class	2—4	4—6	6—8	8—10
Frequency	20	40	30	10

7. Find the H.M. for the following data :

Class	0—7	7—14	14—21	21—28	28—35	35—42
f	2	5	8	8	5	2

Answers

1. 4.86 **2.** 22.2251 **3.** (*a*) 169, (*b*) 80.74, (*c*) 38.2328

4. 11.5803 **5.** 21.9 marks **6.** 4.98 **7.** 14.6917

IV. MEDIAN

3.10. DEFINITION OF MEDIAN

The **median** of a statistical series is defined as the size of the middle most item (or the A.M. of two middle most items), provided the items are in order of magnitude. For example, the median for the series 4, 6, 10, 12, 18 is 10 and for the series 4, 6, 10, 12, 18, 22, the value of median would be $\frac{10+12}{2}$ = 11. It can be observed that 50% items in the series would have value less than or equal to median and 50% items would be with value greater or equal to the value of the median.

For an individual series, the median is given by,

$$\textbf{Median} = \textbf{size of } \frac{n+1}{2} \textbf{th item}$$

where $x_1, x_2, ..., x_n$ are the values of the variable under consideration. The values x_1, $x_2, ..., x_n$ are supposed to have been arranged in order of magnitude. If $\frac{n+1}{2}$ comes out to be in decimal, then we take median as the A.M. of size of $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$th items.

WORKING RULES FOR FINDING MEDIAN FOR AN INDIVIDUAL SERIES

Step I. *Arrange the given items in order of magnitude.*

Step II. *Find the total number 'n' of items.*

Step III. *Write : median = size of* $\frac{n+1}{2}$*th item.*

Step IV. (*i*) *If* $\frac{n+1}{2}$ *is a whole number, then* $\frac{n+1}{2}$*th item gives the value of median.*

(*ii*) *If* $\frac{n+1}{2}$ *is in friction, then the A.M. of* $\frac{n}{2}$*th and* $\left(\frac{n}{2}+1\right)$*th items gives the value of median.*

For a frequency distribution, in which frequencies (*f*) of different values (*x*) of the variable are given, we have

$$\textbf{Median} = \textbf{size of } \frac{N+1}{2} \textbf{ th item.}$$

Remark. The values of the variable are supposed to have been arranged in order of magnitude.

WORKING RULES FOR FINDING MEDIAN FOR A FREQUENCY DISTRIBUTION

Step I. *Arrange the values of the variable in order of magnitude and find the cumulative frequencies (c.f.).*

Step II. *Find the total 'N' of all frequencies and check that it is equal to the last c.f.*

Step III. *Write : median = size of* $\frac{N+1}{2}$*th item.*

Step IV. (*a*) *If* $\frac{N+1}{2}$ *is a whole number, then* $\frac{N+1}{2}$*th item gives the value of median. For this, look at the cumulative frequency column and find that total which is either equal to* $\frac{N+1}{2}$ *or the next higher than*

NOTES

$\frac{N+1}{2}$ *and determine the value of the variable corresponding to this. This gives the value of median.*

(b) *If* $\frac{N+1}{2}$ *is in friction, then the A.M. of* $\frac{N}{2}$*th and* $\left(\frac{N}{2}+1\right)$*th items gives the value of median.*

In case, the values of the variable are given in the form of classes, we shall assume that items in the classes are uniformly distributed in the corresponding classes. We define

$$\textbf{Median} = \textbf{size of } \frac{N}{2} \textbf{th item.}$$

Here we shall get the class in which N/2th item is present. This is called the **median class**. To ascertain the value of median in the median class, the following formula is used.

$$\textbf{Median} = L + \left(\frac{N/2 - c}{f}\right) h$$

where L = lower limit of the median class

c = cumulative frequency of the class preceding the median class

f = simple frequency of the median class

h = width of the median class.

Remark. In problems on **Averages** or in other problems in the following chapters, where we need only the mid values of class intervals in the formula, we need not convert the classes written using *'inclusive method'*.

The following points must be taken care of, while calculating median :

1. The values of the variable must be in order of magnitude. In case of classes of values of the variable, the classes must be strictly *in ascending* order of magnitude.

2. If the classes are in inclusive form, then the actual limits of the median class are to be taken for finding L and h.

3. The classes may not be of equal width *i.e.*, h need not be the common width of all classes. It is the width of the *"median class"*.

4. In case of open end classes, it is advisable to find average by using median.

WORKING RULES FOR FINDING MEDIAN FOR A FREQUENCY DISTRIBUTION WITH CLASS INTERVALS

Step I. *Arrange the classes in the ascending order of magnitude. The classes must be in 'exclusive form'. The widths of classes may not be equal. Find the cumulative frequencies (c.f.).*

Step II. *Find the total 'N' of all frequencies and check that it is equal to the last c.f.*

Step III. *Write : median = size of* $\frac{N}{2}$ *th item.*

Step IV. *Look at the cumulative frequency column and find that total which is either equal to* $\frac{N}{2}$ *or the next higher than* $\frac{N}{2}$ *and determine the class corresponding to this. That gives the 'median class'.*

Step V. *Write : median* $= L + \left(\frac{N/2 - c}{f}\right)$ *h. Put the values of L, N/2, c, f, h and calculate the value of median.*

NOTES

Example 19. *Find the median of the series :*

4, 6, 9, 4, 2, 8, 10.

Solution. The values of the variable arranged in ascending order are

x : 2, 4, 4, 6, 8, 9, 10

Here $n = 7$. $\therefore \quad \frac{n+1}{2} = \frac{7+1}{2} = 4$

$\therefore$ Median = size of 4th item = **6.**

Example 20. *Find the median for the series :*

25, 20, 23, 32, 40, 27, 30, 25, 20, 10, 55, 41.

Solution. The values of the variable arranged in the ascending order are

x : 10, 20, 20, 23, 25, 25, 27, 30, 32, 40, 41, 55

Here $n = 12$

$$\therefore \quad \frac{n+1}{2} = \frac{12+1}{2} = 6.5$$

$\therefore$ Median = 'size of 6.5th' item

$$= \frac{\text{6th item + 7th item}}{2} = \frac{25+27}{2} = \mathbf{26.}$$

Example 21. *Find the median for the following frequency distribution :*

x	*0*	*1*	*2*	*3*	*4*	*5*	*6*
f	*5*	*9*	*10*	*12*	*6*	*4*	*2*

Solution. **Calculation of Median**

x	f	c.f.
0	5	5
1	9	14
2	10	24
3	12	36
4	6	42
5	4	46
6	2	48 = N
	N = 48	

Here, $\quad \frac{N+1}{2} = \frac{48+1}{2} = 24.5$

$\therefore$ Median = size of 24.5th item

$$= \frac{\text{size of 24th item + size of 25th item}}{2} = \frac{2+3}{2} = \mathbf{2.5.}$$

NOTES

Example 22. *Find the median for the following wage distribution in a certain factory :*

Monthly wages (₹)	*50—80*	*80—100*	*100—110*	*110—120*
No. of workers	*30*	*127*	*140*	*240*
Monthly wages (₹)	*120—130*	*130—150*	*150—180*	*180—200*
No. of workers	*176*	*135*	*20*	*3*

Solution. **Calculation of Median**

Monthly wages (₹)	*No. of workers f*	*c. f.*
50—80	30	30
80—100	127	157
100—110	140	$297 = c$
L = 110—120	$240 = f$	537
120—130	176	713
130—150	135	848
150—180	20	868
180—200	3	871 = N
	N = 871	

$$\frac{N}{2} = \frac{871}{2} = 435.5$$

∴ Median = size of 435.5th item

∴ Median class is 110—120.

$$\therefore \quad \text{Median} = L + \left(\frac{N/2 - c}{f}\right)h = 110 + \left(\frac{435.5 - 297}{240}\right)10$$

$$= 110 + 5.77 = ₹\ \mathbf{115.77}.$$

Example 23. *The following table gives the marks obtained by 50 students. Find the median :*

Marks	*10—14*	*15—19*	*20—24*	*25—29*
No. of students	*5*	*8*	*6*	*7*
Marks	*30—34*	*35—39*	*40—44*	*45—49*
No. of students	*6*	*3*	*9*	*6*

Solution. Here the classes are given in 'inclusive form'. These classes in 'exclusive form' are :

9.5—14.5, 14.5—19.5, ..., 44.5—49.5.

Calculation of Median

Marks	*No. of students (f)*	*c.f.*
9.5—14.5	5	5
14.5—19.5	8	13
19.5—24.5	6	19 = *c*
L = 24.5—29.5	7 = *f*	26
29.5—34.5	6	32
34.5—39.5	3	35
39.5—44.5	9	44
44.5—49.5	6	50 = N
	N = 50	

$$\frac{N}{2} = \frac{50}{2} = 25$$

∴ Median = size of 25th item

∴ Median class is 24.5—29.5

$$\therefore \quad \text{Median} = L + \left(\frac{\frac{N}{2} - c}{f}\right)h = 24.5 + \left(\frac{25 - 19}{7}\right)5$$

$$= 24.5 + 4.286 = \mathbf{28.786\ marks.}$$

3.11. MERITS OF MEDIAN

1. It is simple to understand.
2. It is easy to compute.
3. It is well-defined.
4. It is not affected by the extreme items.
5. It is best suited for open end classes.
6. It can also be located graphically.

3.12. DEMERITS OF MEDIAN

1. It is not based on all the items.
2. It is not capable of further algebraic treatment.
3. It can only be calculated when the data is in order of magnitude.

NOTES

EXERCISE 3.4

1. Find the value of the median for the series :

   8, 6, 6, 5, 11, 80, 12.

2. Find the value of median for the following data :

x	1	2	3	4	5	6	7	8	9
f	5	6	7	2	2	1	2	1	1

3. Find the median for the following frequency distribution :

No. of students	6	4	16	7	8	2
Marks	20	9	25	50	40	80

4. Find the median marks for the following distribution :

Marks less than	10	20	30	40	50	60
No. of students	4	10	30	40	47	50

5. Find the value of median for the following series :

Class	1—10	11—20	21—30	31—40	41—50
No. of items	10	21	51	45	26

[**Hint.** The actual limits of classes are :

0.5—10.5, 10.5—20.5, 20.5—30.5, 30.5—40.5 and 40.5—50.5.]

6. Calculate median for the following data :

Mid-value	115	125	135	145	155	165	175	185	195
Frequency	6	25	48	72	116	60	38	22	3

7. Find A.M. and Median for the following frequency distribution:

Marks	*No. of students*
0–7	19
7–14	25
14–21	36
21–28	72
28–35	51
35–42	43
42–49	28

8. An incomplete frequency distribution is given below :

Variable	10—20	20—30	30—40	40—50	50—60	60—70	70—80
Frequency	12	30	?	65	?	25	18

It is given that median value is 46 and the total number of items is 229. You are required to find the missing frequencies.

Answers

1. 8	**2.** 3	**3.** 25 marks	**4.** 27.5 marks
5. 29.4216	**6.** 153.7931	**7.** 26.49, 30.82	**8.** 34, 45

NOTES

V. MODE

3.13. DEFINITION OF MODE

The **mode** of a statistical series is defined as that value of the variable around which the values of the variable tend to be most heavily concentrated. It can also be defined as that value of the variable whose own frequency is dominating and at the same time, the frequencies of its neighbouring items are also dominating. Thus, we see that mode is that value of the variable around which the items of the series cluster densily. Let us consider the data regarding the sale of ready made shirts :

Size (in inches)	30	32	34	36	38	40	42
No. of shirts sold	5	22	24	38	16	8	2

Here we see that the frequency of 36 is highest and the frequencies of its neighbouring items (34, 38) are also dominating. Here the most fashionable, modal size is 36 inches. Technically, we shall say that the mode of the distribution is 36 inches.

In case of mode, we are to deal with the frequencies of values of the items, thus if we are to find the value of mode for an individual series, we will have to see the repetition of different items. *i.e.*, we would be in a way expressing it in the form of frequency distribution. Thus, we start our discussion for evaluating mode for frequency distributions. There are two methods of finding mode of a frequency distribution.

3.14. MODE BY INSPECTION

Sometimes the frequencies in a frequency distribution are so distributed that we would be able to find the value of mode just, by inspection. For example, let us consider the frequency distribution :

x	4	5	6	7	8	9	10	11	12
f	1	2	1	5	12	4	2	2	1

Here we can say, at once, that mode is 8.

3.15. MODE BY GROUPING

Let us consider the distribution :

x	4	5	6	7	8	9	10	11	12
f	4	5	7	14	8	15	2	2	1

NOTES

Here the frequency of 9 is more than the frequency of 7, whereas the frequencies of neighbouring items of 7 are more than that for 9. In such a case, we would not be able to judge the value of mode just by inspecting the data. In case there is even slight doubt as to which is the value of mode, we go for this method. In this method, two tables are drawn. These tables are called 'Grouping Table' and 'Analysis Table'. In the grouping table, six columns are drawn. The column of frequencies is taken as the column I. In the column II, the sum of two frequencies are taken at a time. In the column III, we exclude the first frequency and take the sum of two frequencies at a time. In the column IV, we take the sum of three frequencies at a time. In the column V, we exclude the first frequency and take the sum of frequencies, taking three at a time. In the last column, we exclude the first two frequencies and take the sum of three frequencies at a time. The next step is to mark the maximum sums in each of the six columns.

In the analysis table, six rows are drawn corresponding to each column in the grouping table. In this table, columns are made for those values of the variable whose frequencies accounts for giving maximum totals in the columns of the grouping table. In this table, marks are given to the values of the variable as often as their frequencies are added to make the total maximum.in the columns of the grouping table. The value of the variable which get the maximum marks is declared to be the mode of the distribution.

In case, the values of the variable are given in the form of classes, we shall assume that the items in the classes are uniformly distributed in the corresponding classes. Here we shall get a 'class' either by the method of inspection or the method of grouping. This class is called the **modal class**. To ascertain the value of mode in the modal class, the following formula is used.

$$\textbf{Mode} = \text{L} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \textbf{h}$$

where L = lower limit of modal class

Δ_1 = difference of frequencies of modal class and pre-modal class

Δ_2 = difference of frequencies of modal class and post-modal class

h = width of the modal class.

The following points must be taken care of while calculating mode :

1. The values (or classes of values) of the variable must be in ascending order of magnitude.

2. If the classes are in inclusive form, then the actual limits of the modal class are to be taken for finding L and h.

3. The classes must be of equal width.

It may be noted that while analysing the analysis table, we may find two or more values (or classes of values) of the variable getting equal marks. In such a case, the grouping method fails. Such distribution is called a **multi-modal distribution**.

3.16. EMPIRICAL MODE

In case of a multi-modal distribution, we find the value of mode by using the relation

$$\textbf{Mode = 3 Median − 2 A.M.}$$

This mode is called **empirical mode** in the sense that this relation cannot be established algebraically. But it is generally observed that in distributions, the value of mode is approximately equal to 3 Median – 2 A.M. That is why, this mode is called *empirical mode.*

WORKING RULES FOR FINDING MODE

Step I. *If mode is not evident by the 'method of inspection', then the 'method of grouping' should be used.*

Step II. *In case, the values of variable are given in terms of classes of equal width, then* ***step I****, will give the 'modal class'.*

Step III. *To find value of the mode, use the formula :*

$$\text{mode} = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) h.$$

Step IV. *In case, the distribution is multimodal, then find the value of mode by using the formula : 'mode = 3 median – 2 A.M'.*

Example 24. *Find the mode for the following individual series :*

5, 7, 3, 5, 2, 1, 5, 8, 5.

Solution. **Calculation of Mode**

x	*Tally bars*	*Frequency (f)*
1	|	1
2	|	1
3	|	1
5	||||	4
7	|	1
8	|	1

By inspection, we can say that mode is **5.**

Example 25. *Find the mode for the following distribution :*

Profit ('000 ₹)	*28*	*29*	*30*	*31*	*32*	*33*
No. of firms	*4*	*7*	*10*	*6*	*2*	*1*

Solution. **Calculation of Mode**

Profit ('000 ₹) x	*No. of firms* f
28	4
29	7
30	10
31	6
32	2
33	1

By inspection we can say that mode is ₹ **30,000**. This is so because the frequency of 30,000 is very high as compared with the frequencies of other values of x. Moreover, the frequencies of the neighbouring items are also dominating.

NOTES

Example 26. *Find the value of mode for the following distribution :*

x	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
f	*15*	*18*	*12*	*30*	*27*	*40*	*20*	*20*	*12*

Solution. For the given distribution, we cannot judge the value of mode, just by inspection. In this case, we shall apply the method of grouping. The values of the variable are already in order of magnitude.

Grouping Table

	I	*II*	*III*	*IV*	*V*	*VI*
x	*f*					
4	15	33		45		
5	18		30		60	
6	12	42				69
7	30		57	97		
8	27	**67**			87	
9	**40**		**60**			**80**
10	20	40		52		
11	20		32			
12	12					

Analysis Table

Column	*9*	*8*	*10*	*7*	*11*
I	1				
II	1	1			
III	1		1		
IV	1	1		1	
V	1	1	1		
VI	1		1		1
Total	6	3	3	1	1

∴ Mode = **9**.

Example 27. *Calculate the value of mode for the following frequency distribution :*

Class	*Frequency*	*Class*	*Frequency*
1—4	*2*	*21—24*	*14*
5—8	*5*	*25—28*	*14*
9—12	*8*	*29—32*	*15*
13—16	*9*	*33—36*	*11*
17—20	*12*	*37—40*	*13*

Solution. In this problem, we shall make use of 'grouping method' to find the modal class.

Grouping Table

Class	*Frequency* *I*	*II*	*III*	*IV*	*V*	*VI*
1—4	2	7		15		
5—8	5		13			
9—12	8	17			22	
13—16	9		21	35		
17—20	12	26				29
21—24	14		**28**		**40**	
25—28	14	**29**		**40**		
29—32	**15**		26			**43**
33—36	11	24			39	
37—40	13					

Analysis Table

Column	*29—32*	*25—28*	*21—24*	*33—36*	*17—20*
I	1				
II	1	1			
III		1	1		
IV	1	1		1	
V		1	1		1
VI	1	1	1		
Total	4	5	3	1	1

∴ Modal class is 25—28.

Now $\quad \text{mode} = \text{L} + \left(\dfrac{\Delta_1}{\Delta_1 + \Delta_2}\right) h$

Here $\quad \text{L} = 24.5,\ \Delta_1 = 14 - 14 = 0,\ \Delta_2 = 15 - 14 = 1,\ h = 28.5 - 24.5 = 4$

∴ $\quad \text{Mode} = 24.5 + \left(\dfrac{0}{0+1}\right) 4 = 24.5 + 0 = \mathbf{24.5.}$

Example 28. *If a, b are positive numbers, then show that*

(*i*) *A.M.* ≥ *G.M.* ≥ *H.M.* (*ii*) $G.M. = \sqrt{A.M. \times H.M.}$

Solution We have $\text{A.M.} = \dfrac{a+b}{2}$, $\text{G.M.} = \sqrt{ab}$, $\text{H.M.} = \dfrac{2}{\dfrac{1}{a}+\dfrac{1}{b}} = \dfrac{2ab}{a+b}$.

(*i*) $$\text{A.M.} - \text{G.M.} = \frac{a+b}{2} - \sqrt{ab} = \frac{a+b-2\sqrt{ab}}{2}$$

$$= \frac{(\sqrt{a})^2 + (\sqrt{b})^2 - 2\sqrt{a}\sqrt{b}}{2} = \frac{(\sqrt{a}-\sqrt{b})^2}{2} \geq 0$$

∴ $\quad \text{A.M.} - \text{G.M.} \geq 0$ *i.e.*, $\text{A.M.} \geq \text{G.M.}$ ...(1)

NOTES

$$\text{G.M.} - \text{H.M.} = \sqrt{ab} - \frac{2ab}{a+b} = \frac{\sqrt{ab}\,(a+b-2\sqrt{ab})}{a+b}$$

$$= \frac{\sqrt{ab}\,((\sqrt{a})^2 + (\sqrt{b})^2 - 2\sqrt{a}\sqrt{b})}{a+b} = \frac{\sqrt{ab}\,(\sqrt{a}-\sqrt{b})^2}{a+b} \geq 0.$$

$\therefore$ G.M. – H.M. ≥ 0 *i.e.,* G.M. ≥ H.M. ...(2)

Combining (1) and (2), we have

A.M. ≥ G.M. ≥ H.M.

In particular if $a = b$, then

$$\text{A.M.} = \frac{a+b}{2} = \frac{a+a}{2} = a, \text{ G.M.} = \sqrt{ab} = \sqrt{aa} = a$$

and
$$\text{H.M.} = \frac{2ab}{a+b} = \frac{2aa}{a+a} = \frac{2a^2}{2a} = a$$

i.e., **A.M. = G.M. = H.M.**

(*ii*) $\sqrt{\text{A.M.} \times \text{H.M.}} = \sqrt{\frac{a+b}{2} \times \frac{2ab}{a+b}} = ab$ and G.M. $= (\sqrt{ab})^2 = ab$

$\therefore$ **G.M.** $= \sqrt{\textbf{A.M.} \times \textbf{H.M.}}$

3.17. MODE IN CASE OF CLASSES OF UNEQUAL WIDTHS

When the values of the variable are given in the form of classes and the classes are not of equal width, then we would not be able to proceed directly to find the modal class either by the method of inspection or by the method of grouping. In fact, we are to compare the frequencies of different classes in order to observe the concentration of items about some item. If the classes happen to be of unequal width, then we would not be able to compare the frequencies in different classes. To make the comparison meaningful, we will first make classes of equal width by grouping two or more classes or by breaking classes, as per the need.

Example 29. *Calculate mode for the following distribution :*

Marks	*No. of students*	*Marks*	*No. of students*
0—2	*8*	*25—30*	*45*
2—4	*12*	*30—40*	*60*
4—10	*20*	*40—50*	*20*
10—15	*10*	*50—60*	*13*
15—20	*16*	*60—80*	*15*
20—25	*25*	*80—100*	*4*

Solution. In this frequency distribution, the classes are not of equal width. Before ascertaining the modal class, we shall make the widths of classes equal.

Marks	*No. of students*
0—20	8 + 12 + 20 + 10 + 16 = 66
20—40	25 + 45 + 60 = 130
40—60	20 + 13 = 33
60—80	15
80—100	4

NOTES

By inspection, modal class is 20—40.

Now $$\text{mode} = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) h$$

Here $L = 20,\ \Delta_1 = 130 - 66 = 64,\ \Delta_2 = 130 - 33 = 97,\ h = 20$

$$\therefore \qquad \text{Mode} = 20 + \left(\frac{64}{64 + 97}\right) 20 = 20 + 7.9503 = \mathbf{27.9503.}$$

3.18. MERITS OF MODE

1. It is easy to compute.
2. It is not affected by the extreme items.
3. It can be located graphically.

3.19. DEMERITS OF MODE

1. It is not simple to understand.

2. It is not well defined. There are number of formulae to calculate mode, not necessarily giving the same answer.

3. It is not capable of further algebraic treatment.

EXERCISE 3.5

1. Find the value of mode for the following series :

10, 12, 17, 12, 10, 12, 16, 11.

2. Find the value of mode for the following frequency distribution :

x	10	11	12	13	14	15	16	17	18
f	2	4	6	8	10	9	6	2	1

3. Find the mode for the following frequency distribution :

x	1	2	3	4	5	6	7	8	9	10	11	12
f	3	8	15	23	35	40	32	28	20	45	14	6

4. Find the mode for the following frequency distribution :

Class	0—4	4—8	8—12	12—16
f	4	8	5	6

5. The following table gives the weight of 50 students of a class. Find the modal weight :

Weight (in kg)	37—41	42—46	47—51	52—56	57—61	62—66	67—71
No.of students	3	7	11	14	7	6	2

6. Following table shows the marks obtained by 60 students. Calculate (*i*) median (*ii*) mode and (*iii*) arithmetic average.

Marks		*No. of students*
More than	70%	8
,,	60%	18
,,	50%	40
,,	40%	40
,,	30%	55
,,	20%	60

7. Calculate median and arithmetic average for the following data. Also calculate mode with the help of median and A.M.

Variable	100—110	110—120	120—130	130—140
Frequency	4	6	20	32
Variable	150—160	160—170	170—180	
Frequency	33	8	2	

8. The following table shows the distribution of 100 families according to their expenditure per week. Number of families corresponding to the groups ₹ 1000–2000 and ₹ 3000–4000 are missing. The mode is given to be ₹ 2,400. Calculate the missing frequencies :

Expenditure (₹)	0–1000	1000–2000	2000–3000	3000–4000	4000–5000
No. of families	14	?	27	?	15

[**Hint.** Let the frequencies of the classes 1000—2000 and 3000—4000 be a and b respectively.

$$\therefore \qquad 2400 = 2000 + \left(\frac{27-a}{(27-a)+(27-b)}\right)1000 \text{ i.e., } 400 = \frac{(27-a)\,100}{54-a-b}$$

Also $\qquad a + b = 44.$]

9. Find the mode for the following data :

Size of items	*Frequency*	*Size of items*	*Frequency*
0—4	5	20—24	14
4—8	7	24—28	6
8—12	9	28—32	3
12—16	17	32—36	1
16—20	15	36—40	0

10. Calculate media and mode from the following table:

Income	100–200	100–300	100–100	100–300	100–600
No. of persons	15	33	63	53	100

Answers

1. 12 **2.** 14 **3.** 6 **4.** 6.286

5. 53 **6.** (*i*) 54.54% marks (*ii*) 56.67% marks (*iii*) 51.67% marks

7. 137.03, 140.14, 130.81. **8.** 23, 21 **9.** 18.67

10. 356.67, 354.55

EXERCISE 3.7

1. What is meant by 'Central Tendency' ? Describe the various methods of measuring it and point out the usefulness of each method.
2. Describe the merits and demerits of arithmetic mean, median and mode.
3. What is the relationship among mode, median and arithmetic average in a symmetrical series ?
4. What purpose is served by an average ? Discuss the relative merits and shortcomings of various types of statistical averages.
5. Define geometric mean, for individual series and frequency distribution and give their computational formulae.

3.20. SUMMARY

- This is the most popular and widely used measure of central tendency. The popularity of this average can be judged from the fact that it is generally referred to as 'mean'. The **arithmetic mean** of a statistical data is defined as the quotient of the sum of all the values of the variable by the total number of items and is generally denoted by $\overline{x}$.
- If all the values of the variable are not of equal importance, or in other words, these are of varying significance, then we calculate **weighted A.M.**
- Geometric mean is specially used to find the average rate of increase or decrease in sale, production, population etc.
- The **harmonic mean** of a statistical data is defined as the quotient of the number of items by the sum of the reciprocals of all the values of the variable.
- The **mode** of a statistical series is defined as that value of the variable around which the values of the variable tend to be most heavily concentrated. It can also be defined as that value of the variable whose own frequency is dominating and at the same time, the frequencies of its neighbouring items are also dominating.
- In case, the values of the variable are given in the form of classes, we shall assume that the items in the classes are uniformly distributed in the corresponding classes. Here we shall get a 'class' either by the method of inspection or the method of grouping. This class is called the **modal class**.
- It may be noted that while analysing the analysis table, we may find two or more values (or classes of values) of the variable getting equal marks. In such a case, the grouping method fails. Such distribution is called a **multi-modal distribution**.
- It is generally observed that in distributions, the value of mode is approximately equal to 3 Median – 2 A.M. That is why, this mode is called *empirical mode.*

NOTES

4. MEASURES OF DISPERSION

STRUCTURE

4.1. REQUISITES OF A GOOD MEASURE OF DISPERSION

The requisites of a good measure of a dispersion are the same as those for a good measure of central tendency. For the sake of completeness, we list the requisites as under :

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be well-defined.
4. It should be based on all the items.
5. It should not be unduly affected by the extreme items.
6. It should be capable of further algebraic treatment.
7. It should have sampling stability.

4.2. METHODS OF MEASURING DISPERSION

I. Range
II. Quartile Deviation (Q.D.)
III. Mean Deviation (M.D.)
IV. Standard Deviation (S.D.)
V. Lorenz Curve.

I. RANGE

4.3. DEFINITION OF RANGE

The **range** of a statistical data is defined as the difference between the largest and the smallest values of the variable.

$\therefore$ **Range = L – S,**

where L = largest value of the variable

S = smallest value of the variable.

In case, the values of the variable are given in the form of classes, then L is taken as the upper limit of the largest value class and S as the lower limit of the smallest value class.

Example 1. *Find the range of the series :*

4, 2, 6, 8, 10.

Solution. Here L = 10, S = 2.

$\therefore$ Range = L – S = 10 – 2 = **8.**

Example 2. *Find the range of the following distribution :*

Age (in years)	*16—18*	*18—20*	*20—22*	*22—24*	*24—26*	*26—28*
No. of students	*0*	*4*	*6*	*8*	*2*	*2*

Solution. Here L = 28, S = 18

$\therefore$ Range = L – S = 28 – 18 = **10 years.**

It may be noted that S ≠ 16, though it is the lower limit of the smallest value class, but there is no item in this class and so this class is meaningless so far as the calculation of range is concerned.

4.3.1 Merits of Range

1. It is simple to understand.

2. It is easy to compute.

3. It is well-defined.

4. It helps in giving an idea about the variation, just by giving the lowest value and the greatest value of variable.

4.3.2 Demerits of Range

NOTES

1. It is not based on all the items.

2. It is highly affected by the extreme items. In fact, if extreme items are present, then range would be calculated by taking only extreme items.

3. It does not take into account the frequencies of items in the middle of the series.

4. It is not capable of further algebraic treatment.

5. It does not have sampling stability.

EXERCISE 4.1

1. Calculate the range for the following series :

17, 10, 12, 8, 12, 16, 19.

2. Find the range for the following data :

Profit ('000 ₹)	0—10	10—20	20—30	30—40	40—50
No. of firms	0	6	0	7	15

3. Find the coefficient of range for the following data :

Marks less than	10	20	30	40	50	60	70
No. of students	0	4	7	10	14	18	20

Answers

1. 11 **2.** ₹ 40,000 **3.** 0.75

II. QUARTILE DEVIATION (Q.D.)

4.4. INADEQUACY OF RANGE

Consider the series

I : 4, 4, 4, 5, 5, 6, 4, 5, 5, 1000.

II : 4, 4, 4, 5, 5, 6, 4, 5, 5.

$$\text{For series I, Coeff. of Range} = \frac{1000-4}{1000+4} = \frac{996}{1004} = 0.992$$

$$\text{For series II, Coeff. of Range} = \frac{6-4}{6+4} = \frac{2}{10} = 0.200.$$

On comparing the values of coeff. of range for these series, one is likely to conclude that these is marked difference in variability in the series. In fact, the series II is obtained from the series I, just by ignoring the extreme item 1000. Thus, we see that extreme items can distort the value of range and even the coefficient of range. If we have a glance at the definitions of these measures, we would find that only extreme items are required in their calculation, if at all extreme items are present. Even if extreme items are present in a series, the middle 50% values of the variable would be expected to vary quite smoothly, keeping this in view, we define another measure of dispersion, called 'Quartile Deviation'.

NOTES

4.5. DEFINITION OF QUARTILE DEVIATION

The **quartile deviation** of a statistical data is defined as

$$\frac{Q_3 - Q_1}{2} \text{ and is denoted as Q.D.}$$

This is also called *semi-inter quartile* range. We have already studied the method of calculating quartiles. The value of Q.D. is obtained by subtracting Q_1 from Q_3 and then dividing it by 2.

For comparing two or more series for variability, the absolute measure Q.D. would not work. For this purpose, the corresponding relative measure, called coeff. of Q.D. is calculated. This is defined as :

$$\textbf{Coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Example 3. *Find Q.D. and its coefficient for the following series :*

x (in ₹) : 4, 7, 6, 5, 9, 12, 19.

Solution. The values of the variable arranged in ascending order are

x (in ₹) : 4, 5, 6, 7, 9, 12, 19.

Here $n = 7$.

$$\mathbf{Q_1} : \quad \frac{n+1}{4} = \frac{7+1}{4} = 2 \qquad \therefore \quad Q_1 = \text{size of 2nd item} = ₹\ 5$$

$$\mathbf{Q_3} : \quad 3\left(\frac{n+1}{4}\right) = 3\left(\frac{7+1}{4}\right) = 6 \qquad \therefore \quad Q_3 = \text{size of 6th item} = ₹\ 12$$

$$\therefore \quad \text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{12-5}{2} = ₹\ \mathbf{3.5.}$$

$$\text{Coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{12-5}{12+5} = \frac{7}{17} = \mathbf{0.4118.}$$

Example 4. *Find the quartile deviation for the following distribution :*

Marks	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
No. of students	*10*	*11*	*12*	*13*	*5*	*12*	*7*	*5*

NOTES

Solution. **Calculation of Quartiles**

Marks	*No. of students* *f*	*c.f.*
2	10	10
3	11	21
4	12	33
5	13	46
6	5	51
7	12	63
8	7	70
9	5	75 = N
	N = 75	

Q_1 : $\frac{N+1}{4} = \frac{75+1}{4} = 19$ ∴ Q_1 = size of 19th item = 3 marks

Q_3 : $3\left(\frac{N+1}{4}\right) = 3\left(\frac{75+1}{4}\right) = 57$ ∴ Q_3 = size of 57th item = 7 marks

∴ Q.D. $= \frac{Q_3 - Q_1}{2} = \frac{7-3}{2} =$ **2 marks.**

Example 5. *Find the coeff. of Q.D. for the following distribution :*

Marks	*0—4*	*4—8*	*8—12*	*12—14*
No. of students	*10*	*12*	*18*	*7*
Marks	*14—18*	*18—20*	*20—25*	*25 and above*
No. of students	*5*	*8*	*4*	*6*

Solution. **Calculation of Quartiles**

Marks	*No. of students* *f*	*c.f.*
0—4	10	10
4—8	12	22
8—12	18	40
12—14	7	47
14—18	5	52
18—20	8	60
20—25	4	64
25 and above	6	70 = N
	N = 70	

Q_1 : $\frac{N}{4} = \frac{70}{4} = 17.5$ ∴ Q_1 = size of 17.5th item

∴ Q_1 class is 4—8.

$$\therefore \quad Q_1 = L + \left(\frac{N/4 - c}{f}\right)h = 4 + \left(\frac{17.5 - 10}{12}\right)4 = 4 + 2.5 = 6.5 \text{ marks.}$$

$$Q_3 : \quad 3\left(\frac{N}{4}\right) = 3\left(\frac{70}{4}\right) = 52.5 \qquad \therefore \quad Q_3 = \text{size of 52.5th item}$$

$\therefore$ Q_3 class is 18—20.

$$\therefore \quad Q_3 = L + \left(\frac{3N/4 - c}{f}\right)h = 18 + \left(\frac{52.5 - 52}{8}\right)2$$

$$= 18 + 0.125 = 18.125 \text{ marks.}$$

$$\text{Coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{18.125 - 6.5}{18.125 + 6.5} = \frac{11.625}{24.625} = \mathbf{0.4721.}$$

4.5.1 Merits of Q.D.

1. It is simple to understand.
2. It is easy to calculate.
3. It is well-defined.
4. It helps in studying the middle 50% items in the series.
5. It is not affected by the extreme items.
6. It is useful in the case of open end classes.

4.5.2 Demerits of Q.D.

1. It is not based on all the items.
2. It is not capable of further algebraic treatment.
3. It does not have sampling stability.

EXERCISE 4.2

1. Find the Q.D. and its coefficient for the given data regarding the age of 7 students.

Age (in years) : 17, 19, 22, 26, 19, 28, 17.

2. Find the coefficient of Q.D. for the following frequency distribution :

Age (in years)	15	16	17	18	19	20
No. of students	241	500	600	550	700	750

3. For the following data, calculate Q.D. and its coefficient :

Class	10—20	20—30	30—40	40—50	50—60	60—70
Frequency	3	5	15	10	4	2

4. Calculate Quartile Deviation and its coefficient for the data given below :

Daily wages (in ₹)	1—5	6—10	11—15	16—20
No. of workers	3	8	14	11
Daily wages (in ₹)	21—25	26—30	31—35	36—40
No. of workers	7	6	5	2

NOTES

Answers

1. Q.D. = 4.5 years, Coeff. of Q.D. = 0.2093
2. 0.0555
3. Q.D. = 7.5416, Coeff. of Q.D. = 0.1948
4. Q.D. = 6.6072, Coeff. of Q.D. = 0.3635

III. MEAN DEVIATION (M.D.)

4.6. DEFINITION OF MEAN DEVIATION

Mean deviation is also called **average deviation**. The **mean deviation** of a statistical data is defined as the arithmetic mean of the numerical values of the deviations of items from some average. Generally, A.M. and median are used in calculating mean deviation. Let '*a*' stand for the average used for calculating M.D.

For an **individual series**, the M.D. is given by

$$\text{M.D.} = \frac{\sum_{i=1}^{n} |x_i - a|}{n} = \frac{\Sigma|x-a|}{n}$$

where $x_1, x_2, \ldots\ldots, x_n$ are the values of the variable, under consideration.

For a **frequency distribution,**

$$\text{M.D.} = \frac{\sum_{i=1}^{n} f_i |x_i - a|}{N} = \frac{\Sigma f|x-a|}{N}$$

where f_i is the frequency of x_i $(1 \le i \le n)$.

When the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

Median is used in calculating M.D., because of its property that the sum of numerical values of deviations of items from median is always least. So, if median is used in the calculation of M.D., its value would come out to be least. M.D. is also calculated by using A.M. because of its simplicity and popularity. In problems, it is generally given as to which average is to be used in the calculation of M.D. If it is not given, then either of the two can be made use of.

4.7. COEFFICIENT OF M.D.

For comparing two or more series for variability, the corresponding relative measure, 'Coefficient of M.D.', is used. This is defined as :

$$\textbf{Coeff. of M.D.} = \frac{\textbf{M.D.}}{\textbf{Average}}.$$

If M.D. is calculated about A.M., then M.D. is written as M.D.$(\bar{x})$. Similarly, M.D.(Median) would mean that median has been used in calculating M.D.

∴ We can write

$$\textbf{Coeff. of M.D.}(\overline{\mathbf{x}}) = \frac{\textbf{M.D.}(\overline{\mathbf{x}})}{\overline{\mathbf{x}}}$$

$$\textbf{Coeff. of M.D.(Median)} = \frac{\textbf{M.D.(Median)}}{\textbf{Median}}$$

WORKING RULES TO FIND M.D. ($\overline{\mathbf{x}}$)

Rule I. *In case of an individual series, first find $\overline{x}$ by using the formula $\overline{x} = \frac{\Sigma x}{n}$. In the second step, find the values of $x - \overline{x}$. In the next step, find the numerical values $|x - \overline{x}|$ of $x - \overline{x}$. Find the sum $\Sigma|x - \overline{x}|$ of these numerical values $|x - \overline{x}|$. Divide this sum by n to get the value of M.D.$(\overline{x})$.*

Rule II. *In case of a frequency distribution, first find $\overline{x}$ by using the formula $\overline{x} = \frac{\Sigma fx}{N}$. In the second step, find the values of $x - \overline{x}$. In the next step, find the numerical values $|x - \overline{x}|$ of $x - \overline{x}$. Find the products of the values of $|x - \overline{x}|$ and their corresponding frequencies. Find the sum $\Sigma f|x - \overline{x}|$ of these products. Divide this sum by N to get the value of M.D.$(\overline{x})$.*

Rule III. *If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.*

Rule IV. *To find the coefficient of M.D.$(\overline{x})$, divide M.D.$(\overline{x})$ by $\overline{x}$.*

Remark. Similar working rules are followed to find the values of M.D.(median) and coefficient of M.D. (median).

Example 6. *Find M.D. $(\overline{x})$ and M.D.(median) for the following statistical series :*

7, 10, 12, 13, 15, 20, 21, 27, 30, 35.

Solution. **Calculation of M.D. ($\overline{\mathbf{x}}$)**

S. No.	x	$x - \overline{x}$ $\overline{x} = 19$	$\lvert x - \overline{x}\rvert$
1	7	– 12	12
2	10	– 9	9
3	12	– 7	7
4	13	– 6	6
5	15	– 4	4
6	20	1	1
7	21	2	2
8	27	8	8
9	30	11	11
10	35	16	16
$n = 10$	$\Sigma x = 190$		$\Sigma\lvert x - \overline{x}\rvert = 76$

NOTES

$$\bar{x} = \frac{\Sigma x}{n} = \frac{190}{10} = 19.$$

$$\therefore \quad \text{M.D.}(\bar{x}) = \frac{\Sigma|x-\bar{x}|}{n} = \frac{76}{10} = \mathbf{7.6.}$$

Calculation of M.D.(median)

S. No.	*x*	*x – median* *median = 17.5*	*| x – median |*
1	7	– 10.5	10.5
2	10	– 7.5	7.5
3	12	– 5.5	5.5
4	13	– 4.5	4.5
5	15	– 2.5	2.5
6	20	2.5	2.5
7	21	3.5	3.5
8	27	9.5	9.5
9	30	12.5	12.5
10	35	17.5	17.5
$n = 10$			Σ/x – median = 76

$$\frac{n+1}{2} = \frac{10+1}{2} = \mathbf{5.5}$$

$$\therefore \quad \text{Median} = \frac{\text{Size of 5th item + size of 6th item}}{2} = \frac{15+20}{2} = 17.5$$

$$\therefore \quad \text{M.D.(median)} = \frac{\Sigma|x-\text{median}|}{n} = \frac{76}{10} = \mathbf{7.6.}$$

Example 7. *Find the M.D. from A.M. for the following data :*

x	*3*	*5*	*7*	*9*	*11*	*13*
f	*2*	*7*	*10*	*9*	*5*	*2*

Solution. **Calculation of M.D. $(\bar{x})$**

x	*f*	*fx*	$x-\bar{x}$	$\|x-\bar{x}\|$	$f\|x-\bar{x}\|$
3	2	6	– 4.8	4.8	9.6
5	7	35	– 2.8	2.8	19.6
7	10	70	– 0.8	0.8	8.0
9	9	81	1.2	1.2	10.8
11	5	55	3.2	3.2	16.0
13	2	26	5.2	5.2	10.4
	N = 35	Σ*fx* = 273			$\Sigma f\|x-\bar{x}\| = 74.4$

$$\bar{x} = \frac{\Sigma fx}{\text{N}} = \frac{273}{35} = 7.8$$

Now $$\text{M.D.}(\bar{x}) = \frac{\Sigma f|x-\bar{x}|}{\text{N}} = \frac{74.4}{35} = \mathbf{2.1257.}$$

4.8. SHORT-CUT METHOD FOR M.D.

We know that the calculation of M.D. involve taking of deviations of items from some average. If the value of the average under consideration is a whole number, we can easily take the deviations and proceed without any difficulty. But in case, the value of the average comes out to be in decimal like 18.6747, the calculation of M.D. would become quite tedious. In such a case, we would have to approximate the value of the average up to one or two places of decimal for otherwise we would have to bear the heavy calculation work involved. If the value of the average is in decimal, the following short-cut method is preferred.

$$\textbf{M.D.} = \frac{(\Sigma \mathbf{fx})_A - (\Sigma \mathbf{fx})_B - ((\Sigma \mathbf{f})_A - (\Sigma \mathbf{f})_B)\,\mathbf{a}}{\mathbf{N}}$$

where 'a' is the average about which M.D. is to be calculated. In this formula, suffixes A and B denote the sums corresponding to the values of $x \geq a$ and $x < a$ respectively.

This formula can also be used for an individual series, by taking 'f' equal to 1 for each x, in the series. In this case, the formula reduces to

$$\textbf{M.D.} = \frac{(\Sigma \mathbf{x})_A - (\Sigma \mathbf{x})_B - ((\mathbf{n})_A - (\mathbf{n})_B)\,\mathbf{a}}{\mathbf{n}}$$

where $(n)_A$ and $(n)_B$ are the number of items whose values are greater than or equal to a and less than a respectively.

If short-cut method is to be used to find M.D.$(\bar{x})$, then it is advisable to use *direct method* to find $\bar{x}$, because we would be needing $(\Sigma fx)_A$ and $(\Sigma fx)_B$ in the calculation of M.D.$(\bar{x})$.

Example 8. *Calculate M.D.(Median) for the following data :*

x : *4, 6, 10, 12, 18, 19.*

Solution. **Calculation of M.D. (Median)**

S. No.	x	x – *median*	| x – *median* |
1	4	– 7	7
2	6 $(\Sigma x)_B$	– 5	5
3	10 = 20	– 1	1
4	12	1	1
5	18 $(\Sigma x)_A$	7	7
6	19 = 49	8	8
$n = 6$			$\Sigma \mid x - \text{median} \mid = 29$

$$\text{Median} = \text{size of } \frac{6+1}{2} \text{ th item} = \text{size of 3.5th item} = \frac{10+12}{2} = 11.$$

Direct Method

$$\text{M.D. (Median)} = \frac{\Sigma \mid x - \text{median} \mid}{n} = \frac{29}{6} = \mathbf{4.8333.}$$

Short-cut Method

$$\text{M.D. (Median)} = \frac{(\Sigma x)_A - (\Sigma x)_B - ((n)_A - (n)_B)\,\text{median}}{n}$$

$$= \frac{49 - 20 - (3-3).11}{6} = \frac{29}{6} = \mathbf{4.8333.}$$

Example 9. *Calculate the mean deviation from the median for the following distribution :*

x	*20*	*40*	*60*	*80*	*100*	*120*	*140*	*160*	*180*	*200*	*220*	*240*
f	*3*	*13*	*43*	*102*	*175*	*220*	*204*	*139*	*69*	*25*	*6*	*1*

Solution. **Calculation of M.D. (median)**

S. No.	*x*	*f*		*c.f.*	*fx*	
1	20	3	$(\Sigma f)_B = 556$	3	60	$(\Sigma fx)_B = 55220$
2	40	13		16	520	
3	60	43		59	2580	
4	80	102		161	8160	
5	100	175		336	17500	
6	120	220		556	26400	
7	140	204	$(\Sigma f)_A = 444$	760	28560	$(\Sigma fx)_A = 69780$
8	160	139		899	22240	
9	180	69		968	12420	
10	200	25		993	5000	
11	220	6		999	1320	
12	240	1		1000	240	
		N = 1000			$\Sigma fx = 125000$	

$$\frac{N+1}{2} = \frac{1000+1}{2} = 500.5$$

$$\therefore \quad \text{Median} = \text{size of 500.5th item} = \frac{120+120}{2} = 120.$$

$$\text{Now, M.D. (median)} = \frac{(\Sigma fx)_A - (\Sigma fx)_B - [(\Sigma f)_A - (\Sigma f)_B] \text{ median}}{N}$$

$$= \frac{69780 - 55220 - (444 - 556)\,120}{1000} = \frac{28000}{1000} = \mathbf{28.}$$

4.8.1 Merits of M.D.

1. It is simple to understand.
2. It is easy to compute.
3. It is well-defined.
4. It is based on all the items.
5. It is not unduly affected by the extreme items.
6. It can be calculated by using any average.

4.8.2. Demerits of M.D.

1. It is not capable of further algebraic treatment.

2. It does not take into account the signs of the deviations of items from the average value.

EXERCISE 4.3

1. Calculate M.D.($\bar{x}$) and its coefficient for the following individual series :

21, 23, 25, 28, 30, 32, 38, 39, 46, 48.

2. Find the mean deviation about A.M. for the following data :

x	2	3	5	9	10
f	3	6	10	7	4

3. The following table gives the monthly distribution of wages of 1000 employees in a certain factory :

Wages (in ₹)	20	40	60	80	100	120
No. of employees	3	13	43	102	175	220
Wages (in ₹)	140	160	180	200	220	240
No. of employees	204	139	69	25	6	1

4. Calculate the mean deviation about median and its coefficient for the following frequency distribution :

Marks	0—10	10—20	20—30	30—40	40—50
No. of students	6	7	15	16	6

5. Calculate the M.D.($\bar{x}$) for the following data regarding the difference in age between husbands and wives :

Difference in age (in years)	0—5	5—10	10—15	15—20
No. of couples	449	705	507	281
Difference in age (in years)	20—25	25—30	30—35	35—40
No. of couples	109	52	16	4

6. Find M.D.($\bar{x}$) for the following distribution :

Class	15—24	25—34	35—44	45—54	55—64
Frequency	4000	16000	28000	33000	28000

7. Calculate median and M.D.(Median) for the following frequency distribution :

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

NOTES

Answers

1. M.D.$(\bar{x})$ = 7.8. coeff. of M.D.$(\bar{x})$ = 0.2364.
2. 2.54
3. Coeff. of M.D.$(\bar{x})$ = 0.2285, Coeff. of M.D.(median) = 0.2333
4. 9.76 marks, 0.3485.
5. M.D.$(\bar{x})$ = 5.34 years.
6. M.D.$(\bar{x})$ = 9.656.
7. Median = 19.95 years, M.D.(median) = 7.1 years.

IV. STANDARD DEVIATION (S.D.)

4.9. DEFINITION OF STANDARD DEVIATION

It is the most important measure of dispersion. It finds indispensable place in advanced statistical methods. The **standard deviation** of a statistical data is defined as the positive square root of the A.M. of the squared deviations of items from the A.M. of the series under consideration. The S.D. is often denoted by the greek letter 'σ'.

For an **individual series**, the S.D. is given by

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

where $x_1, x_2, \ldots\ldots, x_n$ are the value of the variable, under consideration.

For a **frequency distribution**,

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N}} = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{N}}$$

where f_i is the frequency of x_i $(1 \le i \le n)$.

When the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.

4.10. COEFFICIENT OF S.D., C.V., VARIANCE

For comparing two or more series for variability, the corresponding relative measure, called coefficient of S.D. is calculated. This measure is defined as :

$$\textbf{Coefficient of S.D.} = \frac{\text{S.D.}}{\bar{x}}.$$

The product of coefficient of S.D. and 100 is called as the *coefficient of variation*.

$$\therefore \quad \textbf{Coefficient of variation} = \left(\frac{\text{S.D.}}{\bar{x}}\right)100.$$

This measure is denoted as C.V.

$$\therefore \quad \text{C.V.} = \left(\frac{\text{S.D.}}{\bar{x}}\right) 100.$$

In practical problems, we prefer comparing C.V. instead of comparing coefficient of S.D. The coefficient of variation is also represented as percentage. The square of S.D. is called the **variance** of the distribution.

NOTES

WORKING RULES TO FIND S.D.

Rule I. *In case of an individual series, first find $\bar{x}$ by using the formula $\bar{x} = \frac{\Sigma x}{n}$. In the second step, find the values of $x - \bar{x}$. In the next step, find the squares $(x - \bar{x})^2$ of the values of $x - \bar{x}$. Find the sum $\Sigma (x - \bar{x})^2$ of the values of $(x - \bar{x})^2$. Divide this sum by n. Take the positive square root of this to get the value of S.D.*

Rule II. *In case of a frequency distribution, first find $\bar{x}$ by using the formula $\bar{x} = \frac{\Sigma fx}{N}$. In the second step, find the values of $x - \bar{x}$. In the next step, find the squares $(x - \bar{x})^2$ of the values of $x - \bar{x}$. Find the products of the values of $(x - \bar{x})^2$ and their corresponding frequencies. Find the sum $\Sigma f(x - \bar{x})^2$ of these products. Divide this sum by N. Take the positive square root of this to get the value of S.D.*

Rule III. *If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.*

Rule IV. (*i*) *Coeff. of S.D.* $= \frac{S.D.}{A.M.}$

(*ii*) *Coeff. of variation (C.V.)* $= \frac{S.D.}{A.M.} \times 100$

(*iii*) *Variance* $= (S.D.)^2$.

Example 10. *Find the S.D. and C.V. for the following data :*

4, 6, 10, 12, 18.

Solution. **Calculation of S.D. and C.V.**

S. No.	*x*	$x - \bar{x}$ $\bar{x} = 10$	$(x - \bar{x})^2$
1	4	– 6	36
2	6	– 4	16
3	10	0	0
4	12	2	4
5	18	8	64
$n = 5$	$\Sigma x = 50$		$\Sigma(x - \bar{x})^2 = 120$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{50}{5} = 10.$$

NOTES

Now, $$\text{S.D.} = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}} = \sqrt{\frac{120}{5}} = \sqrt{24} = \mathbf{4.8989.}$$

$$\text{C.V.} = \left(\frac{\text{S.D.}}{\bar{x}}\right) 100 = \left(\frac{4.8989}{10}\right) 100 = \mathbf{48.989\%.}$$

Example 11. *Calculate S.D. and C.V. for the following frequency distribution :*

Class	*Frequency*	*Class*	*Frequency*
4—8	*11*	*24—28*	*9*
8—12	*13*	*28—32*	*17*
12—16	*16*	*32—36*	*6*
16—20	*14*	*36—40*	*4*
20—24	*14*		

Solution. **Calculation of S.D. and C.V.**

Class	x	f	fx	$x-\bar{x}$	$(x-\bar{x})^2$	$f(x-\bar{x})^2$
4—8	6	11	66	– 14	196	2156
8—12	10	13	130	– 10	100	1300
12—16	14	16	224	– 6	36	576
16—20	18	14	252	– 2	4	56
20—24	22	14	308	2	4	56
24—28	26	9	234	6	36	324
28—32	30	17	510	10	100	1700
32—36	34	6	204	14	196	1176
36—40	38	4	152	18	324	1296
		N = 104	Σfx = 2080			$\Sigma f(x-\bar{x})^2$ = 8640

$$\bar{x} = \frac{\Sigma fx}{\text{N}} = \frac{2080}{104} = 20.$$

Now $$\text{S.D.} = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\text{N}}} = \sqrt{\frac{8640}{104}} = \mathbf{9.1146.}$$

$$\text{C.V.} = \left(\frac{\text{S.D.}}{\bar{x}}\right) 100 = \left(\frac{9.1146}{20}\right) 100 = \mathbf{45.573\%.}$$

4.11. SHORT-CUT METHOD FOR S.D.

We have seen in the above examples that the calculations of S.D. involves a lot of computation work. Even if the value of A.M. is a whole number, the calculations are not so simple. In case, A.M. is in decimal, then the calculation work would become more tedious. In problems, where A.M. is expected to be in decimal, we shall use this method, which is based on deviations (or step deviations) of items in the series.

For an individual series $x_1, x_2, ..., x_n$, we have

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n} - \left(\frac{\sum_{i=1}^{n} u_i}{n}\right)^2} \;.\, h = \sqrt{\frac{\Sigma u^2}{n} - \left(\frac{\Sigma u}{n}\right)^2} \;.\, h$$

NOTES

where $u_i = \dfrac{x_i - A}{h}$, $1 \le i \le n$.

For a frequency distribution, this formula takes the form

$$\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{n} f_i u_i^2}{N} - \left(\frac{\sum_{i=1}^{n} f_i u_i}{N}\right)^2} \;.\, h = \sqrt{\frac{\Sigma\, fu^2}{N} - \left(\frac{\Sigma\, fu}{N}\right)^2} \;.\, h$$

where f_i is the frequency of x_i $(1 \le i \le n)$ and $u_i = \dfrac{x_i - A}{h}$, $1 \le i \le n$.

A and h are constants to be chosen suitably. This method is also known as *step deviation method.*

In practical problems, it is advisable to first take deviations 'd' of the values of the variable (x) from some suitable number 'A'. Then we see if there is any common factor greater than one, in the values of the deviations. If there is a common factor h (> 1), then we calculate $u = \dfrac{d}{h} = \dfrac{x - A}{h}$ in the next column. In case, there is no common factor greater one. then we take $h = 1$ and u becomes $u = \dfrac{d}{1} = x - A$.

In this case, the formula reduces as given below :

$$\text{S.D.} = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} \qquad \textbf{(Individual Series)}$$

$$\text{S.D.} = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \qquad \textbf{(Frequency Distribution)}$$

where d = x – A and A is any constant, to be chosen suitably.

WORKING RULES TO FIND S.D.

Rule I. *In case of an individual series, choose a number A. Find deviations d(= x – A) of items from A. Find the squares 'd²' of the values of d. Find S.D. by using the formula*

$$\sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}.$$

If some common factor h (> 1) is available in the values of d, then we calculate 'u' by dividing the values of d by h. Find the squares 'u²' of the values of u. Find S.D. by using the formula : $\sqrt{\dfrac{\Sigma u^2}{n} - \left(\dfrac{\Sigma u}{n}\right)^2} \times h.$

Rule II. *In case of a frequency distribution, choose a number A. Find deviations d(= x – A) of items from A. Find the products fd of f and d. Next, find the products of fd and d. Find the sums Σfd and Σfd². Find S.D. by using the formula :*

$$\sqrt{\frac{\Sigma fd^2}{N}-\left(\frac{\Sigma fd}{N}\right)^2}.$$

If some common factor h(> 1) is available in the values of d, then we calculate 'u' by dividing the values of d by h. Find the product fu of f and u. Next find the products of fu and u. Find the sums Σfu and Σfu². Find S.D. by using the formula :

$$\sqrt{\frac{\Sigma fu^2}{N}-\left(\frac{\Sigma fu}{N}\right)^2}\times h.$$

Rule III. *If the values of the variable are given in the form of classes, then their respective mid-points are taken as the values of the variable.*

Example 12. *Find the C.V. of the following individual series :*

2.76398, 4.76398, 6.76398, 8.76398, 12.76398, 10.76398.

Solution. **Calculation of S.D. and $\overline{x}$**

S.No.	*x*	*d = x – A* *A = 8.76398*	*u = d/h* *h = 2*	*u²*
1	2.76398	– 6	– 3	9
2	4.76398	– 4	– 2	4
3	6.76398	– 2	– 1	1
4	8.76398	0	0	0
5	12.76398	4	2	4
6	10.76398	2	1	1
n = 6			Σ*u* = – 3	Σ*u*² = 19

Now $$\overline{x} = A+\left(\frac{\Sigma u}{n}\right)h = 8.76398+\left(\frac{-3}{6}\right).\ 2 = 7.76398$$

$$\text{S.D.} = \sqrt{\frac{\Sigma u^2}{n}-\left(\frac{\Sigma u}{n}\right)^2}\ .\ h = \sqrt{\frac{19}{6}-\left(\frac{-3}{6}\right)^2}\times 2 = \sqrt{2.9167}\times 2 = 3.4157.$$

$$\therefore \quad \text{C.V.} = \left(\frac{\text{S.D.}}{\overline{x}}\right)100 = \left(\frac{3.4157}{7.76398}\right)100 = \mathbf{43.9942\%}.$$

Example 13. *Find the value of coefficient of variation for the following frequency distribution :*

Class	*0—5*	*5—10*	*10—15*	*15—20*	*20—25*
No. of items	*20*	*24*	*32*	*28*	*20*
Class	*25—30*	*30—35*	*35—40*	*40—45*	
No. of items	*16*	*34*	*10*	*16*	

Solution. **Calculation of S.D. and $\bar{x}$**

Class	f	x	$d = x - A$ $A = 22.5$	$u = d/h$ $h = 5$	fu	fu^2
0—5	20	2.5	− 20	− 4	− 80	320
5—10	24	7.5	− 15	− 3	− 72	216
10—15	32	12.5	− 10	− 2	− 64	128
15—20	28	17.5	− 5	− 1	− 28	28
20—25	20	22.5	0	0	0	0
25—30	16	27.5	5	1	16	16
30—35	34	32.5	10	2	68	136
35—40	10	37.5	15	3	30	90
40—45	16	42.5	20	4	64	256
	N = 200				Σfu $= -66$	Σfu^2 $= 1190$

Now $$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = 22.5 + \left(\frac{-66}{200}\right)5 = 20.85$$

$$\text{S.D.} = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2}\ .\ h = \sqrt{\frac{1190}{200} - \left(\frac{-66}{200}\right)^2} \times 5$$

$$= \sqrt{5.8411} \times 5 = 12.0842.$$

$$\therefore \quad \text{C.V.} = \left(\frac{\text{S.D.}}{\bar{x}}\right) 100 = \left(\frac{12.0842}{20.85}\right) 100 = \mathbf{57.9578\%}.$$

Example 14. *A student obtained the A.M. and S.D. of 100 observations as 40 and 5.1 respectively. Later on, it was discovered that he had wrongly copied down an observation as 50 instead of 40. Calculate the correct value of S.D.*

Solution. We have

No. of items	= 100
Incorrect $\bar{x}$	= 40
Incorrect S.D.	= 5.1
Correct item	= 40
Incorrect item	= 50

Now $$\bar{x} = \frac{\Sigma x}{n}$$

$$\therefore \quad 40 = \frac{\text{Incorrect } \Sigma x}{100} \quad \text{or} \quad \text{Incorrect } \Sigma x = 4000$$

$$\therefore \quad \text{Correct } \Sigma x = 4000 - 50 + 40 = 3990$$

$$\therefore \quad \text{Correct } \bar{x} = \frac{3990}{100} = 39.9.$$

Now $$\text{S.D.} = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2}\ .$$

Note. The reader is advised to note this form of S.D. carefully.

$$\therefore \qquad 5.1 = \sqrt{\frac{\text{Incorrect } \Sigma x^2}{100} - (40)^2}$$

or
$$26.01 = \frac{\text{Incorrect } \Sigma x^2}{100} - 1600$$

$$\therefore \qquad \text{Incorrect } \Sigma x^2 = (1626.01)\ 100 = 162601$$

$$\therefore \qquad \text{Correct } \Sigma x^2 = 162601 - (50)^2 + (40)^2 = 161701$$

Now
$$\text{Correct S.D.} = \sqrt{\frac{161701}{100} - (39.9)^2} = \sqrt{1617.01 - 1592.01} = 5.$$

4.12. RELATION BETWEEN MEASURES OF DISPERSION

It has been observed that in frequency distribution, the following relations hold.

1. Q.D. is approximately equal to $\frac{2}{3}$ S.D.
2. M.D. is approximately equal to $\frac{4}{5}$ S.D.

4.12.1 Merits of S.D.

1. It is simple to understand.
2. It is well-defined.
3. In the calculation of S.D., the signs of deviations of items are also taken into account.
4. It is based on all the items.
5. It is capable of further algebraic treatment.
6. It has sampling stability.
7. It is very useful in the study of "Tests of Significance".

4.12.2 Demerits oF S.D.

1. It is not easy to calculate.
2. It is unduly affected by the extreme items, because the squares of deviations of extreme items would be either extremely low or extremely high.

EXERCISE 4.4

1. (*a*) Find S.D. and C.V. for the following individual series :

4, 4, 4, 4, 4, 4, 4.

(*b*) The A.M. of the runs scored by three batsmen A, B, C in the same innings are 58, 48, 12 respectively. The S.D. of their runs are 15, 12 and 2 respectively. Find who is most consistent of the three.

NOTES

2.

Class interval	60—70	50—60	40—50	30—40	20—30	10—20
Frequency	3	6	10	12	15	6

Find coefficient of variation for the above data.

3. Find which of the following batsman is more consistent in scoring :

Batsman A	5	7	16	27	53	80
Batsman B	0	4	16	21	43	83

4. A group of 100 selected students is with average height 168.8 cm and coefficient of variation 3.2%. What is the S.D. of their height ?

5. Goals scored by two teams in a football season were as follows :

No. of goals scored in a match	*No. of matches*	
	Team A	*Team B*
0	15	20
1	10	10
2	07	05
3	05	04
4	03	02
5	02	01
Total	42	42

Calculate coefficient of variation and state which team is more consistent.

6. From the data given below, state which series is more variable :

Variable	10—20	20—30	30—40	40—50	50—60	60—70
Group A	10	18	32	22	40	18
Group B	18	22	40	18	32	10

7. The following is the data relating to two models of televisions :

Life (No. of years)	*No. of televisions*	
	Model : Crown	*Dyanora*
0—2	5	2
2—4	16	7
4—6	13	12
6—8	7	19
8—10	5	9
10—12	4	1

(*i*) What is the average life of each model of these televisions ?

(*ii*) Which model has more uniformity ?

NOTES

8. The following table gives the distribution of wages in the two branches of an industrial concern. Find out which branch has greater variability in wages relating to the average wage :

Monthly wages (in ₹)	*No. of workers*	
	Branch A	*Branch B*
50—100	20	8
100—150	35	17
150—200	42	43
200—250	45	20
250—300	58	12
Total	200	100

9. Mean and standard deviation of 200 items are found to be 60 and 20. If at the time of calculations two items are wrongly taken as 3 and 67 instead of 13 and 17, find the correct mean and the standard deviation.

10. The analysis of the results of a budget survey of 150 families gave an average monthly expenditure of ₹ 120 on food items with a S.D. of ₹ 15. After the analysis was completed it was noticed that the figure recorded for one household was wrongly taken as ₹ 15 instead of ₹ 105. Determine the correct value of the average expenditure and its S.D.

11. Following is the data related to two factories:

	Factory A	*Factory B*
Numbers of Workers	200	250
Average wage per hour	₹ 15.00	₹ 12.00
Variance (σ^2)	₹ 16	₹ 9

Find the following:

(*i*) Which factory pays larger amount as total wages per hour?

(*ii*) Which factory is more variable?

(*iii*) Find combined S.D.

Answers

1. (*a*) S.D. = 0, C.V. = 0

(*b*) C.V. (A) = 25.86%, C.V. (B) = 25.00%,

C.V. (C) = 16.67% ∴ C is most consistent.

2. C.V. = 38.708

3. C.V. for A = 81.0306%, C.V. for B = 101.6502%, A is consistent.

4. S.D. = 5.4016 cm

5. $\left.\begin{array}{l}\text{C.V. for A} = 102.1259\\ \text{C.V. for B} = 124.5434\end{array}\right\}$ A is consistent.

6. C.V. for A = 33.8496%, C.V. for B = 38.2442%

Group B is more variable.

7. (*i*) A.M. for Crown = 5.12 yrs, A.M. for Dyanora = 6.16 yrs

(*ii*) C.V. for Crown = 54.9158%, C.V. for Dyanora = 36.2068%

∴ Dyanora model has more uniformity..

8. C.V. for A = 33.9015%. C.V. for B =.29.8077%
   Variability is more in branch A.
9. 59.8, 20.0938
10. Correct A.M. = ₹ 120.60, Correct S.D. = ₹ 12.35
11. (*i*) Same amount of ₹ 3,000 (*ii*) A (*iii*) ₹ 3.786

NOTES

EXERCISE 4.5

1. Define dispersion and discuss its various measures.
2. What are the requisites of a good measure of dispersion ?

4.12. SUMMARY

- The requisites of a good measure of a dispersion are the same as those for a good measure of central tendency.
- The **range** of a statistical data is defined as the difference between the largest and the smallest values of the variable.
- The **mean deviation** of a statistical data is defined as the arithmetic mean of the numerical values of the deviations of items from some average.
- The **standard deviation** of a statistical data is defined as the positive square root of the A.M. of the squared deviations of items from the A.M. of the series under consideration.

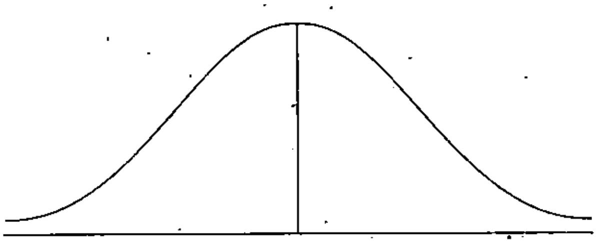
NOTES

5. SKEWNESS

STRUCTURE

5.1. INTRODUCTION

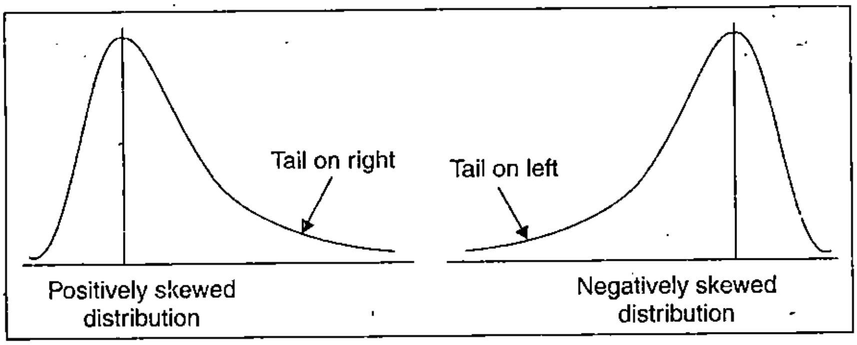
In symmetrical distribution, the values of mean, mode and median, would coincide. If the curve of the distribution is not symmetrical, it may admit of tail on either side of the distribution. Such a distribution lack in symmetry. **Skewness** is the word used for lack of symmetry. A distribution which is not symmetrical is called **asymmetrical** or **skewed**. We can define '**skewness**' of a distribution as the tendency of a distribution to depart from symmetry.



$\bar{x}$ = Mode = Median

Symmetrical distribution

NOTES


Tail on right
Tail on left
Positively skewed distribution
Negatively skewed distribution

If the tail of an asymmetrical distribution is on the right side, then the distribution is called a **positively skewed distribution**. If the tail is on left side, then the distribution is defined to be **negatively skewed distribution**. Now we shall account for the situations when skewness can be expected in a distribution.

5.2. TESTS OF SKEWNESS

1. If A.M. = mode = median, then there is no skewness in the distribution. In other words, the curve of the frequency distribution would be symmetrical, bell-shaped.

2. If A.M. is less than (greater than), the value of mode, the tail would on left (right) side, *i.e.,* the distribution is negatively (positively) skewed.

3. If sum of frequencies of values less than mode is equal to the sum of frequencies of values greater than mode, then there would be no skewness.

4. If quartiles are equidistant from median, then there would be no skewness.

5.3. KARL PEARSON'S METHOD

This method is based on the fact that in a symmetrical distribution, the value of A.M. is equal to that of mode. As we have already noted that the distribution is positively skewed if A.M. > Mode and negatively skewed if A.M. < Mode. The Karl Pearson's coefficient of skewness is given by

$$\textbf{Karl Pearson's coefficient of skewness} = \frac{\textbf{A.M.} - \textbf{Mode}}{\textbf{S.D.}}.$$

We have already studied the methods of calculating A.M., mode and S.D. of frequency distributions. If mode is ill-defined in some frequency distribution, then the value of empirical mode is used in the formula.

Empirical mode = 3 Median – 2 A.M.

$$\therefore \quad \text{Coeff. of skewness} = \frac{\text{A.M.} - \text{Mode}}{\text{S.D.}}$$

$$= \frac{\text{A.M.} - (3\text{ Median} - 2\text{ A.M.})}{\text{S.D.}} = \frac{3\text{ A.M.} - 3\text{ Median}}{\text{S.D.}}$$

$$\therefore \quad \textbf{Karl Pearson's coefficient of skewness} = \frac{\textbf{3 (A.M.} - \textbf{Median)}}{\textbf{S.D.}}$$

The coefficient of skewness as calculated by using this method would give magnitude as well as direction of skewness, present in the distribution. Practically, its value lies between – 1 and 1. For a symmetrical distribution, its value comes out to be zero.

The Karl Pearson's coefficient of skewness is generally denoted by 'SK_P'.

WORKING RULES FOR SOLVING PROBLEMS

Rule I. *If the values of $\bar{x}$, σ and mode are given, then find SK_P by using the formula :*

$$SK_P = \frac{\bar{x} - mode}{\sigma}.$$

Rule II. *If the values of $\bar{x}$, σ and median are given, then find SK_P by using the formula :*

$$SK_P = \frac{3\,(\bar{x} - median)}{\sigma}.$$

Rule III. *If the values of $\bar{x}$, σ and mode are not given, then calculate these. If mode is ill-defined, then find median.*

Rule IV. *Find SK_P by using formulae given in above rules.*

Example 1. *Karl Pearson's coefficient of skewness of a distribution is 0.32, its standard deviation is 6.5 and mean is 29.6. Find the mode of the distribution.*

Solution. We have $SK_P = 0.32$, S.D. = 6.5, $\bar{x} = 29.6$.

Now $$SK_P = \frac{\bar{x} - \text{Mode}}{\text{S.D.}}$$

$\therefore$ $$0.32 = \frac{29.6 - \text{Mode}}{6.5}.$$

$\Rightarrow$ $$29.6 - \text{Mode} = 0.32 \times 6.5 = 2.08$$

$\Rightarrow$ $$\text{Mode} = 29.6 - 2.08 = \mathbf{27.52.}$$

Example 2. *For a moderately skewed data, the arithmetic mean is 100, the variance is 35 and Karl Pearson's coefficient of skewness is 0.2. Find its mode and median.*

Solution. We have $\bar{x} = 100$, variance = 35, $SK_P = 0.2$.

Now $$SK_P = \frac{\bar{x} - \text{Mode}}{\sigma}.$$

$\therefore$ $$0.2 = \frac{100 - \text{Mode}}{\sqrt{35}} \qquad (\because\ \text{S.D.} = \sqrt{\text{variance}})$$

$\Rightarrow$ $$100 - \text{Mode} = 0.2 \times 5.92 = 1.184$$

$\Rightarrow$ $$\text{Mode} = 100 - 1.184 = \mathbf{98.816.}$$

Also $$\text{Mode} = 3\ \text{Median} - 2\bar{x}$$

$\Rightarrow$ $$98.816 = 3\ \text{Median} - 2(100).$$

$\therefore$ $$3\ \text{Median} = 98.816 + 200 = 298.816$$

$\therefore$ $$\text{Median} = \frac{298.816}{3} = \mathbf{99.61.}$$

Example 3. *Find the coefficient of skewness by Karl Pearson's method for the following data :*

Value	*6*	*12*	*18*	*24*	*30*	*36*	*42*
Frequency	*4*	*7*	*9*	*18*	*15*	*10*	*3*

Solution. **Calculation of $\overline{\mathbf{x}}$, S.D.**

Value x	f	$d = x - A$ $A = 24$	$u = d/h$ $h = 6$	fu	fu^2
6	4	– 18	– 3	– 12	36
12	7	– 12	– 2	– 14	28
18	9	– 6	– 1	– 9	9
24	18	0	0	0	0
30	15	6	1	15	15
36	10	12	2	20	40
42	3	18	3	9	27
	N = 66			$\Sigma fu = 9$	$\Sigma fu^2 = 155$

A.M.

$$\bar{x} = \text{A} + \left(\frac{\Sigma fu}{\text{N}}\right)h = 24 + \left(\frac{9}{66}\right)6 = 24.82$$

S.D.

$$\text{SD} = \sqrt{\frac{\Sigma fu^2}{\text{N}} - \left(\frac{(\Sigma fu)}{\text{N}}\right)^2} \times h = \sqrt{\frac{155}{66} - \left(\frac{9}{66}\right)^2} \times 6$$

$$= \sqrt{2.35 - 0.12} \times 6 = 1.49 \times 6 = 8.94.$$

Mode. **Grouping Table**

x	*I* f	*II*	*III*	*IV*	*V*	*VI*
6	4	11		20		
12	7		16		**34**	
18	9	**27**				**42**
24	**18**		**33**	**43**		
30	15	25			28	
36	10		13			
42	3					

Analysis Table

Column	*24*	*18*	*30*	*36*	*12*
I	1				
II	1	1			
III	1		1		
IV	1		1	1	
V	1	1			1
VI	1	1	1		
Total	6	3	3	1	1

$$\therefore \quad \text{Mode} = 24$$

$$\therefore \quad \text{SK}_P = \frac{\bar{x} - \text{Mode}}{\text{S.D}} = \frac{24.82 - 24}{8.94} = \frac{0.82}{8.94} = \mathbf{0.092.}$$

Example 4. *Calculate Karl Pearson's coefficient of skewness for the following data :*

Income (in ₹)	*500—600*	*600—700*	*700—800*
No. of employees	*8*	*12*	*4*
Income (in ₹)	*800—900*	*900—1000*	*1000—1100*
No. of employees	*2*	*1*	*1*

Solution. **Calculation of $\bar{\mathbf{x}}$, Mode, S.D.**

Income (in ₹)	*No. of employees* f	*Mid-points of classes* x	$d = x - A$ $A = 750$	$u = d/h$ $h = 100$	fu	fu^2
500—600	8	550	– 200	– 2	– 16	32
600—700	12	650	– 100	– 1	– 12	12
700—800	4	750	0	0	0	0
800—900	2	850	100	1	2	2
900—1000	1	950	200	2	2	4
1000—1100	1	1050	300	3	3	9
	N = 28				$\Sigma fu = -21$	$\Sigma fu^2 = 59$

A.M.,
$$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = 750 + \left(-\frac{21}{28}\right)(100) = 750 - 75 = ₹\ 675$$

Mode. By inspection, modal class is 600—700.

$$\therefore \quad \text{Mode} = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)h$$

Here $\quad L = 600,\ \Delta_1 = 12 - 8 = 4,\ \Delta_2 = 12 - 4 = 8,\ h = 100$

$$\therefore \quad \text{Mode} = 600 + \left(\frac{4}{4+8}\right)(100) = 600 + 33.33 = ₹\ 633.33$$

S.D.,
$$\text{S.D.} = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h = \sqrt{\frac{59}{28} - \left(-\frac{21}{28}\right)^2} \times 100$$

$$= \sqrt{2.1071 - 0.5625} \times 100 = \sqrt{1.5446} \times 100$$

$$= 1.2428 \times 100 = ₹\ 124.28$$

Now, Karl Pearson's coeff. of skewness

$$= \frac{\bar{x} - \text{Mode}}{\text{S.D.}} = \frac{675 - 633.33}{124.28} = \mathbf{0.34.}$$

EXERCISE 5.1

1. A frequency distribution gives the following results :

   Coeff. of variation = 5
   Karl Pearson's Coeff. of Skewness = 0.5
   S.D. = 2

   Find A.M. and Mode of the distribution.

2. For the following data, find Karl Pearson's coefficient of skewness :

Height (in inches)	58	59	60	61	62	63
No. of persons	10	18	30	42	35	28

3. Calculate skewness and its coefficient from the following data (Use Karl Pearson's Formula):

Wage (in ₹)	10	11	12	13	14	15	16
No. of Workers	4	7	9	15	8	5	2

4. Calculate Karl Pearson's coefficient of skewness for the following frequency distribution :

Marks more than	0	10	20	30	40	50	60	70
No. of students	100	90	75	50	25	15	5	0

5. For the following data, calculate Karl Pearson's coefficient of skewness :

Marks (above)	0	10	20	30	40	50	60	70	80
No. of students	150	140	100	80	80	70	30	14	0

Answers

1. $\bar{x} = 40$, Mode = 39 2. – 0.0208 3. 1.52, – 0.14

4. – 0.0627 5. – 0.7539

5.4. BOWLEY'S METHOD

This method is based on the fact that in a symmetrical distribution, the quartiles are equidistant from the median. In a skewed distribution, this would not happen. The Bowley's coefficient of skewness is given by

$$\textbf{Bowley's coefficient of skewness} = \frac{\mathbf{Q_3 + Q_1 - 2\,Median}}{\mathbf{Q_3 - Q_1}}.$$

For a symmetrical distribution, its values would come out to be zero. The value of Bowley's coefficient of skewness lies between – 1 and + 1. The coefficient of skewness as calculated by using this, would give magnitude as well as direction of skewness present in the distribution. In problems, it is generally given as to which method is to be used. But in case, the method to be used is not specifically mentioned, then it is advisable to use Bowley's method. The calculation of Bowley's coefficient of skewness would involve the calculation of Q_1, Q_3 and median. The calculation of these measures would definitely take lesser time than for the calculation of $\bar{x}$, mode and S.D. It may also be noted that the values of coefficient of skewness as calculated by using different formulae may not be same. This method is also useful in case of open end classes in the distribution.

NOTES

The Bowley's coefficient of skewness is generally denoted by 'SK$_B$'.

WORKING RULES FOR SOLVING PROBLEMS

Rule I. *If the values of medium, Q_1 and Q_3 are given, then find SK_B by using the formula :* $SK_B = \frac{Q_3 + Q_1 - 2Median}{Q_3 - Q_1}$.

Rule II. *If the values of median, Q_1 and Q_3 are not given, then find these by using cumulative frequencies of the distribution.*

Rule III. *If the name of the method is not mentioned, then the coefficient should be calculated by Bowley's method. This method will take less time.*

Example 5. *For a distribution, the Bowley's coefficient of skewness = – 0.36, Q_1 = 8.6 and median 12.3. What is the value of the coefficient of Q.D. ?*

Solution. We have, median = 12.3, Q_1 = 8.6 and Bowley's coeff. of skewness = – 0.36.

$$\text{Bowley's coeff. of skewness} = \frac{Q_3 + Q_1 - 2\text{ Median}}{Q_3 - Q_1}$$

$$\therefore \quad -0.36 = \frac{Q_3 + 8.6 - 2(12.3)}{Q_3 - 8.6}$$

$$\therefore \quad -0.36\,Q_3 + 3.096 = Q_3 - 16$$

$$\therefore \quad 1.36\,Q_3 = 19.096$$

$$\therefore \quad Q_3 = \frac{19.096}{1.36} = 14.0412$$

Now $\text{coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{14.0412 - 8.6}{14.0412 + 8.6} = \frac{5.4412}{22.6412} = \mathbf{0.2403.}$

Example 6. *In a frequency distribution, the following measures were calculated :*

Bowley's coeff. of skewness	*= 0.35*
Median	*= 75*
Q.D.	*= 6*

Find the value of the coefficient of Q.D.

Solution. We have coeff. of skewness = 0.35, median = 75 and Q.D. = 6.

Now $\text{Q.D.} = \frac{Q_3 - Q_1}{2}$

$$\Rightarrow \quad \frac{Q_3 - Q_1}{2} = 6 \Rightarrow Q_3 - Q_1 = 12.$$

Also $\text{coeff. of skewness} = \frac{Q_3 + Q_1 - 2\text{ Median}}{Q_3 - Q_1}$

$$\Rightarrow \quad 0.35 = \frac{Q_3 + Q_1 - 2(75)}{12}$$

$$\Rightarrow \quad 4.2 = Q_3 + Q_1 - 150 \Rightarrow Q_3 + Q_1 = 154.2$$

Now $\text{coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{12}{154.2} = \mathbf{0.0778.}$

NOTES

Example 7. *From the information given below, calculate Karl Pearson's coefficient of skewness and also Bowley's coefficient of skewness :*

Measure	*Place A*	*Place B*
Mean	*150*	*140*
Median	*142*	*155*
S.D.	*30*	*55*
Third Quartile	*195*	*260*
First Quartile	*62*	*80*

Solution. Place A

We have $\bar{x}$ = 150, median = 142, S.D. = 30, Q_3 = 195, Q_1 = 62.

$$\therefore \quad SK_P(A) = \frac{3(\bar{x} - \text{Median})}{\text{S.D.}} = \frac{3(150-142)}{30} = \frac{24}{30} = \mathbf{0.8.}$$

$$SK_B(A) = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1} = \frac{195+62-2(142)}{195-62} = \frac{-27}{133} = \mathbf{-\,0.203.}$$

Place B

We have $\bar{x}$ = 140, median = 155, S.D. = 55, Q_3 = 260, Q_1 = 80.

$$\therefore \quad SK_P(B) = \frac{3(\bar{x} - \text{Median})}{\text{S.D.}} = \frac{3(140-155)}{55} = \frac{-45}{55} = \mathbf{-\,0.818.}$$

$$SK_B(B) = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1} = \frac{260+80-2(155)}{260-80} = \frac{30}{180} = \mathbf{0.167.}$$

EXERCISE 5.2

1. In a frequency distribution, the difference of quartiles is 25 and their sum is 45. The median is found to be 15. Find Bowley's coefficient of skewness.
2. In a frequency distribution, the Bowley's coefficient of skewness is 0.80. The sum of upper and lower quartiles is 130. The median is 60. Calculate the values of the quartiles.
3. Calculate the Bowley's coefficient of skewness for the following frequency distribution :

More than	*No. of items*	*More than*	*No. of items*
0	5474	60	2718
10	5426	70	1406
20	5259	80	764
30	5023	90	370
40	4475	100	160
50	3712	110	39

4. Find Bowley's coefficient of skewness for the following data and show which section of carpenters is more skewed ?

Daily wage (in ₹)	55—58	58—61	61—64	64—67	67—70
No. of carpenters (Locality A)	12	17	23	18	11
No. of carpenters (Locality B)	20	22	25	13	7

NOTES

Answers

1. 0.6 **2.** $Q_1 = 58.75, Q_3 = 71.25$
3. – 0.1628
4. Coeff. of skewness (for A) = – 0.01429

   Coeff. of skewness (for B) = – 0.0597

   ∴ Skewness is greater for section B

EXERCISE 5.3

1. "Averages, Measures of Disperson and Skewness are complementary to one another in understanding a frequency distribution". Elucidate.
2. Define skewness. Explain the difference between positive skewness and negative skewness.
3. Explain what do you undertand by "Skewness".
4. Explain the use of moments in studying skewness in frequency distributions.
5. How does 'Skewness' differ from 'Dispersion' ? Explain the different methods of studying skewness.
6. Explain the use of quartiles in studying skewness in frequency distributions.

5.5. SUMMARY

- **Skewness** is the word used for lack of symmetry. A distribution which is not symmetrical is called **asymmetrical** or **skewed**. We can define '**skewness**' of a distribution as the tendency of a distribution to depart from symmetry.
- This method is based on the fact that in a symmetrical distribution, the value of A.M. is equal to that of mode. As we have already noted that the distribution is positively skewed if A.M. > Mode and negatively skewed if A.M. < Mode. The Karl Pearson's coefficient of skewness is given by

  $$\text{Karl Pearson's coefficient of skewness} = \frac{\text{A.M.} - \text{Mode}}{\text{S.D.}}.$$

- $$\text{Bowley's coefficient of skewness} = \frac{Q_3 + Q_1 - 2\ \text{Median}}{Q_3 - Q_1}.$$

  For a symmetrical distribution, its values would come out to be zero. The value of Bowley's coefficient of skewness lies between – 1 and + 1. The coefficient of skewness as calculated by using this, would give magnitude as well as direction of skewness present in the distribution.

NOTES

6. ANALYSIS OF TIME SERIES

STRUCTURE

6.1. INTRODUCTION

We know that a **time series** is a collection of values of a variable taken at different time periods. If $y_1, y_2, \ldots\ldots, y_n$ be the values of a variable y taken at time periods $t_1, t_2, \ldots\ldots t_n$, then we write this time series as $\{(t_i, y_i) ; i = 1, 2, \ldots\ldots, n\}$. The given time series data is arranged chronologically. If we consider the sale figures of a company for over 20 years, the data will constitute a time series. Population of a town, taken annually for 15 years, would form a time series. There are plenty of variables whose value depends on time.

6.2. MEANING OF TIME SERIES

NOTES

In a time series, the values of the concerned variable is not expected to be same for every time period. For example, if we consider the price of 1 kg tea of a particular brand, for over twenty years, we will note that the price is not the same for every year. What has caused the price to vary ? In fact, there is nothing special with tea, this can happen for any variable, we consider.

There are number of economic, psychological, sociological and other forces which may cause the value of the variable to change with time. In this chapter, we shall locate, measure and interpret the changes in the values of the variable, in a time series. We shall investigate the factors, which may be held responsible for causing changes in the values of the variable with respect to time.

6.3. COMPONENTS OF TIME SERIES

We have already noted that the value of variable in a time series are very rarely constant. The graph of its time series will be a zig-zag line. The variation in the values of time series are due to psychological, sociological, economic etc. forces. The variations in a time series are classified in to four types and are called **components** of the time series. The components are as follows :

(*i*) Secular trend or long-term variations

(*ii*) Seasonal variations

(*iii*) Cyclical variations

(*iv*) Irregular variations.

6.4. SECULAR TREND OR LONG TERM VARIATIONS

The general tendency of the values of the variable in a time series to grow or to decline over a long period of time is called **secular trend** of the times series. It indicates the general direction in which the graph of the time series appears to be going over a long period of time. The graph of the secular trend is either a straight line or a curve. This graph depends upon the nature of data and the method used to determine secular trend.

The secular trend of a time series depends much on factors which changes very slowly, *e.g.*, population, habits, technical development, scientific research etc.

If the secular trend for a particular time series is upward (downward), it does not necessarily imply that the values of the variable must be strictly increasing (decreasing). For example, consider the data :

Year	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
Profit *('000 ₹)*	18	17	20	21	25	22	26	27	28	35

We observe that the profit figures for the years 1979 and 1983 are less than those of their corresponding previous years, but for all other years the profit figures are greater than their corresponding previous years. In this time series, the general tendency of the profit figures is to grow.

If from the definition of secular trend, we drop the condition of having time series data for a long period of time, the definition will become meaningless. For example, if we consider the data :

Year	2002	2003
Price of sugar (1 kg)	₹ 14	₹ 14.50

From this time series, we cannot have the idea of the general tendency of the time series. In this connection, it is not justified to assert that the values of the variable must be taken for time periods covering 6 months or 10 years or 15 years. Rather we must see that the values of the variable are sufficient in number. Thus, in estimating trend, it is not the total time period that matters, but it is the number of time periods for which the values of the variable are known.

6.5. SEASONAL VARIATIONS

The **seasonal variations** in a time series counts for those variations in the series which occur annually. In a time series, seasonal variations occurs quite regularly. These variations play a very important role in business activities. There are number of factors which causes such variations. We know that the demand for raincoats rises automatically during rainy season. Producers of tea and coffee feels that the demand of their products is more in winter season rather than in summer season. Similarly, there is greater demand for cold drinks during summer season. Retailers on Hill stations are also affected by the seasonal variations. Their profits are heavily increased during summer season.

Even Banks have not escaped from seasonal variations. Banks observe heavy withdrawals in the first week of every month. Agricultural yield is also seasonal and so the farmers income is unevenly divided over the year. This has direct effect on business activities.

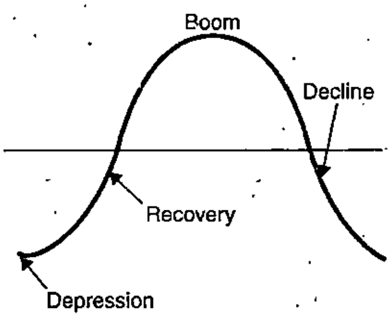
Customs and habits also plays an important role in causing seasonal variations in time series. On the eve of festivals, we are accustumed of purchasing sweets and new clothes. Generally, people get their houses white washed before Deepawali. Sale figures of retailers dealing with fireworks immediately boost up on the eve of Deepawali and in the season of marriages.

The study of seasonal variations in a time series is also very useful. By studying the seasonal variations, the businessman can adjust his stock holding during the year. He will not feel the danger of shortfall of stock during any particular period, in the year.

6.6. CYCLICAL VARIATIONS

The **cyclical variations** in a time series counts for the swings of graph of time series about its trend line (curve). Cyclical variations are seldom periodic and they may or may not follow same pattern after equal interval of time.

NOTES



In particular, business and economic time series are said to have cyclical variations if these variations recur after time interval of more than one year. In business and economic time seriés, *business cycles* are example of cyclical variations. There are four phases of a business cycle. These are :

(*a*) Depression (*b*) Recovery
(*c*) Boom (*d*) Decline.

These four phases of business cycle follows each other in this order.

(*a*) **Depression.** We start with the situation of depression in business cycle. In this phase, the employment is very limited. Employees get, very low wages. The purchasing power of money is high. This is the period of pessimism in business. New equilibrium is achieved in business at low level of cost, profit and prices.

(*b*) **Recovery.** The new equilibrium in the depression phase of a cycle; last for few years. This phase is not going to continue for ever. In the phase of depression, even efficient workers are available at very low wages. In the depression period, prices are low and the costs also too low. These factors replaces pessimism by optimism. Businessman, with good financial support is optimistic in such circumstances. He invests money in repairing plants. New plants are purchased. This also boost the business of allied industries. People get employment and spend money on consumers good. So, the situation changes altogether. This is called the phase of recovery in business cycle.

(*c*) **Boom.** There is also limit to recovery. Investment is revived in recovery phase. Investment in one industry affects investment in other industries. People get employment. Extension in demand is felt. Prices go high. Profits are made very easily. All these leads to over development of business. This phase of business cycle is described as *boom*.

(*d*) **Decline.** In the phase of boom, the business is over developed. This is because of heavy profits. Wages are increased and on the contrary their efficiency decreases. Money is demanded everywhere. This results in the increase in rate of interest. In other words, the demand for production factors increases very much and this results in increase in their prices. This results in the increases in the cost of production. Profits are decreased. Banks insists for repayment of loans under these circumstances. Businessmen give concession in prices so that cash may be secured. Consumers start expecting more reduction in prices. Condition become more worse. Products accumulates with businessmen and repayment of loan does not take place. Many business houses fails. All these leads to depression phase and the business cycle continues itself.

The length of a business cycle is in general between 3 to 10 years. Moreover, the lengths of business cycles are not equal.

6.7. IRREGULAR VARIATIONS

The **irregular variations** in a time series counts for those variations which cannot be predicted before hand. This component is different from the other three components in the sense that irregular variations in a time series are very irregular. Nothing can be predicted about the occurrence of irregular variations. It is very true that floods, famines, wars, earthquakes, strikes etc. do affect the economic and business activities.

The component *irregular variations* refers to the variations in time series which are caused due to the occurrence of events like flood, famine, war, earthquake, strike, etc.

6.8. ADDITIVE AND MULTIPLICATIVE MODELS OF DECOMPOSITION OF TIME SERIES

Let T, S, C and I represent the trend component, seasonal component, cyclical component and irregular component of a time series, respectively. Let the variable of the time series be denoted by Y. There are mainly two models of decomposition of time series.

1. Additive model. In this model, we have

$$Y = T + S + C + I.$$

In this case, the components T, S, C and I represent absolute values. Here S, C and I may admit of negative values. In this model, we assume that all the four components are independent of each other.

2. Multiplicative model. In this model, we have

$$Y = T \times S \times C \times I.$$

In this case, the components T is in abosolute value where as the components S, C and I represent relative indices with base value unity. In this model, the four components are not necessarily independent of each other.

6.9. DETERMINATION OF TREND

Before we go in the detail of methods of measuring secular trend, we must be clear about the purpose of measuring trend. We know that the secular trend is the tendency of time series to grow or to decline over a long period of time. By studying the trend line (or curve) of the profits of a company for a number of years, it can be well-decided as to whether the company is progressing or not. Similarly, by studying the trend of *consumer price index numbers*, we can have an idea about the rate of growth (or decline) in the prices of commodities.

We can also make use of trend characteristics in comparing the behaviour of two different industries in India. It can equally be used for comparing the growth of industries in India with those functioning in some other country.

The secular trend is also used for forecasting. This is achieved by projecting the trend line (curve) for the required future value.

The secular trend is also measured in order to eliminate itself from the given time series. After this, only three components are left and these are studied separately. The following are the methods of measuring the secular trend of a time series :

(*i*) Free Hand Graphic Method

(*ii*) Semi-Average Method

(*iii*) Moving Average Method

(*iv*) Least Squares Method.

6.10. FREE HAND GRAPHIC METHOD

NOTES

This is a graphic method. Let $\{(t_i, y_i) : i = 1, 2, ..., n\}$ be the given time series. On the graph paper, time is measured horizontally, whereas the values of the variable y are measured vertically. Points (t_1, y_1) (t_2, y_2), ..., (t_n, y_n) are plotted on the graph paper. These plotted points are joined by straight lines to get the graph of actual time series data.

In this method, trend line (or curve) is fitted by inspection. This is a subjective method. The trend line (or curve) is drawn through the graph of actual data so that the following are satisfied as far as possible :

(*i*) The algebraic sum of the deviations of actual values from the trend values is zero.

(*ii*) The sum of the squares of the deviations of actual values from the trend values is least.

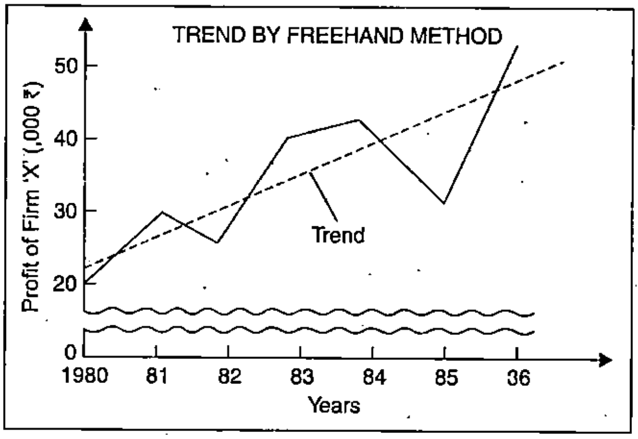
(*iii*) The area above the trend is equal to area below it.

(*iv*) The trend line (or curve) is smooth.

Example 1. *Fit a straight line trend to the following data, by using free hand graphic method :*

Year	*1980*	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*
Profit of Firm X ('000 ₹)	*20*	*30*	*25*	*40*	*42*	*30*	*50*

Solution.



6.10.1 Merits and Demerits

Merits

1. This is the simplest of all the methods of measuring trend.

2. This is a non-mathematical method and it can be used by any one who does not have mathematical background.

3. This method proves very useful for one who is well acquainted with the economic history of the concern, under consideration.

4. For rough estimates, this method is best suited.

Demerits

1. This method is not rigidly defined.
2. This method is not suited when accurate results are desired.
3. This is a subjective method and can be affected by the personal bias of the person, drawing it.

NOTES

EXERCISE 6.1

1. Fit a straight line trend to the following data by using free hand graphic method:

Year	1992	1993	1994	1995	1996
Import (in crores of ₹)	45	47	30	32	27

2. Fit a straight line trend to the following data by using free hand graphic method:

Year	1931	1941	1951	1961	1971	1981
Population of city X (in lakhs)	45	47	50	55	60	70

3. Fit a straight line trend to the following data by using free hand graphic method:

Year	1992	1993	1994	1995	1996	1997	1998
X	10	8	7	15	16	25	30

6.11. SEMI AVERAGE METHOD

This is a method of fitting trend line to the given time series. In this method, we divide the given values of the variable (*y*) into two parts. If the number of items is odd, then we make two equal parts by leaving the middle most value. And in case, the number of items is even, then we will not have to leave any item. After making two equal parts, the A.M. of both parts are calculated.

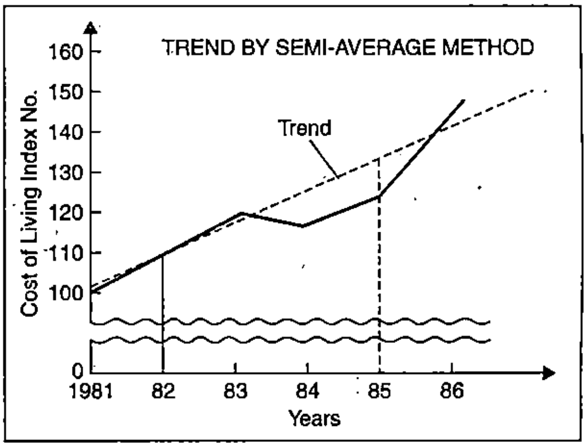
On graph paper. the graph of actual data is plotted. The A.M. of two parts are considered to correspond to the mid-points of the time interval considered in making the parts. The points corresponding to these averages of two parts are also plotted on the graph paper. These points are then joined by a straight line. This line represents the trend by semi average method. From the trend line, we can easily get the trend values. This trend line can also be used for predicting the value of the variable for any future period.

Example 2. *Fit a straight line trend to the following data by using semi average method :*

Year	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*
Cost of Living Index No.	*100*	*110*	*120*	*118*	*130*	*159*

Solution. **Trend line by Semi-Average Method**

Year	*Cost of Living Index*	*Year*	*Cost of Living Index*
1981	100	1984	118
1982	110 $\Big\}\ \frac{330}{3} = 110$	1985	130 $\Big\}\ \frac{407}{3} = 135.67$
1983	120	1986	159



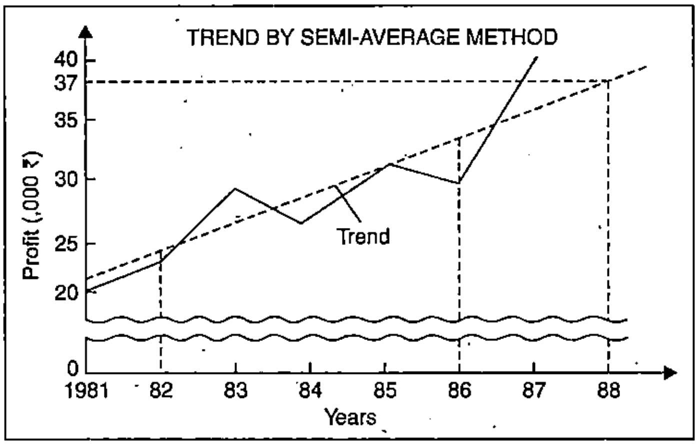
Example 3. *Fit a straight line trend by using the following data:*

Year	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*	*1987*
Profit ('000 ₹)	*20*	*22*	*27*	*26*	*30*	*29*	*40*

Semi-Average Method is to be used. Also estimate the profit for the year 1988.

Solution. **Trend Line by Semi-Average Method**

Year	*Profit (,000 ₹)*	*Year*	*Profit (,000 ₹)*
1981	20	1985	30
1982	22 $\Big\}\ \frac{69}{3} = 23$	1986	29 $\Big\}\ \frac{99}{3} = 33.$
1983	27	1987	40
1984	26		



The estimated profit for the year 1988 is ₹ **37000.**

6.11.1 Merits and Demerits

Merits

1. This method is rigidly defined.
2. This method is simple to understand.

Demerits

1. This method assumes a straight line trend, which is not always true.
2. Since this method is based on A.M., all the demerits of A.M. becomes the demerits of this method also.

NOTES

EXERCISE 6.2

1. Fit a straight line trend by the method of semi-average from the following data:

Year	1993	1994	1995	1996	1997	1998
Sales ('000 Units)	20	24	22	30	28	32

2. Fit a straight line trend for the following data, by using semi-average method:

Year	1980	1981	1982	1983	1984	1985	1986
Profit ('000 ₹)	80	82	85	70	89	95	105

3. Fit a straight line trend for the following data, by using semi-average method:

Year	1980	1981	1982	1983	1984	1985	1986	1987
Y	12	14	16	20	25	29	31	28

Also estimate the value of the variable Y for the year 1988.

4. Determine the trend of the following by semi-average method. Graph should be neat.

Year	*Sales (,000 ₹)*	*Year*	*Sales (,000 ₹)*
1965	18	1971	30
1966	25	1972	20
1967	21	1973	35
1968	15	1974	32
1969	26	1975	23
1970	31		

6.12. MOVING AVERAGE METHOD

Let $\{(t_i, y_i) : i = 1, 2, \ldots\ldots, n\}$ be the given time series. Here $y_1, y_2, \ldots\ldots, y_n$ are the values of the variable (y) corresponding to time periods $t_1, t_2, \ldots\ldots, t_n$ respectively.

We define **moving totals of order m** as $y_1 + y_2 + \ldots\ldots + y_m$, $y_2 + y_3 + \ldots\ldots + y_{m+1}$, $y_3 + y_4 + \ldots\ldots + y_{m+2}$, ...

The **moving averages of order m** are defined as

$$\frac{y_1 + y_2 + ... + y_m}{m}, \quad \frac{y_2 + y_3 + ... + y_{m+1}}{m}, \quad \frac{y_3 + y_4 + ... + y_{m+2}}{m}, ...$$

NOTES

These moving averages will be called **m yearly moving averages** if the values, $y_1, y_2, ... y_n$ of y are given annually. Similarly, if the data are given monthly, then the moving averages will be called **m monthly moving averages.**

In using moving averages in estimating the trend, we shall have to decide as to what should be the order of the moving averages. The order of the moving averages should be equal to the length of the cycles in the time series. In case, the order of the moving averages is given in the problem itself, then we shall use that order for computing the moving averages. The order of the moving averages may either be odd or even.

Let the order of moving averages be 3. The moving averages will be

$$\frac{y_1 + y_2 + y_3}{3}, \quad \frac{y_2 + y_3 + y_4}{3}, \quad \frac{y_3 + y_4 + y_5}{3}, ..., \frac{y_{n-2} + y_{n-1} + y_n}{3}.$$

These moving averages will be considered to correspond to 2nd, 3rd, 4th, ... $(n - 1)$th years respectively.

Similarly, the 5 yearly moving averages will be

$$\frac{y_1 + y_3 + y_3 + y_4 + y_5}{5}, \quad \frac{y_2 + ... + y_6}{5}, ..., \frac{y_{n-4} + ... + y_n}{5}.$$

These 5 yearly moving averages will be considered to correspond to 3rd, 4th, ..., ... $(n - 2)$th years respectively. These moving averages are called the trend values.

Calculation of trend values, by using moving averages of *even* order is slightly complicated. Suppose we are to find trend values by using 4 yearly moving averages. The 4 yearly moving averages are :

$$\frac{y_1 + y_2 + y_3 + y_4}{4}, \quad \frac{y_2 + y_3 + y_4 + y_5}{4}, ..., \frac{y_{n-3} + y_{n-2} + y_{n-1} + y_n}{4}.$$

These moving averages will not correspond to time periods, under consideration. The first moving average will correspond to the mid of t_2 and t_3. Similarly, others.

In order that these moving averages may correspond to original periods, we will have to resort to a process, called *centering of moving averages.* There are two methods of finding centered moving averages. Suppose we are to find 4 yearly centered moving averages for the times series :

$$\{(t_i, y_i)\} : i = 1, 2, ..., n\}.$$

6.12.1 Method I

In this method, we first calculate 4 yearly moving totals from the given data. Of these 4 year moving totals, 2 yearly moving totals are computed. These 2 yearly moving totals are then divided by 8 to get 4 yearly *centered moving averages*. These centered moving averages will correspond to 3rd, 4th, ... $(n - 2)$th years, in the table.

6.12.2 Method II

In this method, we first calculate 4 yearly moving averages. The first 4 yearly moving average will correspond to the mid of 2nd and 3rd years. Similarly, others. We now calculate 2 yearly moving averages of these 4 yearly moving averages. These averages will be 4 yearly *centered moving averages*. These averages will correspond to 3rd, 4th, ..., $(n-2)$th years, in the table.

NOTES

It may be carefully noted that the centered moving averages as calculated by using these methods will be exactly same.

In the moving average method of finding trend, the moving averages will be the trend values. These trend values may be plotted on the graph. The graph of the trend values will not be a straight line, in general.

Example 4. *Find the trend values by taking three yearly moving averages for the following data :*

Year	*1980*	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*	*1987*
Profit ('000 ₹)	*18*	*21*	*20*	*25*	*29*	*27*	*35*	*42*

Also find short term fluctuations assuming additive model.

Solution. **Trend by 3 year Moving Averages**

Year	*Profit ('000 ₹)*	*3 yearly moving total*	*Trend value 3 yearly moving average*
1980	18	—	—
1981	21	18 + 21 + 20 = 59	**19.667**
1982	20	21 + 20 + 25 = 66	**22**
1983	25	20 + 25 + 29 = 74	**24.667**
1984	29	25 + 29 + 27 = 81	**27**
1985	27	29 + 27 + 35 = 91	**30.333**
1986	35	27 + 35 + 42 = 104	**34.667**
1987	42	—	—

Short term fluctuations

Year	*Actual value* y	*Trend value* y_c	*Short term fluctuations* $y - y_c$
1980	18	—	—
1981	21	19.667	**1.333**
1982	20	22	**– 2**
1983	25	24.667	**0.333**
1984	29	27	**2**
1985	27	30.333	**– 3.333**
1986	35	34.667	**0.333**
1987	42	—	—

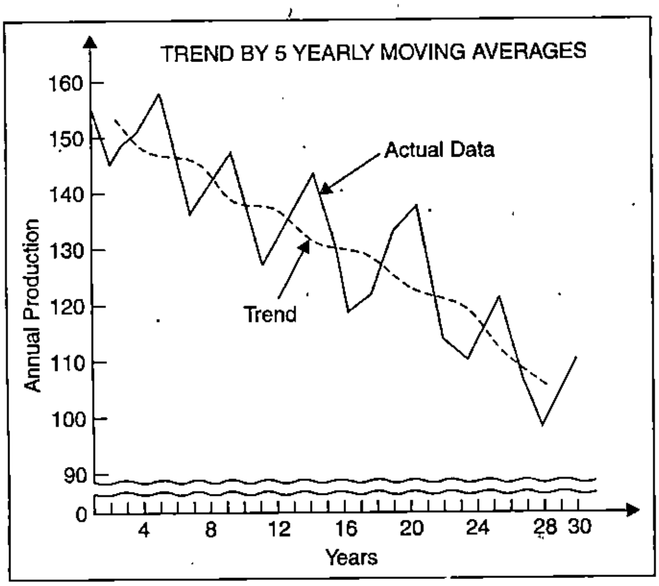
NOTES

Example 5. *From the following data, find trend, using five yearly moving averages and plot the results on a graph paper :*

Year	*Annual production*	*Year*	*Annual production*	*Year*	*Annual production*
1	*156*	*11*	*135*	*21*	*122*
2	*143*	*12*	*127*	*22*	*114*
3	*148*	*13*	*130*	*23*	*110*
4	*150*	*14*	*135*	*24*	*115*
5	*155*	*15*	*140*	*25*	*120*
6	*145*	*16*	*130*	*26*	*109*
7	*135*	*17*	*118*	*27*	*105*
8	*138*	*18*	*122*	*28*	*98*
9	*142*	*19*	*127*	*29*	*104*
10	*146*	*20*	*130*	*30*	*110*

Solution. **Trend by 5 yearly Moving Averages**

Year	*Annual Production*	*5 yearly moving total*	*5 yearly moving average i.e. trend value*
1	156	—	—
2	143	—	—
3	148	752	**150.4**
4	150	741	**148.2**
5	155	733	**146.6**
6	145	723	**144.6**
7	135	715	**143**
8	138	706	**141.2**
9	142	696	**139.2**
10	146	688	**137.6**
11	135	680	**136**
12	127	673	**134.6**
13	130	667	**133.4**
14	135	662	**132.4**
15	140	653	**130.6**
16	130	645	**129**
17	118	637	**127.4**
18	122	627	**125.4**
19	127	619	**123.8**
20	130	615	**123**
21	122	603	**120.6**
22	114	591	**118.2**
23	110	581	**116.2**
24	115	568	**113.6**
25	120	559	**111.8**
26	109	547	**109.4**
27	105	536	**107.2**
28	98	526	**105.2**
29	104	—	—
30	110	—	—



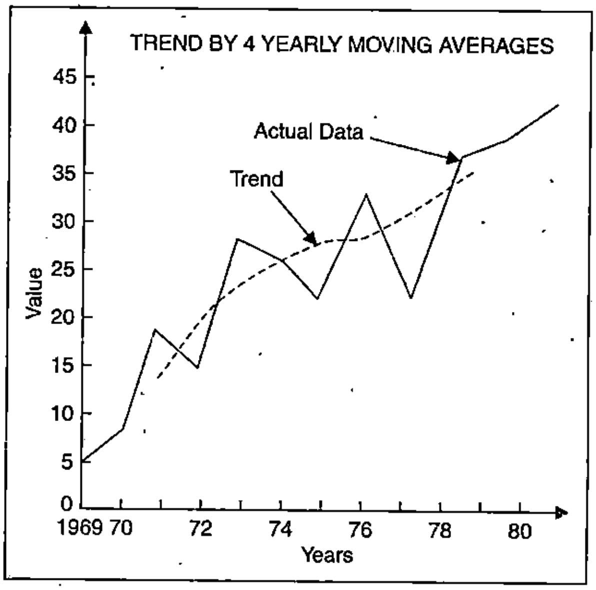
Example 6. *Estimate the trend values using the data given below by taking a four yearly moving average. Also plot the actual and trend values on the graph paper :*

Year	*Value*	*Year*	*Value*
1969	*4*	*1975*	*24*
1970	*7*	*1976*	*36*
1971	*20*	*1977*	*25*
1972	*15*	*1978*	*40*
1973	*30*	*1979*	*42*
1974	*28*	*1980*	*45*

Solution. **Trend by Moving Average Method**

Year	*Value*	*4 yearly moving total*	*4 yearly moving average*	*4 yearly centered moving average*
1969	4			—
1970	7			—
		46	11.5	
1971	20			**14.75**
		72	18	
1972	15			**20.625**
		93	23.25	
1973	30			**23.75**
		97	24.25	
1974	28			**26.875**
		118	29.5	
1975	24			**28.875**
		113	28.25	
1976	36			**29.75**
		125	31.25	
1977	25			**33.5**
		143	35.75	
1978	40			**36.875**
		152	38	
1979	42			—
1980	45			—

NOTES



Remark. In the above two examples, centering of moving averages has been done by adopting **method 1** and **method 2** respectively.

Example 7. *Calculate 4 yearly moving averages from the following :*

Year	*1972*	*1973*	*1974*	*1975*	*1976*	*1977*	*1978*	*1979*	*1980*	*1981*
Sale	*30*	*40*	*80*	*60*	*70*	*110*	*90*	*100*	*140*	*120*

Solution. **Trend by Moving Average Method**

Year	*Sale*	*4 yearly moving total*	*2 yearly moving totals of column 3*	*4 yearly centered moving average*
1972	30		—	—
1973	40		—	—
		210		
1974	80		460	**57.5**
		250		
1975	60		570	**71.25**
		320		
1976	70		650	**81.95**
		330		
1977	110		700	**87.5**
		370		
1978	90		810	**101.25**
		440		
1979	100		890	**111.25**
		450		
1980	140		—	—
1981	120		—	—

6.12.3 Merits and Demerits

Merits

1. This method is rigidily defined, so it cannot be affected by the personal prejudice of the person computing it.

2. If the order of the moving averages is exactly equal to the length of the cycle in the time series, the cyclical variations are eliminated.

3. If some more values of the variable are added at the end of the time series, the entire calculations are not changed.

4. This method is best suited for the time series whose trend is not linear. For such series, the general movement of the variable will be best shown by moving averages.

Demerits

1. Moving averages are strongly affected by the presence of extreme items, in the series.

2. It is difficult to decide the order of the moving averages, because the cycles in time series are seldom regular in duration.

3. In this method, we lose trend values at each end of the series. For example, if the order of the moving averages is five, we lose trend values for two years at each end of the series.

4. Forecasting is not possible in this method, because we cannot objectively project the graph of the trend values, for a future period.

EXERCISE 6.3

1. Calculate three yearly moving averages for the following data :

Year	1982	1983	1984	1985	1986
Production (in tonnes)	15	17	20	28	30

2. Find trend values for the following data, by using 5 yearly moving averages. Also plot the actual data and trend values on a graph :

Year	*Profit ('000 ₹)*	*Year*	*Profit ('000 ₹)*
1970	80	1977	90
1971	82	1978	92
1972	84	1979	97
1973	88	1980	95
1974	70	1981	99
1975	72	1982	80
1976	90	1983	99

3. Compute 3 yearly and 5 yearly moving averages of the following data. Also plot them on a graph paper along with original data :

Year	*Value*	*Year*	*Value*
1	1	11	3
2	2	12	4
3	3	13	5
4	2	14	4
5	1	15	3
6	2	16	4
7	3	17	5
8	4	18	6
9	3	19	5
10	2	20	4

4. The following table gives the number of workers employed in a small industry during 1980–87. Calculate the trend values by using 3 yearly moving averages :

Year	1980	1981	1982	1983	1984	1985	1986	1987
No. of workers	20	24	25	18	27	26	28	30

5. Compute 4 yearly and 5 yearly moving averages from the following data :

Year	*Value (in ₹)*	*Year*	*Value (in ₹)*
1960	365	1971	255
1961	360	1972	250
1962	355	1973	245
1963	330	1974	225
1964	300	1975	210
1965	330	1976	200
1966	340	1977	230
1967	290	1978	225
1968	280	1979	200
1969	250	1980	195
1970	235		

6. From the given data, compute trend by moving average method assuming 'a four-yearly cycle' Also find short term variations assuming additive model.

Year	*Sales*	*Year*	*Sales*
1984	75	1990	70
1985	60	1991	75
1986	55	1992	85
1987	60	1993	100
1988	65	1994	70
1989	70		

7. Assuming a four year cycle calcualate the trend by the method of moving averages from the following data :

Year	*Value*	*Year*	*Value*
1991	464	1996	540
1992	515	1997	557
1993	518	1998	571
1994	467	1999	586
1995	502	2000	612

Answers

1. 17.333, 21.667, 26
2. 80.8, 79.2, 80.8, 82, 82.8, 88.2, 92.8, 94.6, 92.6, 94
3. 3 yearly moving averages : 2, 2.33, 2, 1.67, 2, 3, 3.33, 3, 2.67, 3, 4, 4.33, 4, 3.67, 4, 5, 5.33, 5
   5 yearly moving averages : 1.8, 2, 2.2, 2.4, 2.6, 2.8, 3, 3.2, 3.4, 3.6, 3.8, 4, 4.2, 4.4, 4.6, 4.8
4. 23, 22.333, 23.333, 23.667, 27, 28
5. 4 yearly moving averages : 344.375, 332.5, 326.875, 320, 312.5, 300, 276.875, 259.375, 251.25, 246.875, 245, 238.125, 226.25, 218.125, 216.25, 215, 213.125

5 yearly moving averages : 342, 335, 331, 318, 308, 298, 279, 262. 254, 247, 242, 237, 226, 222, 218, 213, 210

6. 61.25, 61.25, 64.38, 68.13, 72.50, 78.75, 82.5 ; – 6.25, – 1.25, 0.62, 1.87, 2.50, – 3.75, 2.50

7. 495.75, 503.625, 511.625, 529.5, 553, 572.5.

6.13. LEAST SQUARES METHOD

This is a mathemetical method. Let $\{(t_i, y_i) : i = 1, 2, ..., n\}$ be the given time series. By using this method, we can find linear trend as well as non-linear trend of the corresponding data.

In this method, trend values (y_e) of the variable (y) are computed so as to satisfy the following two conditions :

(*i*) The sum of the deviations of values of y ($= y_1, y_2, ..., y_n$) from their corresponding trend values, is zero, *i.e.*, $\Sigma(y - y_e) = 0$.

(*ii*) The sum of the squares of the deviations of the values of y from their corresponding trend values is least *i.e.*, $\Sigma(y - y_e)^2$ is least.

On the graph paper, we shall measure the actual values and the estimated values (trend values) of the variable y, along the vertical axis. Let x denote the deviations of the time periods ($t_1, t_2, ..., t_n$) from some fixed time period. The fixed time period is called the *origin*.

6.14. LINEAR TREND

From the knowledge of coordinate geometry, we know that the equation of the required trend line can be expressed as

$$y_e = a + bx,$$

where a and b are constants. We have already mentioned that our trend line will satisfy the conditions :

(*i*) $\Sigma(y - y_e) = 0$ and (*ii*) $\Sigma(y - y_e)^2$ is least.

In order to meet these requirements, we will have to use those values of a and b in the trend line equation which satisfies the following *normal equations :*

$$\Sigma y = an + b\Sigma x \qquad ...(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad ...(2)$$

In the equation $y_e = a + bx$, of the trend, a represents the trend value of the variable when $x = 0$ and b represents the *slope* of the trend line. If b is positive, the trend will be upward and if b is negative, the trend of the time series will be downward.

It is very important to mention the origin and the x unit with the trend line equation. If either of the two is not given with the equation of the trend, we will not be able to get the trend values of the variable, under consideration.

Example 8. *Below are given the figures of production (in thousand maunds) of a sugar factory :*

Year	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*	*1987*
Production	*80*	*90*	*92*	*83*	*94*	*99*	*92*

(i) Find the slope of a straight line trend of these figures. Also find trend values.

(ii) Plot these figures on a graph and show the trend line.

(iii) Do these figures show a rising or a falling trend ?

Solution. We shall fit a straight line trend by the method of least squares. Let y denote the variable production (in thousand maunds).

Linear Trend by Least Squares Method

S. No.	*Year*	y	x = *Year – 1981*	x^2	xy
1	1981	80	0	0	0
2	1982	90	1	1	90
3	1983	92	2	4	184
4	1984	83	3	9	249
5	1985	94	4	16	376
6	1986	99	5	25	495
$n = 7$	1987	92	6	36	552
Total		630	21	91	1946

Let the equation of the trend line by $y = a + bx$.

$\therefore$ The normal equations are :

$$\Sigma y = an + b\Sigma x \quad \text{...(1)}$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2. \quad \text{...(2)}$$

(1) $\Rightarrow$ $630 = 7a + 21b$ ...(3)

(2) $\Rightarrow$ $1946 = 21a + 91b$ ...(4)

(3) × 3 $\Rightarrow$ $1890 = 21a + 63b$ ...(5)

(4) – (5) $\Rightarrow$ $56 = 28b$ $\Rightarrow$ $b = 2.$

$\therefore$ (3) $\Rightarrow$ $630 = 7a + 21(2)$ $\Rightarrow$ $a = 588/7 = 84.$

$\therefore$ The equation of trend is $y_e = 84 + 2x$, with origin 1981 and x unit = 1 year.

(i) The slope of straight line trend = **2.**

For 1981, $x = 1981 - 1981 = 0,$ $\therefore$ $y_e\,(1981) = 84 + 2(0) = \mathbf{84}$

For 1982, $x = 1982 - 1981 = 1.$ $\therefore$ $y_e\,(1982) = 84 + 2(1) = \mathbf{86}$

For 1983, $x = 1983 - 1981 = 2.$ $\therefore$ $y_e\,(1983) = 84 + 2(2) = \mathbf{88}$

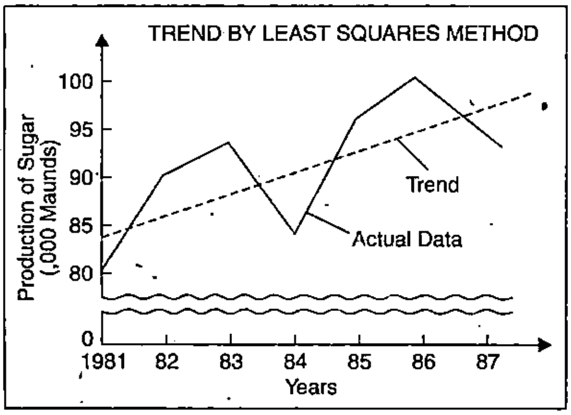
For 1984, $x = 1984 - 1981 = 3.$ $\therefore$ $y_e\,(1984) = 84 + 2(3) = \mathbf{90}$

For 1985, $x = 1985 - 1981 = 4.$ $\therefore$ $y_e\,(1985) = 84 + 2(4) = \mathbf{92}$

For 1986, $x = 1986 - 1981 = 5.$ $\therefore$ $y_e\,(1986) = 84 + 2(5) = \mathbf{94}$

For 1987, $x = 1987 - 1981 = 6.$ $\therefore$ $y_e\,(1987) = 84 + 2(6) = \mathbf{96.}$

(ii)



NOTES

(iii) The figures shows a *rising* trend.

Example 9. *Below are given figures of production (in thousand tonnes) of a sugar factory :*

Year	*1976*	*1977*	*1978*	*1979*	*1980*	*1981*	*1982*
Production	*77*	*88*	*94*	*85*	*91*	*98*	*90*

(i) *Fit a straight line trend by the method of least squares and calculate the trend values.*

(ii) *What is the monthly increase in production.*

(iii) *Eliminate the trend by assuming*

(a) *additive model* (b) *multiplicative model.*

Solution. Here the number of periods, 7 is odd.

We take the middle most period 1979 as the origin.

$\therefore$ $x = \text{year} - 1979$

Let y denotes the variable 'production (in thousand tonnes)'

Trend Line by Least Squares Method

S. No.	*Year*	*y*	*x = year – 1979*	x^2	*xy*
1	1976	77	–3	9	–231
2	1977	88	–2	4	–176
3	1978	94	–1	1	–94
4	1979	85	0	0	0
5	1980	91	1	1	91
6	1981	98	2	4	196
$n = 7$	1982	90	3	9	270
Total		623	0	28	56

Let the equation of trend line be $y_e = a + bx$.

The normal equations are :

$$\Sigma y = an + b\Sigma x \qquad ...(1)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2. \qquad ...(2)$$

NOTES

$$(1) \Rightarrow \qquad 623 = a(7) + b(0) \quad \Rightarrow \quad a = \frac{623}{7} = 89$$

$$(2) \Rightarrow \qquad 56 = a(0) + b(28) \quad \Rightarrow \quad b = \frac{56}{28} = 2.$$

∴ The equation of the trend line is y_e **= 89 + 2x, with origin 1979 and x unit = 1 year.**

Trend values

For 1976,	$x = -3$	∴	y_e (1976) = 89 + 2(– 3) = **83**
For 1977,	$x = -2.$	∴	y_e (1977) = 89 + 2(– 2) = **85**
For 1978,	$x = -1.$	∴	y_e (1978) = 89 + 2(– 1) = **87**
For 1979,	$x = 0$	∴	y_e (1979) = 89 + 2(0) = **89**
For 1980,	$x = 1.$	∴	y_e (1980) = 89 + 2(1) = **91**
For 1981,	$x = 2.$	∴	y_e (1981) = 89 + 2(2) = **93**
For 1982,	$x = 3.$	∴	y_e (1982) = 89 + 2(3) = **95.**

(*ii*) Value of 'b' = 2

∴ Annual increase of production = 2 × 1000 tonnes = 2000 tonnes

because the unit of y is thousand tonnes

∴ Monthly increase in production = $\frac{2000}{12}$ tonnes = **166.67 tonnes.**

(*iii*) (*a*) Time series model is **additive.**

∴ The trend eliminated values are given by $y - y_e$.

Year	y	y_e	*Trend eliminated value* $y - y_e$
1976	77	83	77—83 = **– 6**
1977	88	85	88—85 = **3**
1978	94	87	94—87 = **7**
1979	85	89	85—89 = **– 4**
1980	91	91	91—91 = **0**
1981	98	93	98—93 = **5**
1982	90	95	90—95 = **– 5**

(*b*) Trend series model is *multiplicative.*

∴ The trend eliminated values are given by y/y_e.

Year	y	y_e	*Trend eleminated value* y/y_e
1976	77	83	77 ÷ 83 = **0.928**
1977	88	85	88 ÷ 85 = **1.035**
1978	94	87	94 ÷ 87 = **1.080**
1979	85	89	85 ÷ 89 = **0.955**
1980	91	91	91 ÷ 91 = **1.000**
1981	98	93	98 ÷ 93 = **1.054**
1982	90	95	90 ÷ 95 = **0.947**

Example 10. *The number of units of a product exported during 1980 – 1987 are given below. Fit a straight line trend to the data. Plot the data showing also the trend line. Find an estimate for 1988.*

Year	*1980*	*1981*	*1982*	*1983*	*1984*	*1985*	*1986*	*1987*
No. of units (in thousands)	*12*	*13*	*13*	*16*	*19*	*23*	*21*	*23*

NOTES

Solution. Here the number of periods is eight. We take $\frac{1983 + 1984}{2} = 1983.5$ as the origin. In order to avoid decimals in the deviations, we define

$$x = (\text{year} - 1983.5)2.$$

Let y denote the variable 'no. of units in thousands'

Trend Line by Least Squares Method

S. No.	Year	y	x	x^2	xy
1	1980	12	−7	49	−84
2	1981	13	−5	25	−65
3	1982	13	−3	9	−39
4	1983	16	−1	1	−16
5	1984	19	1	1	19
6	1985	23	3	9	69
7	1986	21	5	25	105
$n = 8$	1987	23	7	49	161
Total		140	0	168	150

Let the equation of the trend line be $y_e = a + bx$.

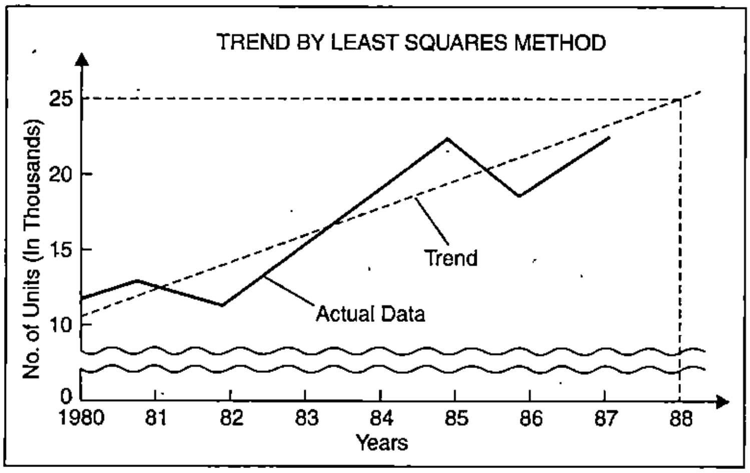
The normal equations are :

$$\Sigma y = an + b\Sigma x \qquad ...(1)$$

and $$\Sigma xy = a\Sigma x + b\Sigma x^2. \qquad ...(2)$$

(1) ⇒ $140 = 8a + b.0 \Rightarrow a = 140/8 = 17.5$

(2) ⇒ $150 = a.0 + 168b \Rightarrow b = 150/168 = 0.89.$



∴ The equation of trend line is **y_e = 17.5 + 0.89x with origin = 1983.5 and x unit = 1/2 year.**

For 1988, $x = (1988 - 1983.5)2 = 4.5 \times 2 = 9.$

∴ The estimate value of $y(1988) = 17.5 + (0.89)9 =$ **25.51 thousand units.**

NOTES

Example 11. *Find trend values by least squares method :*

Year	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
Production	*102.3*	*101.9*	*105.8*	*112.0*	*114.8*	*118.7*	*124.5*	*102.9*

Solution. Here the number of periods is even, we take $\frac{1988 + 1989}{2} = 1988.5$ as the origin. In order to avoid decimals in the deviations, we take x = (year – 1988.5)2.

Let y denote the variable 'production'.

Trend Line by Least Squares Method

S. No.	Year	y	x	x^2	xy
1	1985	102.3	– 7	49	– 716.1
2	1986	101.9	– 5	25	– 509.5
3	1987	105.8	– 3	9	– 317.4
4	1988	112.0	– 1	1	– 112.0
5	1989	114.8	1	1	114.8
6	1990	118.7	3	9	356.1
7	1991	124.5	5	25	622.5
$n = 8$	1992	102.9	7	49	720.3
Total		882.9	0	168	158.7

Let the equation of the trend line by $y_e = a + bx$.

∴ The normal equations are :

$$Sy = an + bSx \quad ...(1)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad ...(2)$$

$$(1) \Rightarrow \quad 882.9 = a(8) + b(0) \quad \Rightarrow \quad a = \frac{882.9}{8} = 110.363$$

$$(2) \Rightarrow \quad 158.7 = a(0) + b(170) \quad \Rightarrow \quad b = \frac{158.7}{168} = 0.944$$

∴ The equation of the trend line is $y_e = 110.363 + 0.944x$ with origin = 1988.5 and x unit = 1/2 year.

Trend values

For 1985, $x = -7$. ∴ y_e (1985) = 110.363 + (0.944)(– 7) = **103.755**

For 1986, $x = -5$. ∴ y_e (1986) = 110.363 + (0.944)(– 5) = **105.643**

For 1987, $x = -3$. ∴ y_e (1987) = 110.363 + (0.944)(– 3) = **107.531**

For 1988, $x = -1$. ∴ y_e (1988) = 110.363 + (0.944)(– 1) = **109.419**

For 1989, $x = 1$. ∴ y_e (1989) = 110.363 + (0.944)(1) = **111.307**

For 1990, $x = 3$. ∴ y_e (1990) = 110.363 + (0.944)(3) = **113.195**

For 1991, $x = 5$. ∴ y_e (1991) = 110.363 + (0.944)(5) = **115.083**

For 1992, $x = 7$. ∴ y_e (1992) = 110.363 + (0.944)(7) = **116.971.**

EXERCISE 6.4

NOTES

1. Fit a straight line trend and find trend values for the following data by the method of least squares :

Year	1990	1991	1992	1993	1994
Profit *(in thousand Rupees)*	4	7	3	6	8

2. The production figures of a sugar factory are given below. Fit a straight line trend by the method of least squares and draw it on a graph paper along with the actual production figures :

Year	1951	1952	1953	1954	1955	1956	1957
Production *(in thousand quintals)*	80	90	92	83	94	99	92

3. Compute the straight line trend for the following data by using the method of least squares and show it graphically :

Year	1973	1974	1975	1976	1977	1978	1979	1980
Production *(million tonnes)*	38	40	65	72	69	60	87	95

4. Fit a straight line trend by the method of least squares for the following data :

Year	*Milk consumption (million gallons)*	*Year*	*Milk consumption (million gallons)*
1960	102.3	1965	118.7
1961	101.9	1966	124.5
1962	105.8	1967	129.9
1963	112.0	1968	134.8
1964	114.8		

5. From the following data, determine the long-term trend, using the method of least squares :

Year	1963	1964	1965	1966	1967	1968	1969	1970
Income (in Lakh ₹)	38	40	65	72	69	87	95	106

6. Compute the trend values by the method of least squares from the data given below :

Year	*Output*	*Year*	*Output*
1972	5600	1976	4200
1973	5500	1977	3800
1974	5100	1978	3500
1975	4700	1979	3200

7. Calculate trend values by the method of least squares and estimate production for the year 1995 :

Year	1984	1985	1986	1987	1988	1989	1990
Production of steel (in 10 lakh tonnes)	60	72	75	65	80	85	95

8. Fit a straight line trend to the following data on the domestic demand for motor fuel :

Year	*Average monthly demand (million barrels)*	*Year*	*Average monthly demand (million barrels)*
1978	61	1984	96
1979	66	1985	100
1980	72	1986	103
1981	76	1987	110
1982	82	1988	114
1983	90		

9. From the data given below, fit a curve of the type $y = a + bx$:

Year	1980	1981	1982	1983	1984
Population (in ,000's)	132	142	157	174	191

10. Fit a trend line by least squares method to the following data :

Year	1997	1998	1999	2000	2001	2002	2003
Production ('000 tonnes)	70	75	90	91	95	98	100

What is the rate of growth of production?

11. Find the trend values by using least squares method. Also find the trend value for 2005.

Year	1995	1996	1997	1998	1999	2000	2001
Production (in Qtl.)	700	743	816	834	907	860	961

Answers

1. $y_e = 5.6 + 0.7x$, where origin = 1992 and x unit = 1 year. Trend values : 4.2, 4.9, 5.6, 6.3, 7
2. $y_e = 84 + 2x$, with origin = 1951 and x unit = 1 year
3. $y_e = 65.75 + 7.3333\ x$, with origin = 1976.5, x unit = 1 year
4. $y_e = 116.0778 + 4.3017\ x$ with origin = 1964 and x unit = 1 year
5. $y_e = 71.5 + 9.69\ x$, where origin = 1966.5 and x unit = 1 year
6. 5750, 5378.57, 5007.14, 4635.71, 4264.28, 3892.85 3521.42, 3149.99
7. 61.429, 66.286, 71.143, 76, 80.857, 85.714, 90.571 ; 114.856
8. $y_e = 88.18 + 5.418\ x$, where origin = 1983 and x unit = 1 year.
9. $y_e = 159.2 + 15\ x$, where origin = 1982 and x unit = 1 year
10. $y_e = 88.429 + 5.036\ x$, where origin = 2000 and x unit = 1 year ; 0.428
11. 712.86 Qtl., 752.43 Qtl., 792 Qtl., 831.57 Qtl., 871.14 Qtl., 910.71 Qtl., 950.28 Qtl., Estimated production of 2005 = 1108.56 Qtl.

6.15. NON-LINEAR TREND (PARABOLIC)

NOTES

There are situations where linear trend is not found suitable. Linear trend is suitable when the tendency of the actual data is to move approximately in one direction. There are number of curves representing non-linear trend. In the present section, use shall consider parabolic trends. Parabolic trends will give better trend then the straight line trends.

Let $\{(t_i, y_i) : i = 1, 2, \ldots\ldots, n\}$ be the given time series. Let x denote the deviations of the time periods $(t_1, t_2 \ldots\ldots, t_n)$ from some fixed time period, called the origin. Let y_e denote the estimated (trend) values of the variable.

Let the equation of the required parabolic trend curve be

$$y_e = a + bx + cx^2$$

where, a, b, c are constants. This trend curve will satisfy the conditions :

(*i*) $\Sigma(y - y_e) = 0$

(*ii*) $\Sigma(y - y_e)^2$ is least.

In order to meet these requirements, we will have to use those values of a, b and c in the trend curve equation which satisfies the following *normal equations* :

$$\Sigma y = an + b\Sigma x + c\Sigma x^2 \quad \ldots(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \ldots(2)$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4. \quad \ldots(3)$$

Here also, it is very important to mention the origin and the x unit with the trend curve equation.

There is no specific rule for choosing the origin. But if we manage to choose the origin so as to make $\Sigma x = 0$, then we shall be reducing the calculation involved in computing a, b and c. In case the time periods $t_1, t_2, \ldots\ldots t_n$ advances by equal intervals and $\Sigma x = 0$, then we will also have $\Sigma x^3 = 0$. Here, the normal equations will reduce to :

$$\Sigma y = an + b.0 + c\Sigma x^2$$

$$\Sigma xy = a.0 + b\Sigma x^2 + c.0$$

$$\Sigma x^2 y = a\Sigma x^2 + b\ .0 + c\Sigma x^4$$

or

$$\Sigma y = an + c\Sigma x^2 \quad \ldots(1')$$

$$\Sigma xy = b\Sigma x^2 \quad \ldots(2')$$

$$\Sigma x^2 y = a\Sigma x^2 + c\Sigma x^4. \quad \ldots(3')$$

(2′) $\Rightarrow$ $b = \Sigma xy/\Sigma x^2$. The values of a and c will be obtained by solving the equations (1′) and (3′).

Example 12. *The prices of commodities during 1978 – 1983 are given below. Fit a parabola $y = a + bx + cx^2$ to this data. Estimate the price of commodity for the year 1984.*

Year	*1978*	*1979*	*1980*	*1981*	*1982*	*1983*
Price	*100*	*107*	*128*	*140*	*181*	*192*

Also plot actual and trend values on the graph.

NOTES

Sol. Here the number of periods is six. Therefore, we take $\frac{1980 + 1981}{2} = 1980.5$ as the origin. In order to avoid decimals in the deviations, we define

$$x = (\text{year} - 1980.5)2.$$

Let y denote the variable 'price'.

Parabolic Trend by Least Squares Method

S. No.	*Year*	y	x	x^2	x^3	x^4	xy	x^2y
1	1978	100	– 5	25	– 125	625	– 500	2500
2	1979	107	– 3	9	– 27	81	– 321	963
3	1980	128	– 1	1	– 1	1	– 128	128
4	1981	140	1	1	1	1	140	140
5	1982	181	3	9	27	81	543	1629
$n = 6$	1983	192	5	25	125	625	960	4800
Total		848	0	70	0	1414	694	10160

Let the equation of parabolic trend be

$$y_e = a + bx + cx^2.$$

The normal equations are :

$$\Sigma y = an + b\Sigma x + c\Sigma x^2 \qquad ...(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \qquad ...(2)$$

$$\Sigma x^2y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \qquad ...(3)$$

or $\quad 848 = 6a + b.0 + 70c$

$694 = a.0 + 70b + c.0$

$10160 = 70a + b.0 + 1414c$

or $\quad 848 = 6a + 70c \qquad ...(4)$

$694 = 70b \qquad ...(5)$

$10160 = 70a + 1444c \qquad ...(6)$

(5) $\Rightarrow$ $b = 694/70 = 9.914$

(4) × 35 $\Rightarrow$ $29680 = 210a + 2450c \qquad ....(7)$

(6) × 3 $\Rightarrow$ $30480 = 210a + 4242c \qquad ...(8)$

(7) – (8) $\Rightarrow$ $-800 = 0 - 1792c \quad \Rightarrow \quad c = 0.446.$

∴ (4) $\Rightarrow$ $848 = 6a - 70\ (0.446) \quad \Rightarrow \quad a = 136.13.$

∴ The equation of parabolic trend is

$\boldsymbol{y_e = 136.13 + 9.914x + 0.446x^2}$ **with origin = 1980.5 and** $\boldsymbol{x}$ **unit =** $\boldsymbol{\frac{1}{2}}$**year.**

Trend values

For 1978, $\quad x = -5$

∴ $\quad y_e\ (1978) = 136.13 + 9.914\ (-5) + (0.446)\ (-5)^2 = 97.710$

For 1979, $\quad x = -3.$

∴ $\quad y_e\ (1979) = 136.13 + 9.914\ (-3) + (0.446)\ (-3)^2 = 110.402$

For 1980, $\quad x = -1.$

∴ $\quad y_e\ (1980) = 136.13 + 9.914\ (-1) + (0.446)\ (-1)^2 = 126.662$

For 1981, $\quad x = 1.$

∴ $\quad y_e\ (1981) = 136.13 + 9.914\ (1) + (0.446)\ (1)^2 = 146.490$

For 1982, $x = 3$

$\therefore$ y_e (1982) = 136.13 + 9.914 (3) + (0.446) $(3)^2$ = 169.886.

For 1983, $x = 5$.

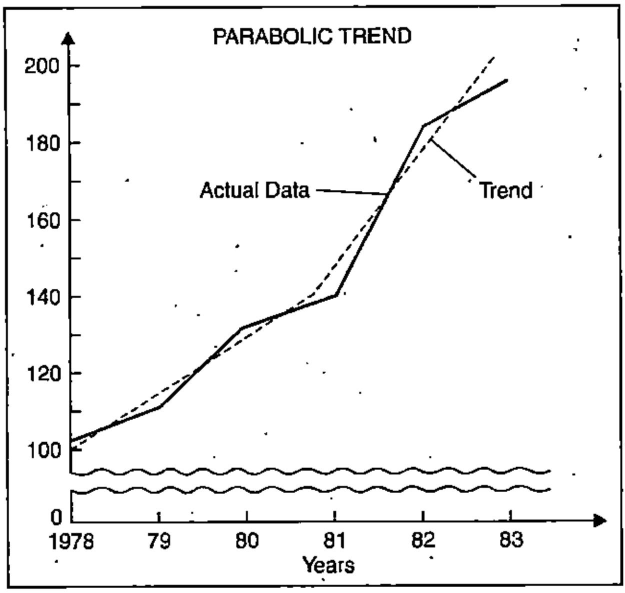
$\therefore$ y_e (1983) = 136.13 + 9.914 (5) + (0.446) $(5)^2$ = 196.85

For 1984, $x = 2\,(1984 - 1980.5) = 7$.

The estimated value of 'price' for 1984

$= y_e$ (1984) = 136.13 + 9.914(7) + (0.446) $(7)^2$ = **₹ 227.382.**

The graph of actual and trend values is given below :



EXERCISE 6.5

1. Find the equation of parabolic trend of second degree to the following data :

Year	1983	1984	1985	1986	1987
Variable	5	7	4	9	12

2. Estimate the value of y for the year 1979, by using the following data :

Year	1980	1981	1982	1983	1984
Variable (y)	10	12	13	10	8

For estimation, a parabolic trend is to be used.

3. Fit a straight line trend and a parabolic trend to the following data :

Year	1993	1994	1995	1996	1997	1998	1999
y	12	14	12	26	42	40	50

Answers :

1. $y_e = 5.97144 + 1.6x + 0.71428x^2$, where origin = 1985 and x unit = 1 year.
2. 6.4003.
3. $y_e = 25.9 + 7x + 0.524x^2$ where origin = 1996 and x unit = 1 year.

6.16. NON-LINEAR TREND (EXPONENTIAL)

NOTES

In this section, we shall study the method of finding non-linear exponential trend of a given time series.

Let $\{(t_i, y_i) : i = 1, 2, ..., n\}$ be the given time series. Let x denote the deviations of the time periods $\{t_1, t_2, ..., t_n\}$ from some fixed time period, called the origin. Let y_e denote the estimated (trend) values of the variable.

Let the equation of the required exponential trend curve be

$$y_e = ab^x \qquad ...(1)$$

where a, b are constants.

$$(1) \Rightarrow \qquad \log y_e = \log a + x \log b. \qquad ...(2)$$

The exponential trend curve will satisfy the conditions :

(*i*) $\Sigma(\log y - \log y_e) = 0$

(*ii*) $\Sigma(\log y - \log y_e)^2$ is least.

In order to meet these requirements we will have to use those values of a and b in the trend curve equation which satisfies the following *normal equations* :

$$\Sigma\log y = (\log a)n + (\log b)\Sigma x \qquad ...(3)$$

$$\Sigma\, x \log y = (\log a)\, \Sigma x + (\log b)\, \Sigma x^2. \qquad ...(4)$$

Here also, it is very important to mention the origin and the x unit with the trend curve equation.

If origin be chosen so that $\Sigma x = 0$, then the above normal equations reduces to

$$\Sigma \log y = (\log a)n + (\log b).0$$

and $$\Sigma x \log y = (\log a).0 + (\log b)\, \Sigma x^2.$$

$$\therefore \qquad \log a = \frac{\Sigma \log y}{x} \quad \text{and} \quad \log b = \frac{\Sigma x \log y}{\Sigma x^2}.$$

$$\therefore \qquad \mathbf{a = AL\left(\frac{\Sigma log\, y}{n}\right) \quad and \quad b = AL\left(\frac{\Sigma x\ log\, y}{\Sigma x^2}\right)}.$$

In practical problems, we prefer to choose origin in such a way that $\Sigma x = 0$. This will facilitate the computation of constants a and b.

Example 13. *The sales of a company in lakhs of rupees for the years 1996 to 2002 are given below :*

Year	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*
Sales (in lakh ruppes)	*16*	*23*	*33*	*46*	*66*	*95*	*137*

Estimate the sales for the year 2003 using an exponential trend curve.

Solution. Here the number of periods is equal to seven, an odd number.

∴ We take 1999 (the middle most period) as the origin.

Exponential Trend by Least Squares Method

NOTES

S. No.	Year	Sales (in lakhs rupees) y	$\log y$	x = year-1999	x^2	$x \log y$
1	1996	16	1.2041	– 3	9	– 3.6123
2	1997	23	1.3617	– 2	4	– 2.7234
3	1998	33	1.5185	– 1	1	– 1.5185
4	1999	46	1.6628	0	0	0
5	2000	66	1.8195	1	1	1.8195
6	2001	95	1.9777	2	4	3.9554
7	2002	137	2.1367	3	9	6.4101
$x = 7$			$\Sigma \log y = 11.6810$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma x \log y = 4.3308$

Let the equation of the exponential trend be $y_e = ab^x$.

$\therefore \quad \log y_e = \log a + x \log b \quad$ ...(1)

The normal equations are :

$$\Sigma \log y = (\log a)n + (\log b)\Sigma x \quad ...(2)$$

and $$\Sigma x \log y = (\log a)\Sigma x + (\log b)\Sigma x^2. \quad ...(3)$$

(2) $\Rightarrow \quad 11.6810 = 7 \log a + (\log b).\ 0$

$\Rightarrow \quad \log a = \dfrac{11.6810}{7} = 1.6687$

(3) $\Rightarrow \quad 4.3308 = (\log a).0 + (\log b)\ .\ 28$

$\Rightarrow \quad \log b = \dfrac{4.3308}{28} = 0.1547$

$\therefore$ (1) $\Rightarrow \quad \log y_e = 1.6687 + 0.1547x.$

For the year 2003, $x = 2003 - 1999 = 4.$

$\therefore$ For 2003, $\log y_e = 1.6687 + (0.1547)4 = 2.2875$

$\therefore \quad y_e = \text{AL}\ (2.2875) = 193.8$

$\therefore$ For the year 2003, the estimated sales are ₹ **193.8 lakh.**

EXERCISE 6.6

1. Fit an exponential trend to the following data :

Year	1998	1999	2000	2001	2002
y	1.6	4.5	13.8	40.2	135.0

2. Fit an exponential trend to the following data :

Year	1997	1998	1999	2000	2001
y	3.2	9.0	27.6	80.4	250.0

NOTES

3. Given the following population figures of India, estimate the population for 1991 and 2001, using exponential trend :

Year	1921	1931	1941	1951	1961	1971	1981
Population (in crores)	25:1	27.9	31.9	36.1	43.9	54.8	68.5

Answers

1. $y = 13.79\ (2.977)^x$, where x = year – 2000
2. $y = 27.54\ (2.98)^x$, where x = year – 1999.
3. $y = 38.80\ (1.19)^x$, where $x = \dfrac{\text{year} - 1951}{10}$, 77.81 crores, 92.59 crores.

EXERCISE 6.7

1. Define a time series. Explain its utility in Business and Economics.
2. What is secular trend ? Critically examine various methods of measuring trend.
3. Discuss the superiority of least square method over moving average method, in estimating secular trend of time series.
4. What is meant by a trend ? How do you fit a straight line trend by the method of least squares.
5. Distinguish between 'Free Hand Graphic Method' and 'Simi-Average Method' of estimating trend of a time series.

6.17. SUMMARY

- The general tendency of the values of the variable in a time series to grow or to Jecline over a long period of time is called secular trend of the time series. It indicates the general direction in which the graph of the time series aoppears to be going over a long period of time. The gaph of the secular trend is either a straight line or a curve.
- The **seasonal variations** in a time series counts for those variations in the series which occur annually. In a time series, seasonal variations occurs quite regularly. These variations play a very important role in business activities.
- The **cyclical variations** in a time series counts for the swings of graph of time series about its trend line (curve).
- The **irregular variations** in a time series counts for those variations which cannot be predicted before hand. This component is different from the other three components in the sense that irregular variations in a time series are very irregular.

7. MEASURES OF CORRELATION

STRUCTURE

7.1. CORRELATION AND CAUSATION

Two variables may be related in the sense that the changes in the values of one variable are accompanied by changes in the values of the other variable. But this cannot be interpreted in the sense that the changes in one variable has necessarily caused changes in the other variable. Their movement in sympathy may be due to mere chance. A high degree correlation between two variables may not necessarily imply the existence of a cause-effect relationship between the variables. On the other hand, if there is a cause-effect relationship between the variables, then the correlation is sure to exist between the variables under consideration. A high degree correlation between 'income' and 'expenditure' is due to the fact that expenditure is affected by the income.

Now we shall outline the reasons which may be held responsible for the existence of correlation between variables.

The correlation between variables may be due to the effect of some common cause. For example, positive correlation between the number of girls seeking admission in colleges A and B of a city may be due to the effect of increasing interest of girls towards higher education. The correlation between variables may be due to mere chance. Consider the data regarding six students selected at random from a college.

Students	*A*	*B*	*C*	*D*	*E*	*F*
% of marks obtained in the previous exam.	42%	47%	60%	80%	55%	40%
Height (in inches)	60	62	65	70	64	59

Here the variables are moving in the same direction and a high degree of correlation is expected between the variables. We cannot expect this degree of correlation to hold good for any other sample drawn from the concerned population. In this case, the correlation has occurred just due to chance.

The correlation between variables may be due to the presence of some cause-effect relationship between the variables. For example, a high degree correlation between 'temperature' and 'sale of coffee' is due to the fact that people like taking coffee in the winter season.

The correlation between variables may also be due to the presence of interdependent relationship between the variables. For example, the presence of correlation between amount spent on entertainment of family and the total expenditure of family is due to the fact that both variables effects each other. Similarly, the variables, 'total sale' and 'advertisement expenses' are interdependent.

TYPES OF CORRELATION

Correlation is classified in the following ways:

(*i*) Positive and Negative Correlation.

(*ii*) Linear and Non-linear Correlation.

(*iii*) Simple, Multiple and Partial Correlation.

7.2. POSITIVE AND NEGATIVE CORRELATION

The correlation between two variables is said to be **positive** if the variables, on an average, move in the same direction. That is, an increase (or decrease) in the value of one variable is accompanied, on an-average, by an increase (or decrease) in the value of the other variable. We do not stress that the variables should move strictly in the same direction. For example, consider the data :

x	2	3	6	8	11
y	14	15	13	18	22

Here the values of y has increased corresponding to every increasing value of x, except for $x = 6$. The correlation between the variables x and y is positive.

The correlation between two variables is said to be **negative** if the variables, on an average, move in the opposite directions. That is, an increase (or decrease) in the value of one variable is accompanied, on an average, by a decrease (or increase) in the value of the other variable.

Here also, we do not stress that the variables should move strictly in the opposite directions. For example, consider the data:

x	110	107	105	95	80
y	8	15	14	27	36

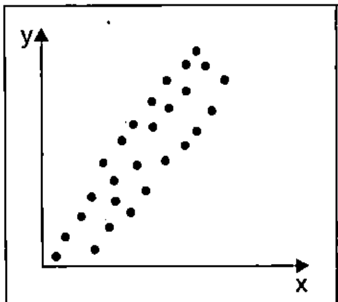
Here, a decrease in the value of x is accompanied by an increase in the value of y, except for $x = 105$. The correlation between x and y is negative.

Thus, we see that the correlation between two variables is positive or negative according as the movements in the variables are in same direction or in the opposite directions, on an average.

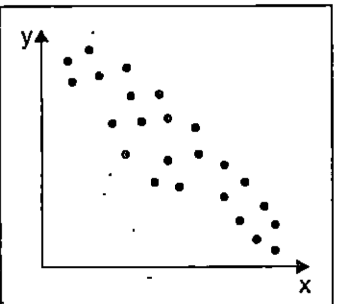
7.3. LINEAR AND NON-LINEAR CORRELATION

The correlation between two variables is said to be **linear** if the ratio of change in one variable to the change in the other variable is almost constant. The correlation between the 'number of students' admitted and the 'monthly fee collected' is linear in nature. Let x and y be two variables such that the ratio of change in x to the change in y is almost constant and if a scatter diagram is prepared corresponding to the variables x and y, the points in the diagrams would be almost along a line.

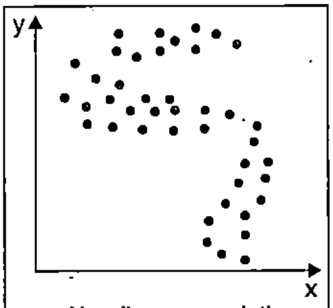
The extent of linear correlation is found by using Karl Pearson's method, Spearman's rank correlation method and concurrent deviation method.



Positive linear correlation



Negative linear correlation



Non-linear correlation

The correlation between two variables is said to be **non-linear** if the ratio of change in one variable to the change in the other variable is not constant. The correlation between 'profit' and 'advertisement expenditure' of a company is non-linear, because if the expenditure on advertisement is doubled, the profit may not be doubled. Let x and y be two variables in which the ratio of change in x to the change in y is not constant and if a scatter diagram is drawn corresponding to the data, the points in the diagram would not be having linear tendency.

7.4. SIMPLE, MULTIPLE AND PARTIAL CORRELATION

The correlation is said to be **simple** if there are only two variables under consideration. The correlation between sale and profit figures of a departmental store is simple. If there are more than two variables under consideration, then the correlation is either multiple or partial. Multiple and partial coefficients of correlation are called into play when the values of one variable are influenced by more than one variable. For example, the expenditure of salaried class of people may be influenced by their monthly incomes, secondary sources of income, legacy (money etc. handed down from ancestors) etc. If we intend to find the effect of all these variables on the expenditure of families, this will be a problem of multiple correlation. In **multiple correlation,** the combined effect of a number of variables on a variable is considered. Let x_1, x_2, x_3 be three variables, then $R_{1.23}$ denotes the multiple correlation coefficient of x_1 on x_2 and x_3. Similarly $R_{2.31}$ denotes the multiple correlation coefficient of x_2 on x_3 and x_1. In **partial correlation,** we study the relationship between any two variables, from a group of more than two variables, after eliminating the effect of other variables mathematically on the variables under consideration. Let x_1, x_2, x_3 be three variables, then $r_{12.3}$ denotes the partial correlation coefficient between x_1 and x_2. Similarly, $r_{13.2}$ denotes the partial correlation coefficient between x_1 and x_3. The methods of computing multiple and partial correlation coefficients are beyond the scope of this book. Thus, we shall be discussing the methods of computing only simple correlation coefficient.

KARL PEARSON'S METHOD

7.5. DEFINITION OF CORRELATION

Let (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) be n pairs of values of two variables x and y with respect to some characteristic (time, place etc.). The Karl Pearson's method is used to study the presence of *linear correlation* between two variables. The Karl Pearson's coefficient of correlation, denoted by $r(x, y)$ is defined as :

$$r(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

or simply, $$\mathbf{r} = \frac{\Sigma(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})}{\sqrt{\Sigma(\mathbf{x} - \bar{\mathbf{x}})^2} \sqrt{\Sigma(\mathbf{y} - \bar{\mathbf{y}})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the A.M.'s of x-series and y-series respectively.

This is called the *direct method* of computing Karl Pearson's coefficient of correlation.

If there is no chance of confusion, we write $r(x, y)$, just as r.

It can be proved mathematically that $-1 \le r \le 1$.

If the correlation between the variables is *linear*, then the value of Karl Pearson's coefficient of correlation is interpreted as follows :

Value of 'r'	*Degree of linear correlation between the variables*
$r = +1$	Perfect positive correlation
$0.75 \le r < 1$.	High degree positive correlation
$0.50 \le r < 0.75$	Moderate degree positive correlation
$0 < r < 0.50$	Low degree positive correlation
$r = 0$	No correlation
$-0.50 < r < 0$	Low degree negative correlation
$-0.75 < r \le -0.50$	Moderate degree negative correlation
$-1 < r \le -0.75$	High degree negative correlation
$r = -1$	Perfect negative correlation

Remark 1. The Karl Pearson's coefficient of correlation is also referred to as **product moment correlation coefficient** or as **Karl Pearson's product moment correlation coefficient.**

Remark 2. The Karl Pearson's coefficient of correlation, r, is also denoted by $\rho(x, y)$ or simply by ρ. The letter ρ is the Greek letter 'rho'.

Remark 3. The square of Karl Pearson's coefficient of correlation is called the **coefficient of determination.**

For example, if $r = 0.753$, then the coefficient of determination is $(0.753)^2 = 0.567$.

The *coefficient of determination* always lies between 0 and 1, both inclusive.

Remark 4.
$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}.$$

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

$$\therefore \qquad \mathbf{r} = \frac{\boldsymbol{\Sigma(x-\bar{x})(y-\bar{y})}}{\boldsymbol{n\sigma_x\,\sigma_y}}.$$

Example 1. *From the data given below calculate coefficient of correlation and interpret it :*

	x	*y*
Number of items	*8*	*8*
Mean	*68*	*69*
Sum of squares of deviations from mean	*36*	*44*

Sum of products of deviations of x and y from their respective means = 24.

Solution. We are given

$n = 8,\ \bar{x} = 68,\ \bar{y} = 69,\ \Sigma(x-\bar{x})^2 = 36,\ \Sigma(y-\bar{y})^2 = 44,\ \Sigma(x-\bar{x})(y-\bar{y}) = 24.$

Coefficient of correlation,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{24}{\sqrt{36}\sqrt{44}} = \frac{24}{39.7995} = +\ \mathbf{0.603}.$$

$\therefore$ There is moderate degree positive linear correlation between the variables x and y.

NOTES

Example 2. *The coefficient of correlation between two variables X and Y is 0.48. The covariance is 36. The variance of X is 16. Find the standard deviation of Y.*

Solution. We have $r = 0.48$, Covariance = 36, $\sigma_X^2 = 16$.

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\,\sqrt{\Sigma(Y-\overline{Y})^2}}$$

$$\Rightarrow \qquad r = \frac{\dfrac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n}}{\sqrt{\dfrac{\Sigma(X-\overline{X})^2}{n}}\,\sqrt{\dfrac{\Sigma(Y-\overline{Y})^2}{n}}} = \frac{\text{Covariance}^*}{\sigma_X\,\sigma_Y}$$

$$\therefore \qquad 0.48 = \frac{36}{\sqrt{16}.\,\sigma_Y} \quad \text{or} \quad \frac{48}{100} = \frac{9}{\sigma_Y} \quad \text{or} \quad \sigma_Y = \mathbf{18.75.}$$

Example 3. *Calculate the Karl Pearson's coefficient of correlation from the data given below :*

x	*4*	*6*	*8*	*10*	*11*
y	*2*	*3*	*4*	*6*	*12*

Solution. **Calculation of 'r'**

S. No.	*x*	*y*	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\ (y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
			$\bar{x} = 7.8$	$\bar{y} = 5.4$			
1	4	2	– 3.8	– 3.4	12.92	14.44	11.16
2	6	3	– 1.8	– 2.4	4.32	3.24	2.76
3	8	4	0.2	– 1.4	– 0.28	0.04	1.96
4	10	6	2.2	0.6	1.32	4.84	0.36
5	11	12	3.2	6.6	21.12	10.24	43.56
$n = 5$	$\Sigma x = 39$	$\Sigma y = 27$			$\Sigma(x-\bar{x})\ (y-\bar{y})$ $= 39.4$	$\Sigma(x-\bar{x})^2$ $= 32.8$	$\Sigma(y-\bar{y})^2$ $= 63.2$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{39}{5} = 7.8, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{27}{5} = 5.4$$

Now
$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\,\sqrt{\Sigma(y-\bar{y})^2}} = \frac{39.4}{\sqrt{32.8}\,\sqrt{63.2}}$$

$$= \frac{39.4}{(5.7271)(7.9498)} = \frac{39.4}{45.5293} = 0.8654.$$

It shows that there is high degree positive linear correlation between the variables.

7.6. ALTERNATIVE FORM OF 'R'

In the above examples, the calculations involved in **Example 3** is much more than in other examples. This is due to the fractional values of $\vec{x}$ and $\bar{y}$ in the data. Suppose for some data, we get $\bar{x} = 27.374$ and $\bar{y} = 14.873$, then it can be well imagined that lot

* **Covariance** between variables X and Y is defined as $\dfrac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n}$.

of time and energy would be consumed in computing the Karl Pearson's coefficient of correlation. There are very few chances to get $\bar{x}$ and $\bar{y}$ as whole numbers. In order to avoid the chance of facing difficulty in computing deviations of the values of variables from their respective arithmetic means, an alternative form is used which is discussed below :

We have $$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}}.$$

Now, $$\begin{aligned}\Sigma(x_i - \bar{x})(y_i - \bar{y}) &= \Sigma(x_i y_i - x_i\bar{y} - \bar{x}y_i + \overline{xy}) \\ &= \Sigma x_i y_i - (\Sigma x_i)\ \bar{y} - \bar{x}\ (\Sigma y_i) + n\bar{x}\,\bar{y} \\ &= \Sigma x_i y_i - \Sigma x_i \left(\frac{\Sigma y_i}{n}\right) - \left(\frac{\Sigma x_i}{n}\right)\Sigma y_i + n\left(\frac{\Sigma x_i}{n}\right)\left(\frac{\Sigma y_i}{n}\right) \\ &= \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n} = \frac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n}.\end{aligned}$$

Also $$\begin{aligned}\Sigma(x_i - \bar{x})^2 &= \Sigma(x_i^2 + \bar{x}^2 - 2x_i\bar{x}) = \Sigma x_i^2 + n\bar{x}^2 - 2(\Sigma x_i)\,\bar{x}. \\ &= \Sigma x_i^2 + n\left(\frac{\Sigma x_i}{n}\right)^2 - 2(\Sigma x_i)\left(\frac{\Sigma x_i}{n}\right) \\ &= \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \frac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n}.\end{aligned}$$

Similarly, $$\Sigma(y_i - \bar{y})^2 = \frac{n\Sigma y_i^2 - (\Sigma y_i)^2}{n}$$

$$\therefore \qquad r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2}\sqrt{\Sigma(y_i - \bar{y})^2}}$$

$$\Rightarrow \qquad r = \frac{\dfrac{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}{n}}{\sqrt{\dfrac{n\Sigma x_i^2 - (\Sigma x_i)^2}{n}}\sqrt{\dfrac{n\Sigma y_i^2 - (\Sigma y_i)^2}{n}}}$$

$$\therefore \qquad \mathbf{r} = \frac{\mathbf{n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)}}{\sqrt{\mathbf{n\Sigma x_i^2 - (\Sigma x_i)^2}}\sqrt{\mathbf{n\Sigma y_i^2 - (\Sigma y_i)^2}}}.$$

For simplicity, we write

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

Example 4. *Find the coefficient of correlation for the following data :*

$n = 10$, $\Sigma x = 50$, $\Sigma y = - 30$, $\Sigma x^2 = 290$, $\Sigma y^2 = 300$, $\Sigma xy = - 115$.

Solution. $$\begin{aligned} r &= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \\ &= \frac{10(-115) - (50)(-30)}{\sqrt{10(290) - (50)^2}\sqrt{10(300) - (-30)^2}} \\ &= \frac{350}{\sqrt{400}\sqrt{2100}} = \frac{35}{\sqrt{8400}} = \text{AL}\left[\log\left(\frac{350}{\sqrt{8400}}\right)\right] \\ &= \text{AL}\left[\log 35 - \frac{1}{2}\log 8400\right] = \text{AL}\left[1.5441 - \frac{1}{2}(3.9243)\right] \\ &= \text{AL}\,(-0.4181) = \text{AL}\,(\bar{1}.5819) = \mathbf{0.3819}.\end{aligned}$$

NOTES

Example 5. *Calculate the Karl Pearson's coefficient for the data given below:*

x	*2*	*3*	*5*	*7*	*3*
y	*15*	*17*	*4*	*5*	*4*

Solution. **Calculation of 'r'**

S. No.	*x*	*y*	*xy*	*x^2*	*y^2*
1	2	15	30	4	225
2	3	17	51	9	289
3	5	4	20	25	16
4	7	5	35	49	25
5	3	4	12	9	16
$n = 5$	$\Sigma x = 20$	$\Sigma y = 45$	$\Sigma xy = 148$	$\Sigma x^2 = 96$	$\Sigma y^2 = 571$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{5(148) - (20)(45)}{\sqrt{5(96) - (20)^2}\,\sqrt{5(571) - (45)^2}}$$

$$= \frac{-160}{\sqrt{80}\,\sqrt{830}} = \frac{-16}{\sqrt{664}} = -\text{AL}\left[\log\left(\frac{16}{\sqrt{664}}\right)\right]$$

$$= -\text{AL}\left[\log 16 - \frac{1}{2}\log 664\right] = -\text{AL}\left[1.2041 - \frac{1}{2}(2.8222)\right]$$

$$= -\text{AL}\,(-0.207) = -\text{AL}\,(-1 + 1 - 0.207)$$

$$= -\text{AL}\,(\bar{1}.793) = \mathbf{-0.6209.}$$

Example 6. *Calculate 'r' for the following data:*

x	*– 3*	*– 2*	*– 1*	*0*	*1*	*2*	*3*
y	*9*	*4*	*1*	*0*	*1*	*4*	*9*

Solution. **Calculation of 'r'**

S. No.	*x*	*y*	*xy*	*x^2*	*y^2*
1	– 3	9	– 27	9	81
2	– 2	4	– 8	4	16
3	– 1	1	– 1	1	1
4	0	0	0	0	0
5	1	1	1	1	1
6	2	4	8	4	16
7	3	9	27	9	81
$n = 7$	$\Sigma x = 0$	$\Sigma y = 28$	$\Sigma xy = 0$	$\Sigma x^2 = 28$	$\Sigma y^2 = 196$

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\,\sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{7(0) - (0)(28)}{\sqrt{7(28) - (0)^2}\,\sqrt{7(196) - (28)^2}} = \mathbf{0.}$$

Remark. In the above example, $r = 0$ indicates that there is no linear correlation between the variables. In fact, the variables x and y are very well correlated and there also exists algebraic relation $y = x^2$ between the variables. The correlation between x and y is curvilinear and Karl Pearson's coefficient of correlation does not help in estimating curvilinear correlation.

Example 7. *Calculate the coefficient of correlation between x and y:*

x	*22*	*24*	*25*	*27*	*21*	*22*	*23*
y	*41*	*44*	*45*	*48*	*40*	*42*	*44*

Solution. **Calculation of 'r'**

S. No.	x	y	xy	x^2	y^2
1	22	41	902	484	1681
2	24	44	1056	576	1936
3	25	45	1125	625	2025
4	27	48	1296	729	2304
5	21	40	840	441	1600
6	22	42	924	484	1764
7	23	44	1012	529	1936
$n = 7$	$\Sigma x = 164$	$\Sigma y = 304$	$\Sigma xy = 7155$	$\Sigma x^2 = 3868$	$\Sigma y^2 = 13246$

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

$$= \frac{7(7155) - (164)(304)}{\sqrt{7(3868) - (164)^2}\,\sqrt{7(13246) - (304)^2}}$$

$$= \frac{229}{\sqrt{180}\,\sqrt{306}} = \frac{229}{234.6913} = \mathbf{0.9757.}$$

Example 8. *Calculate the coefficient of correlation between X and Y from the following data and interpret the result :*

$$\Sigma XY = 8425,\ \overline{X} = 28.5,\ \overline{Y} = 28.0,\ \sigma_X = 10.5,\ \sigma_Y = 5.6, \text{ and } n = 10.$$

Solution. We have

$$\Sigma XY = 8425,\ \overline{X} = 28.5,\ \overline{Y} = 28,\ \sigma_X = 10.5,\ \sigma_Y = 5.6, \quad \text{and} \quad n = 10.$$

Now $\overline{X} = \dfrac{\Sigma X}{n} \Rightarrow 28.5 = \dfrac{\Sigma X}{10} \Rightarrow \Sigma X = 285$

and $\overline{Y} = \dfrac{\Sigma Y}{n} \Rightarrow 28 = \dfrac{\Sigma Y}{10} \Rightarrow \Sigma Y = 280.$

Also $\sigma_X = \sqrt{\dfrac{\Sigma X^2}{n} - \left(\dfrac{\Sigma X}{n}\right)^2} \Rightarrow \sqrt{n\Sigma X^2 - (\Sigma X)^2} = n\,\sigma_X = 10 \times 10.5 = 105$

and $\sigma_Y = \sqrt{\dfrac{\Sigma Y^2}{n} - \left(\dfrac{\Sigma Y}{n}\right)^2} \Rightarrow \sqrt{n\Sigma Y^2 - (\Sigma Y)^2} = n\,\sigma_Y = 10 \times 5.6 = 56.$

Now $r = \dfrac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\,\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$

$\therefore$ $r = \dfrac{10(8425) - (285)(280)}{105 \times 56} = \dfrac{4450}{105 \times 56} = \mathbf{0.7568.}$

The value of r shows that there is high degree positive linear correlation between the variables X and Y.

NOTES

EXERCISE 7.1

1. From the data given below, calculate the coefficient of correlation:

	x	y
Number of items	15	15
A.M.	25	18
Sum of squares of deviations from mean	136	138

Sum of products of deviations of x and y from their respective means = 122.

2. Find the coefficient of correlation for the following data:

x	5	4	3	2	1
y	5	2	10	8	4

3. Find the coefficient of correlation for the following data:

x	1	2	3	4	5
y	7	6	5	4	3

4. Find the Karl Pearson's coefficient of correlation for the following data:

x	1	2	8	4	5
y	10	9	8	8	7

5. Find Karl Pearson's coefficient of correlation between x and y for the following data:

x	1	2	3	4	5
y	2	5	7	8	10

6. Find the coefficient of correlation for the following data:

x	1	3	5	7	8	10
y	8	12	15	17	18	20

7. Find the coefficient of correlation for the following data:

x	1	2	3	4	5	6	7	8	9	10
y	10	9	8	8	6	12	4	3	18	1

8. Calculate the coefficient of correlation between the values of X and Y given below:

X	– 15	+ 18	– 12	– 10	+ 15	– 20	– 25	+ 15	+ 16	– 14
Y	+ 8	– 10	+ 5	+ 12	– 6	+ 4	+ 11	– 9	– 7	+ 13

9. Calculate the Karl Pearson's coefficient of correlation for the following data:

x	28	32	38	42	46	52	54	57	58	63
y	0	1	3	4	2	5	4	6	7	8

Answers

1. 0.8906	2. $r = -0.1980$	3. $r = -1$	4. $r = 0.7206$
5. $r = 0.9851$	6. $r = 0.9879$	7. $r = -0.1840$	8. -0.912
9. 0.9293			

7.7. STEP DEVIATION METHOD

When the values of x and y are numerically high, as in **Example 7** of Article **7.6,** the step deviation method is used.

Deviations of values of variables x and y are calculated from some chosen arbitrary numbers, called A and B. Let h be a *positive* common factor of all the deviations $(x - \text{A})$ of items in the x-series. The definition of h is valid, since at least one common factor "1" exist for all the deviations. Similarly let k be a *positive* factor of all the deviations $(y - \text{B})$ of items in the y-series.

Let $$u = \frac{x - \text{A}}{h} \quad \text{and} \quad v = \frac{y - \text{B}}{k}.$$

$\therefore$ The variables u and v are obtained by changing origin and scale of the variables x and y respectively.

Since correlation coefficient is independent of change of origin and scale, we have

$$r(x, y) = r(u, v).$$

$$\therefore \quad r(x, y) = \frac{\Sigma(u - \bar{u})(v - \bar{v})}{\sqrt{\Sigma(u - \bar{u})^2}\sqrt{\Sigma(v - \bar{v})^2}}$$

On simplification, we get

$$\mathbf{r(x, y)} = \frac{\mathbf{n\Sigma uv - (\Sigma u)(\Sigma v)}}{\sqrt{\mathbf{n\Sigma u^2 - (\Sigma u)^2}}\sqrt{\mathbf{n\Sigma v^2 - (\Sigma v)^2}}}.$$

The values of u and v are called the **step deviations** of the values of x and y respectively. In the above form :

Σu is the sum of step deviations of the items of x-series.

Σv is the sum of step deviations of the items of y-series.

Σuv is the sum of the products of the step deviations of items of x-series with the corresponding step deviations of items of y-series.

Σu^2 is the sum of the squares of the step deviations of items of x-series.

Σv^2 is the sum of the squares of the step deviations of items of y-series.

In practical problems, the choice of common factors h and k would not create any problem. Even if we do not care to compute step deviations, by dividing the deviations of values of x and y by some common factor, the formula would still work. Suppose we have taken deviations (u) of the items of x-series from A,

i.e., $$u = x - \text{A} = \frac{x - \text{A}}{1}.$$

We can consider the values of u as the step deviations of the items of x-series, taking '1' as the common factor. Similar argument would also work for y-series.

Therefore, in solving problems, we first calculate deviations of items of x-series and y-series from some convenient and suitable assumed means A and B respectively. These deviations of x-series and y-series are then divided by positive common factors,

NOTES

if at all desired. If we do not bother to divide these deviations by common factors, then these deviations would be thought of as *step deviations* of items of given series with '1' as the common factor for both series.

Thus if $u = x - A$ and $v = y - B$, then, we have

$$\mathbf{r(x, y)} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}.$$

Example 9. *From the given data, calculate the Karl Pearson's coefficient of correlation :*

Months	*Price of Commodity*	
	A	*B*
Jan.	*35*	*65*
Feb.	36	72
March	*40*	*78*
April	*38*	*77*
May	*37*	*76*
June	*39*	*77*
July	*41*	*80*
August	*40*	*79*
Sept.	*36*	*76*
Oct.	*38*	*75*

Use 38 as assumed mean for commodity A and 75 for commodity B.

Soluton. Let the variables 'price of A' and 'price of B' be denoted by x and y respectively.

Calculation of 'r'

Months	x	y	$u = x - A$ $A = 38$	$v = y - B$ $B = 75$	uv	u^2	v^2
Jan.	35	65	− 3	− 10	30	9	100
Feb.	36	72	− 2	− 3	6	4	9
March	40	78	2	3	6	4	9
April	38	77	0	2	0	0	4
May	37	76	− 1	1	− 1	1	1
June	39	77	1	2	2	1	4
July	41	80	3	5	15	9	25
August	40	79	2	4	8	4	16
Sept.	36	76	− 2	1	− 2	4	1
Oct	38	75	0	0	0	0	0
$n = 10$			$\Sigma u = 0$	$\Sigma v = 5$	$\Sigma uv = 64$	$\Sigma u^2 = 36$	$\Sigma v^2 = 169$

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

$$= \frac{10(64) - (0)(5)}{\sqrt{10(36) - (0)^2}\sqrt{10(169) - (5)^2}}$$

$$= \frac{640}{\sqrt{360}\,\sqrt{1665}} = \text{AL}\left[\log \frac{640}{\sqrt{360 \times 1665}}\right]$$

$$= \text{AL}\left[\log 640 - \frac{1}{2}(\log 360 + \log 1665)\right]$$

$$= \text{AL}\left[\log 2.8064 - \frac{1}{2}(2.5563 + 3.2214)\right]$$

$$= \text{AL}\,(-\,0.08245) = \text{AL}\,(-\,1 + 1 - 0.08245)$$

$$= \text{AL}\,(\bar{1}\,.\,91755) = \mathbf{0.8271.}$$

∴ There is high degree positive linear correlation between the prices of commodities A and B.

Example 10. *Calculate the coefficient of correlation for the data given below :*

Age of husband (in years)	*23*	*27*	*28*	*28*	*29*	*30*	*31*	*33*	*35*	*36*
Age of wife (in years)	*18*	*20*	*22*	*27*	*21*	*29*	*27*	*29*	*28*	*29*

Solution. Let x and y denote the variables 'Age of Husband' and 'Age of Wife' respectively.

Calculation of 'r'

S. No.	x	y	$u = x - A$ $A = 30$	$v = y - B$ $B = 24$	uv	u^2	v^2
1	23	18	− 7	− 6	42	49	36
2	27	20	− 3	− 4	12	9	16
3	28	22	− 2	− 2	4	4	4
4	28	27	− 2	3	− 6	4	9
5	29	21	− 1	− 3	3	1	9
6	30	29	0	5	0	0	25
7	31	27	1	3	3	1	9
8	33	29	3	5	15	9	25
9	35	28	5	4	20	25	16
10	36	29	6	5	30	36	25
$n = 10$			$\Sigma u = 0$	$\Sigma v = 10$	$\Sigma uv = 123$	$\Sigma u^2 = 138$	$\Sigma v^2 = 174$

Now
$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\,\sqrt{n\Sigma v^2 - (\Sigma v)^2}} = \frac{10(123) - (0)(10)}{\sqrt{10(138) - (0)^2}\,\sqrt{10(174) - (-10)^2}}$$

$$= \frac{1230}{\sqrt{1380}\sqrt{1740 - 100}} = \frac{1230}{\sqrt{1380}\sqrt{1640}}$$

$$= \text{AL}\left\{\log 1230 - \frac{1}{2}(\log 1380 + \log 1640)\right\}$$

$$= \text{AL}\left\{3.0899 - \frac{1}{2}(3.1399 + 3.2148)\right\} = \text{AL}\left\{3.0899 - \frac{1}{2}(6.3547)\right\}$$

$= \text{AL}\{3{:}0899 - 3.1773\} = \text{AL}\{-0.0874\} = \text{AL}\{\bar{1}.9126\} = 0.8177$

$\therefore$ **r = 0.8177.**

It shows that there is high degree positive linear correlation between the variables.

NOTES

Example 11. *Complete the coefficient of correlation between the variables x and y from the given data :*

No. of pairs of x and y series	*= 8*
x-series A.M.	*= 74.5*
x-series S.D.	*= 13.07*
x-series assumed mean	*= 69*
y-series A.M.	*= 125.5*
y-series S.D.	*= 15.85*
y-series assumed mean	*= 112*

Sum of products of corresponding deviations of x and y series = 2176.

Solution. Let A = 69, B = 112

and $u = x - 69$, $v = y - 112$.

Let 'r' be the coefficient of correlation between x and y variables.

$$\therefore \quad r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}. \quad \text{...(1)}$$

We are given

$$\Sigma uv = 2176,\ \bar{x} = 74.5,\ \bar{y} = 125.5$$
$$\sigma_x = 13.07,\ \sigma_y = 15.85,\ n = 8,\ A = 69,\ B = 112.$$

We know that $\bar{x} = A + \dfrac{\Sigma u}{n}$.

$$\therefore \quad 74.5 = 69 + \frac{\Sigma u}{8} \quad i.e., \quad \Sigma u = 8(74.5 - 69) = 44$$

Also $\bar{y} = B + \dfrac{\Sigma v}{n}$

$$\therefore \quad 125.5 = 112 + \frac{\Sigma v}{8} \quad i.e., \quad \Sigma v = 8(125.5 - 112) = 108$$

Again $$\sigma_x = \sqrt{\frac{\Sigma u^2}{n} - \left(\frac{\Sigma u}{n}\right)^2} = \sqrt{\frac{n\Sigma u^2 - (\Sigma u)^2}{n^2}}$$

$$\therefore \quad \sqrt{n\Sigma u^2 - (\Sigma u)^2} = n\sigma_x = 8(13.07) = 104.56$$

Also $$\sigma_y = \sqrt{\frac{\Sigma v^2}{n} - \left(\frac{\Sigma v}{n}\right)^2} = \sqrt{\frac{n\Sigma v^2 - (\Sigma v)^2}{n^2}}$$

$$\therefore \quad \sqrt{n\Sigma v^2 - (\Sigma v)^2} = n\sigma_y = 8(15.85) = 126.8$$

$$\therefore \quad (1) \Rightarrow \quad r = \frac{8(2176) - (44)(108)}{(104.56)(126.8)} = \frac{12656}{13258.208} = \mathbf{0.9546.}$$

EXERCISE 7.2

1. Calculate the coefficient of correlation for the following ages of husbands and wives :

Age of husband	23	27	28	28	29	30	31	33	35	36
Age of wife	18	20	20	27	21	29	27	29	28	29

2. Find the coefficient of correlation for the following data. Also explain, what does it express.

x	300	350	400	450	500	550	600	650	700
y	800	900	1000	1100	1200	1300	1400	1500	1600

3. Calculate the coefficient correlation between the values of x and y given below :

x	78	89	96	69	59	79	68	61
y	125	137	156	112	107	136	123	108

(You may use 69 as working mean for x and 112 that for y).

4. Compute the coefficient of correlation between sales tax collected and sales of product '*M*' in ten countries selected at random from those served by the company.

Country	A	B	C	D	E	F	G	H	I	J
Sales tax collected (x) (in pounds)	16	24	32	15	20	12	18	14	10	29
Units of 'M' sold (y)	40	50	68	36	45	27	42	36	29	67

5. Calculate K.P's coefficient of correlation for the given series :

Husband's age	24	27	28	28	29	30	32	33	35	35	40
Wife's age	18	20	22	25	22	28	28	30	27	30	22

6. Find the coefficient of correlation between the variables 'income' and 'expenditure'.

Family	A	B	C	D	E	F	G	H	I	J
Income (₹)	95	90	110	100	85	105	95	100	105	95
Expenditure (₹)	90	95	115	95	85	110	90	95	95	95

7. Calculate the coefficient of correlation between the marks obtained by 12 students in Statistics and Accountancy paper :

Marks in Statistics	52	74	93	55	41	23	92	64	40	71	33	71
Marks in Accountancy	45	80	63	60	35	40	70	58	43	64	51	75

NOTES

8. From the following data, calculate the coefficient of correlation between 'age' and 'playing habit'.

Age group	*No. of employees*	*No. of regular players*
20–30	25	10
30–40	60	30
40–50	40	12
50–60	20	2
60–70	20	1

9. From the following table, given the distribution of students and also regular players among them according to age group, find the correlation between 'age' and 'playing habit'.

Age	15–16	16–17	17–18	18–19	19–20	20–21
No. of students	200	270	340	360	400	300
Regular players	180	162	170	180	180	60

10. Calculate Karl Pearson's coefficient of correlation for the following paired data :

x	28	41	40	38	35	33	40	32	36	33
y	23	34	33	34	30	26	28	31	36	38

Answers

1. $r = 0.8067$ 2. $r = 1$ 3. $r = 0.9544$
4. $r = 0.9854$ 5. $r = 0.504$ 6. $r = 0.8152$
7. $r = 0.7886$ 8. $r = -0.904$ 9. $r = -0.936$
10. $r = 0.4403$.

SPEARMAN'S RANK CORRELATION METHOD

7.8. MEANING OF SPEARMAN'S RANK CORRELATION

In practical life, we come across problems of estimating correlation between variables, which are not quantitative in nature. Suppose, we are interested in deciding if there is any correlation between the variables 'honesty' and 'smartness' among a group of students. Here the variables 'honesty' and 'smartness' are not capable of quantitative measurement. These variables are qualitative in nature. Ranking is possible in case of qualitative variables.

Spearman's rank correlation method is used for studying linear correlation between variables which are not necessarily quantitative in nature. This method works for both quantitative as well as qualitative variables.

Let n pairs of values of variables x and y be given. The first step is to express the values of the variables in ranks. In case of qualitative variables, the data would be given in the desired form. For quantitative variables, the ranks are allotted according to the magnitude of the values of the variables. Generally the I rank is allotted to the item with highest value. If the highest value of the first variable is allotted I rank, then the same method is to be adopted for finding the ranks of the values of the other variable. In allotting ranks, difficulty arises when the values of two or more items in a series are equal. We shall consider this case separately.

7.8.1 Case I. Non-repeated Ranks

Let R_1 and R_2 represent the ranks of the items corresponding to the variables x and y respectively.

The coefficient of rank correlation (r_k) is given by the formula :

$$\mathbf{r_k = 1 - \frac{6\Sigma D^2}{n(n^2-1)}},$$

where n is the number of pairs and D denotes the difference between ranks *i.e.*, $(R_1 - R_2)$ of the corresponding values of the variables.

Example 12. *Ten students of a class are ranked in intelligence tests given by the two teachers. Find the coefficient of rank correlation.*

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Ranks by Teacher A	*6*	*7*	*3*	*2*	*10*	*1*	*9*	*8*	*4*	*5*
Ranks by Teachers B	*7*	*10*	*4*	*1*	*9*	*3*	*8*	*6*	*2*	*5*

Solution. Let R_1 and R_2 denote the ranks allotted by teachers A and B respectively.

Calculation of 'r_k'

S. No.	R_1	R_2	$D = R_1 - R_2$	D^2
1	6	7	−1	1
2	7	10	−3	9
3	3	4	−1	1
4	2	1	1	1
5	10	9	1	1
6	1	3	−2	4
7	9	8	1	1
8	8	6	2	4
9	4	2	2	4
10	5	5	0	0
$n = 10$				$\Sigma D^2 = 26$

Coefficient of rank correlation,

$$r_k = 1 - \frac{6\Sigma D^2}{n(n^2-1)} = 1 - \frac{6(26)}{10(10^2-1)} = 1 - 0.1567 = \mathbf{0.8424.}$$

It shows that there is high degree positive linear correlation between the variables.

Example 13. *Calculate the coefficient of correlation for the following data by the method of rank differences :*

x	*75*	*88*	*95*	*70*	*60*	*80*	*81*	*50*
y	*120*	*130*	*150*	*115*	*110*	*140*	*142*	*100*

Solution. Let R_1 and R_2 denote the ranks of the variables x and y respectively. The first rank is allotted to the greatest item in each series.

NOTES

Calculation of 'r$_k$'

S. No.	x	y	R_1	R_2	$D = R_1 - R_2$	D^2
1	75	120	5	5	0	0
2	88	130	2	4	–2	4
3	95	150	1	1	0	0
4	70	115	6	6	0	0
5	60	110	7	7	0	0
6	80	140	4	3	1	1
7	81	142	3	2	1	1
8	50	100	8	8	0	0
$n = 8$						$SD^2 = 6$

Coefficient of rank correlation,

$$r_k = 1 - \frac{6\Sigma D^2}{n(n^2-1)} = 1 - \frac{6(6)}{8(8^2-1)} = \mathbf{0.9286.}$$

It shows that there is high degree positive linear correlation between the variables.

Example 14. *What is Rank Correlation ? Find the formula for the rank correlation coefficient.*

Solution. Let (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) be n pairs of ranks of two variables x and y.

We have
$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \qquad ...(1)$$

Since the values of x represent ranks, the values $x_1, x_2, ..., x_n$ of x are distinct and take values from 1 to n.

$$\therefore \qquad \Sigma x = 1 + 2 + 3 + ... + n = \frac{n(n+1)}{2}$$

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

$$\therefore \qquad \bar{x} = \frac{\Sigma x}{n} = \frac{1}{n}.\frac{n(n+1)}{2} = \frac{n+1}{2}$$

Now
$$\Sigma(x-\bar{x})^2 = \Sigma(x^2 + \bar{x}^2 - 2x\bar{x}) = \Sigma x^2 + n\bar{x}^2 - 2\bar{x}\Sigma x$$
$$= \Sigma x^2 + n\bar{x}^2 - 2\bar{x}.n\bar{x} = \Sigma x^2 - n\bar{x}^2$$
$$= \frac{n(n+1)(2n+1)}{6} - n.\left(\frac{n+1}{2}\right)^2$$
$$= \frac{n(n+1)}{2}\left[\frac{2n+1}{3} - \frac{n+1}{2}\right] = \frac{n(n+1)}{2} \times \frac{n-1}{6} = \frac{n(n^2-1)}{12}.$$

Similarly, $\bar{y} = \dfrac{n+1}{2}$ and $\Sigma(y-\bar{y})^2 = \dfrac{n(n^2-1)}{12}$.

Let $D = x - y$.

$\therefore$ $D = (x-\bar{x}) - (y-\bar{y}) \qquad (\because \quad \bar{x} = \bar{y})$

$$\Rightarrow \qquad D^2 = (x-\bar{x})^2 + (y-\bar{y})^2 - 2(x-\bar{x})(y-\bar{y})$$

$$\Rightarrow \qquad 2(x-\bar{x})(y-\bar{y}) = (x-\bar{x})^2 + (y-\bar{y})^2 - D^2$$

$$\Rightarrow \qquad 2\Sigma\,(x-\bar{x})(y-\bar{y}) = \Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2 - \Sigma D^2$$

$$= \frac{n(n^2-1)}{12} + \frac{n(n^2-1)}{12} - \Sigma D^2 = \frac{n(n^2-1)}{6} - \Sigma D^2$$

$$\therefore \qquad \Sigma(x-\bar{x})(y-\bar{y}) = \frac{n(n^2-1)}{12} - \frac{\Sigma D^2}{2}.$$

Now,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{\frac{n(n^2-1)}{12} - \frac{\Sigma D^2}{2}}{\sqrt{\frac{n(n^2-1)}{12}}\sqrt{\frac{n(n^2-1)}{12}}}$$

$$= \frac{\frac{n(n^2-1)}{12} - \frac{\Sigma D^2}{2}}{\frac{n(n^2-1)}{12}} = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

$$\therefore \qquad \mathbf{r = 1 - \frac{6\Sigma D^2}{n(n^2-1)}}.$$

7.8.2 Case II. Repeated Ranks

Here we shall consider the case, when the values of two or more items in a series are equal. In such cases, we allot equal ranks to all the items with equal values. Suppose that the values of three items in a series are equal at the fourth place, then each item with equal value would be allotted rank $\frac{4+5+6}{3} = 5$. Similarly, if there happen to be two items in a series with equal values at the seventh place, then each item with equal value would be allotted rank $\frac{7+8}{2} = 7.5$.

In case of repeated ranks, the coefficient of rank correlation is given by the formula,

$$\mathbf{r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \ldots\ldots\right\}}{n(n^2-1)}}$$

where n is the number of pairs and D denote the difference between ranks $(R_1 - R_2)$ of the corresponding values of the variables. In $\frac{1}{12}(m^3 - m)$, m is number of items whose ranks are equal. The term $\frac{1}{12}(m^3 - m)$ is to be added for each group of items with equal ranks. Now, we shall illustrate this method by taking some examples.

Example 15. *Calculate the rank coefficient of correlation for the following data :*

x	*80*	*78*	*75*	*75*	*68*	*67*	*60*	*59*
y	*12*	*13*	*14*	*14*	*14*	*16*	*15*	*17*

Solution. Let R_1 and R_2 denote the ranks of the variables x and y respectively. The first rank is allotted to the greatest item in each series.

Calculation of 'r_k'

S. No.	*x*	*y*	R_1	R_2	$D = R_1 - R_2$	D^2
1	80	12	1	8	– 7	49
2	78	13	2	7	– 5	25
3	75	14	3.5	5	– 1.5	2.25
4	75	14	3.5	5	– 1.5	2.25
5	68	14	5	5	0	0
6	67	16	6	2	4	16
7	60	15	7	3	4	16
8	59	17	8	1	7	49
$n = 8$						$\Sigma D^2 = 159.5$

Now
$$r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + ...\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left\{159.5 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{8(8^2 - 1)}$$

$$= 1 - \frac{6\{159.5 + 0.5 + 2\}}{8 \times 63} = 1 - \frac{6 \times 162}{8 \times 63} = 1 - \frac{27}{14} = -\frac{13}{14} = -\mathbf{0.9286}.$$

It shows that there is a high degree negative linear correlation between the variables.

Example 16. *Calculate the rank coefficient of correlation for the following data :*

X-series	*112*	*106*	*109*	*84*	*95*	*95*	*117*	*97*	*95*	*115*
Y-series	*70*	*68*	*80*	*65*	*71*	*60*	*77*	*68*	*63*	*75*

Solution. Let R_1 and R_2 denote the ranks of the variables X and Y respectively. The first rank is allotted to the greatest item in each series.

Calculation of 'r$_k$'

S. No.	X	Y	R_1	R_2	$D = R_1 - R_2$	D^2
1	112	70	3	5	−2	4
2	106	68	5	6.5	−1.5	2.25
3	109	80	4	1	3	9
4	84	65	10	8	2	4
5	95	71	8	4	4	16
6	95	60	8	10	−2	4
7	117	77	1	2	−1	1
8	97	68	6	6.5	−0.5	0.25
9	95	63	8	9	−1	1
10	115	75	2	3	−1	1
$n = 10$						$\Sigma D^2 = 42.5$

Now, $$r_k = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m^3 - m) + \ldots\ldots\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left\{42.5 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2)\right\}}{10(10^2 - 1)}$$

$$= 1 - \frac{6\left\{42.5 + 2 + \frac{1}{2}\right\}}{990} = 1 - \frac{45}{165} = \mathbf{0.7273.}$$

It shows that there is a moderate degree positive linear correlation between the variables.

Example 17. *The rank correlation coefficient between marks obtained by some students in 'Statistics' and 'Accountancy' is 0.8. If the total of squares of rank difference is 33, find the number of students.*

Solution. We have $r_k = 0.8, \Sigma D^2 = 33.$

Let number of students be n.

$$\therefore \quad r_k = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} \quad \text{or} \quad 0.8 = 1 - \frac{6 \times 33}{n(n^2 - 1)}$$

$$\Rightarrow \quad \frac{198}{n(n^2 - 1)} = 0.2 = \frac{1}{5} \quad \Rightarrow \quad n(n^2 - 1) = 990$$

$$\Rightarrow \quad n(n^2 - 1) = 2 \times 5 \times 3 \times 3 \times 11$$

$$\Rightarrow \quad (n + 1)n\,(n - 1) = 11 \times 10 \times 9$$

i.e., $$(n + 1)n\,(n - 1) = (10 + 1)\,10\,(10 - 1)$$

$$\therefore \quad n = \mathbf{10.}$$

NOTES

7.8.3 Merits

1. This method is applicable to both qualitative and quantitative variables.
2. Only this method in applicable when ranks are given.
3. This method involves less calculation work as compared to Karl Pearson's method.

7.8.4 Demerits

This method is applicable only when the correlation between the variables is linear.

EXERCISE 7.3

1. From the following data, calculate Spearman's Rank Correlation coefficient.

S. No.	1	2	3	4	5	6	7	8	9	10
Rank Difference	– 2	– 4	– 1	+ 3	+ 2	0	– 2	+ 3	+ 3	2

2. The following are the ranks obtained by a group of 7 students in intelligence tests conducted by two teachers separately. Calculate the rank correlation coefficient.

Student	1	2	3	4	5	6	7
Ranks by Teacher 'A'	6	5	7	4	3	2	1
Ranks by Teacher 'B'	3	5	7	1	2	4	6

3. Ten competitors in a beauty contest are ranked by three judges in the following order :

Ist judge	1	6	5	10	3	2	4	9	7	8
IInd judge	3	5	8	4	7	10	2	1	6	9
IIIrd judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to discuss which pair of judges has the nearest approach to common taste in beauty.

4. The following data relates to the monthly income and expenditure of 10 families. Find the coefficient of rank correlation between the variables.

Family	A	B	C	D	E	F	G	H	I	J
Income (in ₹)	1000	700	870	500	900	950	1100	400	1500	800
Expenditure (in ₹)	900	600	800	490	810	860	910	450	1200	750

5. Compute the coefficient of rank correlation between x and y from the data given below :

x	8	10	7	15	3	20	21	5	10	14	8	16	22	19	6
y	3	12	8	13	20	9	14	11	4	16	15	10	18	23	25

6. Calculate the coefficient of rank correlation for the following data of marks of eight students in Statistics and Accountancy :

Marks in Statistics	52	63	45	36	72	65	45	25
Marks in Accountancy	62	53	51	25	79	43	60	30

7. Following are the ranks obtained by 10 students in two subjects Statistics and Economics. To what extent knowledge of students in two subjects is related ?

Statistics	1	2	3	4	5	6	7	8	9	10
Economics	2	4	1	5	3	9	7	10	6	8

8. The coefficient of rank correlation of the marks obtained by 10 students in Mathematics and Accountancy was found to be + 0.91. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 0 instead of 3. Find the correct coefficient of rank correlation.

9. Calculate rank coefficient of correlation for the following data :

A	115	109	112	87	98	98	120	100	98	118
B	75	73	85	70	76	65	82	73	68	80

10. Find the coefficient of correlation between x and y by the method of rank differences :

x	42	48	35	50	50	57	45	40	50	39
y	90	110	95	95	95	120	115	128	116	130

Answers

1. $r_k = 0.6364$ **2.** $r_k = 0.143$ **3.** Ist and IIIrd **4.** $r_k = 1$

5. $r_k = 0.0425$ **6.** $r_k = 0.643$ **7.** $r_k = 0.7576$

8. Correct $r_k = 0.855$. **9.** $r_k = 0.7212$ **10.** $r_k = -0.0556$.

EXERCISE 7.4

1. Explain the meaning of the term 'Correlation'. Does it always signify cause and effect relationship ?
2. How would you interpret the sign and magnitude of a calculated 'r'. Consider in particular the values $r = 0$, $r = -1$ and $r = +1$.
3. Explain the meaning of the term 'correlation'. Name the different measures of correlation and discuss their uses.
4. Define correlation and discuss its significance in statistical analysis.
5. Explain different methods of computing correlation
6. Elucidate the main features of Karl Pearson's coefficient of correlation.
7. What is correlation ?

7.9. SUMMARY

NOTES

- Two variables may be related in the sense that the changes in the values of one variable are accompanied by changes in the values of the other variable. But this cannot be interpreted in the sense that the changes in one variable has necessarily caused changes in the other variable.
- The correlation between two variables is said to be **positive** if the variables, on an average, move in the same direction. That is, an increase (or decrease) in the value of one variable is accompanied, on an average, by an increase (or decrease) in the value of the other variable.
- The correlation between two variables is said to be **linear** if the ratio of change in one variable to the change in the other variable is almost constant.
- The correlation is said to be **simple** if there are only two variables under consideration. The correlation between sale and profit figures of a departmental store is simple. If there are more than two variables under consideration, then the correlation is either multiple or partial.
- Spearman's rank correlation method is used for studying linear correlation between variables which are not necessarily quantitative in nature. This method works for both quantitative as well as qualitative variables.

8. INDEX NUMBERS

NOTES

STRUCTURE

8.1. INTRODUCTION

The **index numbers** are defined as specialized averages used to measure change in a variable or a group of related variables with respect to time or geographical location or some other characteristic.

In our course of discussion, we shall restrict ourselves to the study of changes in a group of related variables with respect to time only. Changes in related variables are expressed clearly by using index numbers, because these are generally expressed as percentages.

The index numbers are used to measure the change in production, prices, values etc., in related variables over time or geographical location. The barometers are used to study changes in whether conditions, similarly the index numbers are used to study the changes in economic and business activities. That is, why, the index numbers are also called **'Economic Barometers'.**

8.2. PURPOSE OF CONSTRUCTING INDEX NUMBERS

NOTES

1. Index numbers are used for computing real incomes from money incomes. The wages, dearness allowances etc., are fixed on the basis of real income. The money income is divided by an appropriate consumer's price index number to get real income.

2. Index numbers are constructed to compare the changes in related variables over time. Index numbers of industrial production can be used to see the change in the production that has occurred in the current period.

3. Index numbers are used to study the changes occurred in the past. This knowledge help in forecasting.

4. Index numbers are used to study the changes in prices, industrial production, purchasing powers of money, agricultural production etc., of different countries. With the use of index numbers, the comparative study is also made possible for such variables.

8.3. TYPES OF INDEX NUMBERS

There are mainly three of index numbers :

I. Price Index Numbers.

II. Quantity Index Numbers.

III. Value Index Numbers.

In our course of discussion, we shall confine mainly to 'Price Index Numbers'. Price index numbers measure the changes is prices of commodities in the current period in comparison with the prices of commodities in the base period.

I. PRICE INDEX NUMBERS

8.4. METHODS OF PRICE INDEX NUMBERS

For constructing price index numbers, the following method are used :

(*i*) Simple Aggregative Method

(*ii*) Simple Average of Price Relatives Method

(*iii*) Laspeyre's Method

(*iv*) Paasche's Method

(*v*) Dorbish and Bowley's Method

(*vi*) Fisher's Method

(*vii*) Marshall Edgeworth's Method

(*viii*) Kelly's Method

(*ix*) Weighted Average of Price Relatives Method

(*x*) Chain Base Method.

First nine methods are fixed base methods of constructing price index number.

NOTES

8.4.1. Simple Aggregative Method

This is the simplest method of computing index number. In this method, we have

$$\mathbf{P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100}$$

where 0 and 1 suffixes stand for base period and current period respectively.

P_{01} = price index number for the current period

Σp_1 = sum of prices of commodities per unit in the current period

Σp_0 = sum of prices of commodities per unit in the base period.

In other words, this price index number is the sum of prices of commodities in the current period expressed as percentage of the sum of prices in the base period. Consider the data :

Item	*Price in base period* *p_0(in ₹)*	*Price in current period* *P_1 (in ₹)*
A	5	6
B	8	10
C	18	27
D	112	84
E	12	15
F	6	9
Total	$\Sigma p_0 = 161$	$\Sigma p_1 = 151$

Here $P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{151}{161} \times 100 = \mathbf{93.79}.$

This index number shows that there is fall in the prices of commodities to the extent of 6.21%. It may be noted that the prices of every item has increased in the current period except for the item *D*. On the other hand, the index number is declaring a decrease in prices on an average. This is not in consistency with the definition of index numbers. In fact, this unwanted result is due to the presence of an extreme item (*D*) in the series. So, in the presence of extreme items, this method is liable to give misleading results. This is a demerit of this method.

Let us find price index number for the data given below :

Item	*Unit*	*Price (in ₹)*	
		1994 *(p_0)*	*1996* *(p_1)*
Sugar	kg	6	7
Milk.	litre	3	4
Ghee	kg	45	50

Here $\Sigma p_0 = 6 + 3 + 45 = 54$

and $\Sigma p_1 = 7 + 4 + 50 = 61$

$\therefore$ $P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{61}{54} \times 100 = \mathbf{112.96}.$

Here we have considered the price of sugar per kg. Now we use the price of sugar per quintal, for calculating index number for the year 1996.

NOTES

Item	*Unit*	*Price (in ₹)*	
		1994 (p_0)	*1996* (p_1)
Sugar	quintal	600	700
Milk	litre	3	4
Ghee	kg	45	50

In this case, $\Sigma p_0 = 600 + 3 + 45 = 648$

and $\Sigma p_1 = 700 + 4 + 50 = 754$

$$\therefore \quad P_{01} = \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{754}{648} \times 120 = \mathbf{116.36.}$$

The index number has changed, whereas we have not affected any change in the data except for writing the price of sugar in a different unit. This type of variation in the value of index numbers is beyond one's expectation. This is another limitation with this method.

8.4.2. Simple Average of Price Relatives Method

Before introducing this method of finding index number, we shall first explain the concept of 'price relative'. The **price relative** of a commodity in the current period with respect to base period is defined as the price of the commodity in the current period expressed as a percentage of the price in the base period. Mathematically,

$$\textbf{Price Relative (P)} = \frac{\mathbf{p_1}}{\mathbf{p_0}} \times \mathbf{100.}$$

For example, if the prices of a commodity be ₹ 5 and ₹ 6 in the years 1995 and 1996 respectively, then the price relative of the commodity in 1996 w.r.t. 1995 is

$$\frac{6}{5} \times 100 = 120.$$

In the simple average of price relatives method of computing index numbers, simple average of price relatives of all the items is the required index number.

Mathematically,

$$P_{01} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{n} \qquad \text{(if A.M. is used)}$$

i.e.,

$$\mathbf{P_{01}} = \frac{\mathbf{\Sigma P}}{\mathbf{n}}$$

where P_{01} is the required price index number,

$\frac{p_1}{p_0} \times 100$ = Price relative = P

n = no. of commodities under consideration.

In averaging price relatives, geometric mean is also used. In this case, the formula is

$$\mathbf{P_{01}} = \textbf{Antilog}\left(\frac{\mathbf{\Sigma \log P}}{\mathbf{n}}\right)$$

It has already been observed that the index number computed by using simple aggregative method is unduly affected by the extreme items, present in the series.

NOTES

We shall just show that this method of computing index number is not at all affected by the extreme items. We compute the index number for the data considered in the previous method:

Index No. by simple A.M. of P.R. Method

Item	*Price in the base period* (p_0) *(in ₹)*	*Price in the current period* (p_1) *(in ₹)*	*Price Relatives* $P = \frac{p_1}{p_0} \times 100$
A	6	6	120
B	8	10	125
C	18	27	150
D	112	84	75
E	12	15	125
F	6	9	150
			$\Sigma P = 745$

$$\therefore \qquad P_{01} = \frac{\Sigma P}{n} = \frac{745}{6} = \mathbf{124.17.}$$

Here the index number is advocating the fact that the prices of commodities have raised on an average.

There is one more advantage of using this method. The index number, computed by averaging the price relatives is not affected by the change in measuring unit of any commodity. We illustrate this by using the data taken in the previous method :

Item	*Unit*	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$
Sugar	kg	6	7	116.67
Milk	litre	3	4	133.33
Ghee	kg	45	50	111.11
				$\Sigma P = 361.11$

$$\therefore \qquad P_{01} = \frac{\Sigma P}{n} = \frac{361.11}{3} = \mathbf{120.37.}$$

Now, we consider this data once again and change the measuring units for sugar :

Item	*Unit*	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$
Sugar	quintal	600	700	116.67
Milk	litre	3	4	133.33
Ghee	kg	45	50	111.11
				$\Sigma P = 361.11$

$$\therefore \qquad P_{01} = \frac{\Sigma P}{n} = \frac{361.11}{3} = \mathbf{120.37.}$$

We see that this index number is same as that for the data when the rate of sugar was expressed in kg.

Thus, the index number as calculated by this method is not affected by changing measuring units.

In averaging the price relatives, we can also make use of median, harmonic mean etc. But, only A.M. and G.M. are generally used for this purpose.

Example 1. *Construct index number for each year from the following annual wholesale prices of cotton with 1984 as base.*

Year	*Wholesale price (in ₹)*	*Year*	*Whole sale price (in ₹)*
1984	*75*	*1989*	*70*
1985	*50*	*1990*	*69*
1986	*65*	*1991*	*75*
1987	*60*	*1992*	*84*
1988	*72*	*1993*	*80*

Solution. **Calculation of Index Nos. (1984 = 100)**

Year	*Whole sale price (in ₹)*	*Index No. (1984 = 100)*
1984	75	**100**
1985	50	$\frac{50}{75} \times 100 = \mathbf{66.67}$
1986	65	$\frac{65}{75} \times 100 = \mathbf{86.67}$
1987	60	$\frac{60}{75} \times 100 = \mathbf{80}$
1988	72	$\frac{72}{75} \times 100 = \mathbf{96}$
1989	70	$\frac{70}{75} \times 100 = \mathbf{93.33}$
1990	69	$\frac{69}{75} \times 100 = \mathbf{92}$
1991	75	$\frac{75}{75} \times 100 = \mathbf{100}$
1992	84	$\frac{84}{75} \times 100 = \mathbf{112}$
1993	80	$\frac{80}{75} \times 100 = \mathbf{106.67}$

Example 2. *Prepare index numbers of price for three years with average price as base.*

Year	*Rate per Rupee*		
	Wheat	*Cotton*	*Oil*
Ist year	*10 seers*	*4 seers*	*3 seers*
IInd year	*9 seers*	*3.5 seers*	*3 seers*
IIIrd year	*9 seers*	*3 seers*	*2.5 seers*

NOTES

Solution. Here the prices of commodities are given in the form of 'quantity prices', we shall convert these quantity prices into money prices.

Price of wheat in the Ist year is 10 seers per rupee.

$\therefore$ Price of 1 maund wheat $= \frac{40}{10} =$ ₹ 4 $\qquad$ ($\because$ 1 maund = 40 seers)

Similarly, we shall express the prices of other commodities per maund.

Index numbers by Simple Aggregative Method

Index no. for Ist year

$$= \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{27.33}{30.10} \times 100 = \mathbf{90.80}$$

Index no. for IInd year

$$= \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{29.20}{30.10} \times 100 = \mathbf{97.01}$$

Index no. for IIIrd year

$$= \frac{\Sigma p_1}{\Sigma p_0} \times 100 = \frac{33.77}{30.10} \times 100 = \mathbf{112.19}$$

Index numbers by Simple A.M. of Price Relatives Method

Index no. for Ist year

$$= \frac{\Sigma P}{n} = \frac{273.26}{3} = \mathbf{91.09}$$

Index no. for IInd year

$$= \frac{\Sigma P}{n} = \frac{295.86}{3} = \mathbf{98.62}$$

Index no. for IIIrd year

$$= \frac{\Sigma P}{n} = \frac{331.03}{3} = \mathbf{110.34.}$$

EXERCISE 8.1

1. Construct price index number for the year 1995, by using the following series. Simple aggregative method is to be used.

Commodity	*A*	*B*	*C*	*D*	*E*
Price (1994) (in ₹)	4	2	6	8	12
Price (1995) (in ₹)	5	2	8	9	10

Commodity	Unit	Ist year		IInd year		IIIrd year		Average price
		p_1	P	p_1	P	p_1	P	p_0
Wheat	Maund	$\frac{40}{10} = 4$	$\frac{4}{4.29} \times 100$ $= 93.24$	$\frac{40}{9} = 4.44$	$\frac{4.44}{4.29} \times 100$ $= 103.50$	$\frac{40}{9} = 4.44$	$\frac{4.44}{4.29} \times 100$ $= 103.50$	$\frac{4 + 4.44 + 4.44}{3}$ $= 4.29$
Cotton	Maund	$\frac{40}{4} = 10$	$\frac{10}{11.59} \times 100$ $= 86.28$	$\frac{40}{3.5} \times 11.43$	$\frac{11.43}{11.59} \times 100$ $= 98.62$	$\frac{40}{3} = 13.33$	$\frac{13.33}{11.59} \times 100$ $= 115.01$	$\frac{10 + 11.43 + 13.33}{3}$ $= 11.59$
Oil	Maund	$\frac{40}{3} = 13.33$	$\frac{13.33}{14.22} \times 100$ $= 93.74$	$\frac{40}{3} = 13.33$	$\frac{13.33}{14.22} \times 100$ $= 93.74$	$\frac{40}{2.5} = 16$	$\frac{16}{14.22} \times 100$ $= 112.52$	$\frac{13.33 + 13.33 + 16}{3}$ $= 14.22$
Total		27.33	273.26	29.20	295.86	33.77	331.03	30.10

2. For the data given below, calculate the index numbers by taking :
   (i) 1980 as the base year
   (ii) 1982 as the base year.

Year	1978	1979	1980	1981	1982	1983	1984	1985
Price of 'x' (in ₹)	4	7	10	10	12	11	15	16

3. Find the price index numbers for three years by simple aggregative method. The average price is to be used as base :

	Price per rupee		
Year	*A*	*B*	*C*
I	4 kg	2 kg	10 kg
II	5 kg	2.5 kg	12 kg
III	3 kg	2.5 kg	8 kg

4. From the following data, construct the price index numbers with average price as base :

	Rate per rupee		
Year	*Wheat*	*Rice*	*Oil*
I	10 kg	5 kg	2 kg
II	8 kg	4 kg	1.33 kg
III	6.67 kg	3.33 kg	1 kg

Answers

1. 106.25
2. (i) 40, 70, 100, 100, 120, 110, 150, 160
   (ii) 33.33, 58.33, 83.33. 83.33; 100, 91.67, 125, 133.33
3. 106.62, 85.71, 107.66
4. 75.23, 100, 124.4 by using simple A.M. of price relative method.

8.5. LASPEYRE'S METHOD

This is a method for finding weighted index numbers. In this method, base period quantities (q_0) are used as weights. If P_{01} is the index number for the current period, then we have

$$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

where '0' and '1' suffixes stand for base period and current period respectively.

$\Sigma p_1 q_0$ = sum of products of prices of the commodities in the current period with their corresponding quantities used in the base period.

$\Sigma p_0 q_0$ = sum of product of prices of the commodities in the base period with their corresponding quantities used in the base period.

NOTES

8.6. PAASCHE'S METHOD

This is a method for finding weighted index numbers. In this methods, current period quantities (q_1) are used as weights.

If P_{01} is the required index number for the current period, then

$$P_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$

where p_0, p_1 represents prices per unit of commodities in the base period and current period respectively.

8.7. DORBISH AND BOWLEY'S METHOD

This is a method for computing weighted index numbers.

If P_{01} is the required index number for the current period, then

$$P_{01} = \frac{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right)}{2} \times 100$$

where p_0, p_1 represents prices per unit of commodities in the base period and current period respectively, q_0, q_1 represents number of units in the base period and current period respectively.

We have $$P_{01} = \frac{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right)}{2} \times 100 = \frac{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100}{2}$$

$$= \frac{\text{Laspeyre's index no} + \text{Paasche's index no.}}{2}$$

∴ Dorbish and Bowely's index number can also be obtained by taking A.M. of Laspeyre's and Paasche's index numbers.

8.8. FISHER'S METHOD

This is a method for computing weighted index numbers.

If P_{01} is the required index number for the current period, then

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$

where symbols p_0, q_0, p_1, q_1 have their usual meaning.

We have $$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100 = \sqrt{\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100\right)\left(\frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100\right)}$$

$$= \sqrt{\left(\begin{matrix}\text{Laspeyre's} \\ \text{Index no.}\end{matrix}\right)\left(\begin{matrix}\text{Paasche's} \\ \text{Index no.}\end{matrix}\right)}$$

∴ Fisher's index numbers can also be obtained by taking G.M. of Laspeyre's and Paasche's index numbers. Fisher's method is considered to be the best method of computing index numbers because this method, satisfies unit test, time reversal test and factor reversal test. That is why, this method is also known as *Fisher's Ideal Method.*

NOTES

8.9. MARSHALL EDGEWORTH'S METHOD

This is a method of computing weighted index numbers. In this method the sum of base period quantities and current period quantities are used as weights.

If P_{01} is the required index number for the current period, then

$$P_{01} = \frac{\Sigma p_1(q_0 + q_1)}{\Sigma p_0(q_0 + q_1)} \times 100$$

where p_0, q_0, p_1, q_1 have their usual meaning.

We can also write this index numbers as

$$P_{0i} = \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

This form is generally used for computing index numbers.

8.10. KELLY'S METHOD

This is a method of computing weighted index numbers. In this method, the quantities (q) corresponding to any period can be used as weights. We can also use the average of quantities for two or more periods as weights.

If P_{01} is the required index numbers for the current period, then

$$P_{01} = \frac{\Sigma p_1 q}{\Sigma p_0 q} \times 100$$

where q represents the quantities which are to be used as weights. p_0, p_1 have their usual meanings. This index number is also known as **Fixed Weights Aggregative Method.**

8.11. WEIGHTED AVERAGE OF PRICE RELATIVES METHOD

This is a method of computing weighted index numbers. In weighted index numbers, we give weights to every commodity in the series so that each commodity may have due influence on the index number. Till now quantity weights were used for constructing price index numbers.

In the weighted average of price relatives method, value weights (W) are used. The values of commodities may correspond to either base period or current period or any other period.

If P_{01} is the required index number for the current period, then

$$P_{01} = \frac{\Sigma WP}{\Sigma W}, \quad \text{where } P = \frac{p_1}{p_0} \times 100.$$

p_0, p_1 have their usual meanings.

In this method, we have infact taken the weighted arithmetic mean of the price relatives. In constructing this index number, geometric mean is also used. In this case, the formula is

$$P_{01} = \text{Antilog}\left(\frac{\Sigma W \log P}{\Sigma W}\right).$$

Example 3. *Calculate Laspeyre's and Paasche's price index numbers for the year 1991 from the following data :*

Commodity	*1981*		*1991*	
	Quantity (in kg)	*Price (in ₹)*	*Quantity (in kg)*	*Price (in ₹)*
Wheat	*60*	*1.00*	*50*	*1.25*
Rice	*25*	*1.50*	*20*	*2.50*
Sugar	*10*	*2.00*	*10*	*3.00*
Ghee	*3*	*12.00*	*4*	*18.00*
Fuel	*40*	*0.10*	*60*	*0.15*

Solution. **Calculation of Index Nos. (1981 = 100)**

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1	p_0q_1	p_1q_0
Wheat	1.00	60	1.25	50	60.0	62.5	50.0	75.0
Rice	1.50	25	2.50	20	37.5	50.5	30.0	62.5
Sugar	2.00	10	3.00	10	20.0	30.0	20.0	30.0
Ghee	12.00	3	18.00	4	36.0	72.0	48.0	54.0
Fuel	0.10	40	0.15	60	4.0	9.0	6.0	6.0
Total					157.5	223.5	154.0	227.5

Laspeyre's price index number $= \dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \dfrac{227.5}{157.5} \times 100 = \mathbf{144.44.}$

Paasche's price index number $= \dfrac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100 = \dfrac{223.5}{154.0} \times 100 = \mathbf{145.13.}$

Example 4. *Calculate Fisher's Ideal Index No. from the following informations and also give three reversibility tests :*

Item	*Base Year*		*Current Year*	
	Price	*Quantity*	*Price*	*Quantity*
A	*2*	*4*	*6*	*5*
B	*4*	*5*	*8*	*4*
C	*6*	*2*	*9*	*3*
D	*8*	*1*	*6*	*2*
E	*10*	*1*	*5*	*2*

NOTES

Solution. **Calculation of Index Number**

Item	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1	p_0q_1	p_1q_0
A	2	4	6	5	8	30	10	24
B	4	5	8	4	20	32	16	40
C	6	2	9	3	12	27	18	18
D	8	1	6	2	8	12	16	6
E	10	1	5	2	10	10	20	5
Total					58	111	80	93

Fisher's price index number

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 = \sqrt{\frac{93}{58} \times \frac{11}{80}} \times 100 = \mathbf{149.16.}$$

The reversibility tests : Time Reversal Test, Factor Reversal Test and Circular Tests are discussed in the section "**Test of adequacy of index numbers**".

Example 5. *Calculate Paasche's index number and Fisher's ideal index number for the year 1995 from the following data :*

Commodity	*1992*		*1995*	
	Price	*Quantity*	*Price*	*Quantiy*
A	*6*	*50*	*10*	*56*
B	*2*	*100*	*2*	*120*
C	*4*	*60*	*6*	*60*
D	*10*	*30*	*12*	*24*
E	*8*	*40*	*12*	*36*

Solution. **Calculation of Index Nos. (1992 = 100)**

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1	p_0q_1	p_1q_0
A	6	50	10	56	300	560	336	500
B	2	100	2	120	200	240	240	200
C	4	60	6	60	240	360	240	360
D	10	30	12	24	300	288	240	360
E	8	40	12	36	320	432	288	480
Total					1360	1880	1344	1900

Paasche's price index number $= \dfrac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100 = \dfrac{1880}{1344} \times 100 = \mathbf{139.88.}$

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

$$= \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}} \times 100 = \mathbf{139.79.}$$

Example 6. *Construct index numbers of price for the year 1994 from the following data by applying :*

1. Laspeyre's method *2. Paasche's method*

3. Bowley's method *4. Fisher's method*

5. Marshall Edgeworth's method

Commodity	*1993*		*1994*	
	Price	*Quantity*	*Price*	*Quantity*
A	*2*	*8*	*4*	*6*
B	*5*	*10*	*6*	*5*
C	*4*	*14*	*5*	*10*
D	*2*	*19*	*2*	*13*

Solution. **Calculation of Index Nos. (1993 = 100)**

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1	p_0q_1	p_1q_0
A	2	8	4	6	16	24	12	32
B	5	10	6	5	50	30	25	60
C	4	14	5	10	56	50	40	70
D	2	19	2	13	38	26	26	38
Total					160	130	103	200

Laspeyre's price index number $= \dfrac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100 = \dfrac{200}{160} \times 100 = \mathbf{125.}$

Paasche's price index number $= \dfrac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100 = \dfrac{130}{103} \times 100 = \mathbf{126.21}$

Bowley's price index number

$$= \frac{\left(\dfrac{\Sigma p_1q_0}{\Sigma p_0q_0} + \dfrac{\Sigma p_1q_1}{\Sigma p_0q_1}\right)}{2} \times 100 = \frac{\left(\dfrac{200}{160} + \dfrac{130}{103}\right)}{2} \times 100 = \mathbf{125.607.}$$

Fisher's price index number

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 = \sqrt{\frac{200}{160} \times \frac{130}{103}} \times 100 = \mathbf{125.605.}$$

Marshall Edgeworth's price index number

$$= \frac{\Sigma p_1(q_0 + q_1)}{\Sigma p_0(q_0 + q_1)} \times 100 = \frac{\Sigma p_1q_0 + \Sigma p_1q_1}{\Sigma p_0q_0 + \Sigma p_0q_1} \times 100$$

$$= \frac{200 + 130}{160 + 103} \times 100 = \mathbf{125.47.}$$

Example 7. *Prepare the index number for 1982 on the basis of 1962 for the following data :*

Year	*Commodity A*		*Commodity B*		*Commodity C*	
	Price	*Expenditure*	*Price*	*Expenditure*	*Price*	*Expenditure*
1962	*5*	*50*	*8*	*48*	*6*	*24*
1982	*4*	*48*	*7*	*49*	*5*	*15*

Solution. We calculate price index number for the year 1982 by using **Fisher's method.**

Calculation of Index Number

Commodity	*1962*			*1982*			p_0q_1	p_1q_0
	p_0	p_0q_0	q_0	p_1	p_1q_1	q_1		
A	5	50	10	4	48	12	60	40
B	8	48	6	7	49	7	56	42
C	6	24	4	5	15	3	18	20
Total		122			112		134	102

NOTES

Fisher's price index number

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 = \sqrt{\frac{102}{122} \times \frac{112}{134}} \times 100 = \mathbf{83.59.}$$

Example 8. *Show that Fisher's price index number lies between Laspeyre's and Paasche's price index numbers.*

Solution. Let L, P and F represent Laspeyre's Pasasche's and Fisher's price index numbers respectively.

$$\therefore \quad L = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100,\ P = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100$$

and
$$F = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100$$

$$\sqrt{LP} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100 \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1} \times 100} = \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} \times 100 = F.$$

Also, L, P, F are positive numbers.

Let $\quad L < P.$

$\therefore \quad L < P \Rightarrow LL < LP \Rightarrow \sqrt{LL} < \sqrt{LP} \Rightarrow L < F$

Also, $\quad L < P \Rightarrow LP < PP \Rightarrow \sqrt{LP} < \sqrt{PP} \Rightarrow F < P$

$\therefore \quad \mathbf{L < F < P.}$

EXERCISE 8.2

1. From the following data calculate price index by using:
   (*i*) Laspeyre's method
   (*ii*) Paasche's method.

Commodity	*Base year*		*Current year*	
	Quantity	*Price*	*Quantity*	*Price*
A	20	4	30	6
B	40	5	60	7
C	60	3	70	4
D	30	2	50	3

NOTES

2. Calculate price index numbers for the year 1990, by using the following methods:
(i) Laspeyre's method (ii) Paasche's method
(iii) Bowley's method (iv) Fisher's method
(v) Marshall's method.

Commodity	1989		1990	
	Price	Quantity	Price	Quantity
A	20	8	40	6
B	50	10	60	5
C	40	15	50	15
D	20	20	20	25

3. Compute price index number by using Fisher's method.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	10	12	12	15
B	7	15	5	20
C	5	24	9	20
D	16	5	14	5

4. From the following data construct a price index number of the group of four commodities by using an appropriate formula:

Commodity	Base year		Current year	
	Price per unit	Expenditure (in ₹)	Price per unit	Expenditure (in ₹)
A	2	40	5	75
B	4	16	8	40
C	1	10	2	24
D	5	25	10	60

5. Calculate weighted aggregative price index number taking 1992 as base, from the following data:

Commodity	Quantity consumed in 1992	Units	Price in Base year 1992	Price in current year 1997
Wheat	4 Qtls	per Qtl	80	100
Rice	1 Qtl	per Qtl	120	250
Gram	1 Qtl	per Qtl	100	150
Pulses	2Qtls	per Qtl	200	300

6. Construct Fisher's and Marshall's price index numbers by using the following data:

Commodity	Base year price	Base year quantity	Current year price	Current year quantity
A	12	100	20	120
B	4	200	4	240
C	8	120	12	120
D	20	60	24	48
E	16	80	24	52

NOTES

7. The prices of four different commodities for 1995 and 1996 are given below. Calculate the price index number for 1996 with 1989 as base, by using weighted A.M. of price relatives:

Commodity	*Weight*	*Price in 1995 (in ₹)*	*Price in 1996 (in ₹)*
A	5	4.5	2.0
B	7	3.2	2.5
C	6	4.5	3.0
D	2	1.8	1.0

8. It is stated that Marshal Edge worth index number is a good approximation to the Fisher's index number. Verify this by using the following data:

Item	*1999*		*2001*	
	Price	*Quantity*	*Price*	*Quantity*
A	2	74	3	82
B	5	125	4	140
C	7	40	6	33

9. Given the data :

Commodity

	A	*B*
p_0	1	1
q_0	10	5
p_1	2	x
q_1	5	2

where p and q respectively stand for price and quantity and subscripts stand for time periods. Find x if the ratio between Laspeyre's (L) and Paasche's (P) index numbers is L : P :: 28 : 27.

Answers

1. (*i*) 140.38 (*ii*) 141.14
2. 124.699, 121.769, 123.234, 123.225, 123.323. **3.** 115.75
4. Fisher's price index no. = 219.12 **5.** 148.94 **6.** 139.729, 139.728
7. 64.01 **9.** 4

8.12. CHAIN BASE METHOD

In this method of computing index numbers, link relatives are required. The prices of commodities in the current period are expressed as the percentages of their prices in the preceding period. These are called **link relatives.**

Mathematically,

$$\textbf{Link Relative (L.R.)} = \frac{\textbf{Price in current period}}{\textbf{Price in preceding period}} \times 100$$

If there are more than one commodity under consideration then averages of link relatives (A.L.R.) are calculated for each period. Generally A.M. is used for averaging link relatives. These averages of link relatives (A.L.R.) for different time periods are called **chain index numbers.** The chain index number of a particular period represent the index number of that period with preceding period as the base period. This would be so except for this first period.

These chain indices can further be used to get index numbers for various periods with a particular period as the base period. These index numbers are called **chain index numbers chained to a fixed base.**

For calculating these index numbers, the following formula is used :

C.B.I. for current period (Base fixed)

$$= \frac{\textbf{A.L.R. for current period} \times \textbf{C.B.I. for preceding period (Base fixed)}}{\textbf{100}}$$

There are certain advantages of using this method. By using chain base method, comparison is possible between any two successive periods. The average of link relatives represent the index number with preceding period as the base period. This characteristic of chain base index numbers benefit businessmen to a good extent. In calculating chain base index number, some items can be introduced or withdrawned during any period. In practice, the chain base index numbers are used only in those circumstances, where the list of items changes very frequently.

Example 9. *From the following data, find index numbers with 1998 as base by using (i) fixed base method (ii) chain base method :*

Year	*1998*	*1999*	*2000*	*2001*	*2002*
Price per unit (in ₹)	*40*	*50*	*60*	*75*	*120*

Solution. Calculation of index numbers (1998 = 100)

Year	*p*	*F.B.I.*	*Link Relative*	*C.B.I.*
1998	40	100	100	100
1999	50	$\frac{50}{40} \times 100 = \mathbf{125}$	$\frac{50}{40} \times 100 = 125$	$\frac{125 \times 100}{100} = \mathbf{125}$
2000	60	$\frac{60}{40} \times 100 = \mathbf{150}$	$\frac{60}{50} \times 100 = 120$	$\frac{120 \times 125}{100} = \mathbf{150}$
2001	75	$\frac{75}{40} \times 100 = \mathbf{187.5}$	$\frac{75}{60} \times 100 = 125$	$\frac{125 \times 150}{100} = \mathbf{187.5}$
2002	120	$\frac{120}{40} \times 100 = \mathbf{300}$	$\frac{120}{75} \times 100 = 160$	$\frac{160 \times 187.5}{100} = \mathbf{300}$

∴ F.B.I. for 1999, 2000, 2001, 2002 with base 1998 are **125, 150, 187.5, 300** respectively.

C.B.I. for 1999, 2000, 2001, 2002 with base 1998 are **125, 150, 187.5, 300** respectively.

Remark. If there is only one series then F.B.I. and C.B.I. with fix base are always same.

Example 10. *The average wholesale prices of three groups of commodities for the years 1988 to 1992 are given below. Compute chain base index numbers with 1988 as base :*

Group	*1988*	*1989*	*1990*	*1991*	*1992*
I	*6*	*9*	*15*	*21*	*24*
II	*24*	*30*	*36*	*42*	*54*
III	*12*	*15*	*21*	*27*	*36*

NOTES

Solution. **Calculation of C.B.I. (1988 = 100)**

Group	*Link Relatives*				
	1988	*1989*	*1990*	*1991*	*1992*
I	100	$\frac{9}{6} \times 100 = 150$	$\frac{15}{9} \times 100 = 166.67$	$\frac{21}{15} \times 100 = 140$	$\frac{24}{21} \times 100 = 114.29$
II	100	$\frac{30}{24} \times 100 = 125$	$\frac{36}{30} \times 100 = 120$	$\frac{42}{36} \times 100 = 116.67$	$\frac{54}{42} \times 100 = 128.57$
III	100	$\frac{15}{12} \times 100 = 125$	$\frac{21}{15} \times 100 = 140$	$\frac{27}{21} \times 100 = 128.57$	$\frac{36}{27} \times 100 = 133.33$
Total	300	400	426.67	385.24	376.19
Average of L.R. or C.B.I.	100	$\frac{400}{3} = 133.33$	$\frac{426.67}{3} = 142.22$	$\frac{385.24}{3} = 128.41$	$\frac{376.19}{3} = 125.40$
C.B.I. (1988 = 100)	100	$\frac{133.33 \times 100}{100}$ **= 133.33**	$\frac{142.22 \times 133.33}{100}$ **= 189.62**	$\frac{128.41 \times 189.62}{100}$ **= 243.49**	$\frac{125.40 \times 243.49}{100}$ **= 305.34**

∴ C.B.I. for years 1989, 1990, 1991, 1992 with base 1988 are **133.33, 189.62, 243.49, 305.34** respectively.

Example 11. *Construct, by chain base method, index number of prices in Kanpur on base 1990, for the following data :*

Commodity	*Year*			
	1990	*1991*	*1992*	*1993*
Rice	*7.5*	*8.0*	*6.0*	*5.5*
Wheat	*8.0*	*6.0*	*5.5*	*5.0*
Pulses	*7.0*	*8.0*	*6.5*	*5.5*
Gur	*6.5*	*7.5*	*6.0*	*5.0*
Cotton	*34.0*	*30.0*	*28.0*	*25.0*

Solution. **Calculation of C.B.I. (1990 = 100)**

Commodity	*Link Relatives*			
	1990	*1991*	*1992*	*1993*
Rice	100	$\frac{8}{7.5} \times 100 = 106.67$	$\frac{6}{8} \times 100 = 75$	$\frac{5.5}{6} \times 100 = 91.67$
Wheat	100	$\frac{6}{8} \times 100 = 75$	$\frac{5.5}{6} \times 100 = 91.67$	$\frac{5}{5.5} \times 100 = 90.91$
Pulses	100	$\frac{8}{7} \times 100 = 114.29$	$\frac{6.5}{8} \times 100 = 81.25$	$\frac{5.5}{6.5} \times 100 = 84.61$
Gur	100	$\frac{7.5}{6.5} \times 100 = 115.38$	$\frac{6}{7.5} \times 100 = 80$	$\frac{5}{6} \times 100 = 83.33$
Cotton	100	$\frac{30}{34} \times 100 = 88.23$	$\frac{28}{30} \times 100 = 93.33$	$\frac{25}{28} \times 100 = 89.29$
Total	500	499.57	421.25	439.81

A.L.R. or C.B.I.	100	$\frac{499.5}{3} = 99.91$	$\frac{421.25}{3} = 84.25$	$\frac{439.81}{3} = 87.96$
C.B.I. (1990 = 100)	100	$\frac{99.91 \times 100}{100} = \mathbf{99.91}$	$\frac{84.25 \times 99.91}{100}$ $= \mathbf{84.17}$	$\frac{87.96 \times 84.17}{100}$ $= \mathbf{74.04}$

∴ C.B.I. for years 1991, 1992, 1993 with base 1990 are **99.91, 84.17, 74.04** respectively.

EXERCISE 8.3

1. From the following data, find index numbers with 1996 as base by using (*i*) fixed base method (*ii*) chain base method. Verify that the index numbers are same :

Year	*1996*	*1997*	*1998*	*1999*	*2000*	*2001*	*2002*
Price per unit (in ₹)	5	7	10	8	15	12	17

2. Compute the fixed base index numbers and chain base index numbers with 1982 as base, for the following data :

Commodity	*Price (in ₹)*				
	1982	*1983*	*1984*	*1985*	*1986*
A	2	3	6	6	9
B	10	10	10	15	15
C	4	6	12	15	18

3. The following table gives the average wholesale prices of three groups of commodities for the year 1993 to 1996. Compute chain base index number chained to 1993.

Group	*Year*			
	1993	*1994*	*1995*	*1996*
I	400	400	550	600
II	225	400	300	350
III	400	400	425	500

Answers

1. 100, 140, 200, 160, 300, 240, 340

2. 100, 133.33, 233.33, 275, 350 ; 100, 133.33, 222.22, 277.77, 342.57

3. 100, 125.93, 133.80, 153.16

II. QUANTITY INDEX NUMBERS

8.13. METHODS OF QUANTITY INDEX NUMBERS

NOTES

Quantity index numbers are used to show the average change in the quantities of related goods with respect to time. These index numbers are also used to measure the level of production. In computing quantiy index numbers, either prices or values are used as weights.

Let Q_{01} denotes the quantity index number for the current period. The formulae for calculating quantity index numbers are obtained by interchanging the role of 'p' and 'q' in the formulae for computing price index numbers. Various methods for computing quantity index-numbers are as follows :

1. Simple Aggregative Method

$$Q_{01} = \frac{\Sigma q_1}{\Sigma q_0} \times 100.$$

2. Simple Average of Quantity Relative Method

$$Q_{01} = \frac{\Sigma Q}{n} \qquad \text{(Using A.M.)}$$

$$= \text{Antilog}\left(\frac{\Sigma \log Q}{n}\right) \qquad \text{(Using G.M.)}$$

where Q = quantity relative = $\frac{q_1}{q_0} \times 100$.

3. Laspeyre's Method

$$Q_{01} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100.$$

4. Paasche's Method

$$Q_{01} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100.$$

5. Dorbish and Bowley's Method

$$Q_{01} = \frac{\left(\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} + \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}\right)}{2} \times 100.$$

6. Fisher's Ideal Method

$$Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100.$$

7. Marshall Edgeworth's Method

$$Q_{01} = \frac{\Sigma q_1 (p_0 + p_1)}{\Sigma q_0 (p_0 + p_1)} \times 100.$$

8. Kelly's Method

$$Q_{01} = \frac{\Sigma q_1 p}{\Sigma q_0 p} \times 100.$$

NOTES

9. Weighted Average of Quantity Relative Method

$$Q_{01} = \frac{\Sigma WQ}{\Sigma W} \qquad \text{(Using A.M.)}$$

$$= \text{Antilog}\left(\frac{\Sigma W \log Q}{\Sigma W}\right) \qquad \text{(Using G.M.)}$$

10. Chain Base Method

Here also, we define chain base quantity index numbers for a period as the average of link relatives (L.R.) for that particular period. These chain indices can be used to obtain quantity index numbers with a common base.

In all the above formulae, suffixes '0' and '1' stand for base period and current period respectively and

p_1 = current period price of an item

p_0 = base period price of an item

q_1 = current period quantity of an item

q_0 = base period quantity of an item

Q = quantity relative of an item = $\frac{q_1}{q_0} \times 100$

W = value weight for an item

p = price of an item in a fixed period

n = no. of item under consideration.

8.14. INDEX NUMBERS OF INDUSTRIAL PRODUCTION

The indices of industrial production are calculated by using the methods of quantity index numbers. In the formulae for quantity index numbers, we shall take *production* in place of quantities.

Example 12. *From the following data, construct the index of industrial production for the year 1996 and 1997 by the methods :*

(*i*) *Simple aggregative method.*

(*ii*) *Simple A.M. of production relatives.*

(*iii*) *Simple G.M. of production relatives.*

Commodity	*Annual Production*		
	1995	*1996*	*1997*
A	*20,000 units*	*25,000 units*	*26,000 units*
B	*4,000 units*	*5,000 units*	*4,000 units*
C	*7,000 units*	*7,000 units*	*12,000 units*

Solution. Let the suffixes 0, 1, 2 refer to the data relating to years 1995, 1996 and 1997 respectively.

Calculation of Index Numbers

Commodity	*Annual Production*			*Production Relatives*			
	q_0	q_1	q_1	$Q_1 = \frac{q_1}{q_0} \times 100$	$Q_2 = \frac{q_2}{q_0} \times 100$	$\log Q_1$	$\log Q_2$
A	20,000	25,000	26,000	125	130	2.0969	2.1139
B	4,000	5,000	4,000	125	100	2.0969	2.0000
C	7,000	7,000	12,000	100	171.43	2.0000	2.2340
Total	31,000	37,000	42,000	350	401.43	6.1938	6.3479

(*i*) Index of industrial production of 1996 with base 1995

$$= Q_{01} = \frac{\Sigma q_1}{\Sigma q_0} \times 100 = \frac{37000}{31000} \times 100 = \mathbf{119.35}$$

Index of industrial production of 1997 with base 1995

$$= Q_{02} = \frac{\Sigma q_2}{\Sigma q_0} \times 100 = \frac{42000}{31000} \times 100 = \mathbf{135.48.}$$

(*ii*) Index of industrial production of 1996 with base 1995

$$= Q_{01} = \frac{\Sigma Q_1}{n} = \frac{350}{3} = \mathbf{116.67}$$

Index of industrial production of 1997 with base 1995

$$= Q_{02} = \frac{\Sigma Q_2}{n} = \frac{40143}{3} = \mathbf{133.81.}$$

(*iii*) Index of industrial production of 1996 with base 1995

$$= Q_{01} = \text{AL}\left(\frac{\Sigma \log Q_1}{n}\right) = \text{AL}\left(\frac{6.1938}{3}\right) = \text{AL}\ (2.0646) = \mathbf{116.10}$$

Index of industrial production of 1997 with base 1995

$$= Q_{02} = \text{AL}\left(\frac{\Sigma \log Q_2}{n}\right) = \text{AL}\left(\frac{6.3479}{3}\right) = \text{AL}\ (2.1160) = \mathbf{130.60.}$$

Example 13. *From the following data, construct quantity index numbers for 1986, by using the following methods :*

(*i*) *Simple aggregative method*
(*ii*) *Laspeyre's method*
(*iii*) *Paasche's method*
(*iv*) *Dorbish and Bowley's method*
(*v*) *Fisher's method*
(*vi*) *Marshall Edgeworth's method*

Commodity	*1995*		*1996*	
	Price	*Value*	*Price*	*Value*
A	*8*	*80*	*10*	*110*
B	*10*	*90*	*12*	*108*
C	*16*	*256*	*20*	*340*

Solution. **Calculation of Quantity Index Nos. (1995 = 100)**

NOTES

Commodity	p_0	Value q_0p_0	q_0	p_1	Value q_1p_1	q_1	q_1p_0	q_0p_1
A	8	80	10	10	110	11	88	100
B	10	90	9	12	108	9	90	108
C	16	256	16	20	340	17	272	320
Total		426	35		558	37	450	528

(*i*) Q_{01} by simple aggregative method

$$= \frac{\Sigma q_1}{\Sigma q_0} \times 100 = \frac{37}{35} \times 100 = \mathbf{105.71}$$

(*ii*) Laspeyre's quantity index no.

$$= \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100 = \frac{450}{426} \times 100 = \mathbf{105.63}$$

(*iii*) Paasche's quantity index no.

$$= \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100 = \frac{558}{528} \times 100 = \mathbf{105.68}$$

(*iv*) Dorbish and Bowley's quantity index no.

$$= \frac{\left(\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} + \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}\right)}{2} \times 100 = \frac{\left(\frac{450}{426} + \frac{558}{528}\right)}{2} \times 100 = \mathbf{105.66}$$

(*v*) Fisher's quantity index no.

$$= \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100 = \sqrt{\frac{450}{426} \times \frac{558}{528}} \times 100 = \mathbf{105.66}$$

(*vi*) Marshall Edgeworth's quantity index no.

$$= \frac{\Sigma q_1(p_0 + p_1)}{\Sigma q_0(p_0 + p_1)} \times 100 = \frac{\Sigma q_1 p_0 + \Sigma q_1 p_1}{\Sigma q_0 p_0 + \Sigma q_0 p_1} \times 100$$

$$= \frac{450 + 558}{426 + 528} \times 100 = \mathbf{105.66.}$$

III. VALUE INDEX NUMBERS

8.15. SIMPLE AGGREGATIVE METHOD OF VALUE INDEX NUMBERS

The simple aggregative method of computing value index number (V_{01}) is given by

$$V_{01} = \frac{\Sigma \mathbf{p}_1 \mathbf{q}_1}{\Sigma \mathbf{p}_0 \mathbf{q}_0} \times 100$$

where $\Sigma p_1 q_1$ = sum of values of items in the current period

$\Sigma p_0 q_0$ = sum of values of items in the base period.

Example 14. *Calculate value index number for 2000 for the following data :*

Item	*1998*		*2000*	
	Price	*quantity*	*Price*	*Quantity*
A	*4*	*12*	*5*	*24*
B	*8*	*15*	*12*	*10*
C	*12*	*6*	*10*	*8*
D	*5*	*10*	*5*	*12*

NOTES

Solution. **Calculation of value index number (1998 = 100)**

Item	p_0	q_0	p_1	q_1	p_0q_0	p_1q_1
A	4	12	5	18	48	120
B	8	15	12	10	120	120
C	12	6	10	8	72	80
D	5	10	5	12	50	60
Total					290	380

$$\text{Value index number} = \frac{\Sigma p_1q_1}{\Sigma p_0q_0} \times 100 = \frac{380}{290} \times 100 = \mathbf{131.03.}$$

EXERCISE 8.4

1. Compute by Fisher's index formula, the quantity index number for the data given below :

Commodity	*Base year*		*Current year*	
	Price	*Total value*	*Price*	*Total value*
A	10	100	8	96
B	16	96	14	98
C	12	36	10	40

2. By using the following methods, calculate quantity index numbers for the year 1995 for the following data :

(*i*) Simple A.M. of quantity relatives (*ii*) Laspeyre's method
(*iii*) Paasche's method (*iv*) Dorbish's method
(*v*) Fisher's method (*vi*) Marshall's method.

Commodity	*1993*		*1995*	
	Price	*Quantity*	*Price*	*Quantity*
A	4	10	5	12
B	6	8	7	10
C	10	5	12	4
D	3	12	4	15
E	5	7	5	8

3. Calculate quantity index numbers for the given data, by using the following methods :

(*i*) Dorbish's method (*ii*) Fisher's method
(*iii*) Marshall's method.

NOTES

Item	*Base year*		*Current year*	
	Price per unit	*Quantity*	*Price per unit*	*Quantity*
A	5	7	7	4
B	3	2	4	3
C	1	5	1	5
D	4	4	3	6
E	2	8	2	10
F	1	2	2	6
G	4	5	6	4

4. Calculate value index number for the following data :

Commodity	*Base year*		*Current year*	
	Price	*Quantity*	*Price*	*Quantity*
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Answers

1. 120.654
2. 112.857, 111.483, 111.647, 111.565, 111.565, 111.572
3. 97.984, 97.963, 97.768 4. 81.25.

8.16. MEAN OF INDEX NUMBERS

If $I_1, I_2, ..., I_n$ are the index numbers of n groups of related items, then the index numbers of all the items of n group taken together is calculated by taking the average of these index numbers. Generally, A.M. is used for averaging the index numbers. If weights are attached with different index numbers, then weighted A.M. is to be calculated.

Let I be the index number of all the items of n groups taken together, then

$$I = \frac{I_1 + I_2 + ... + I_n}{n} \quad i.e., \quad I = \frac{\Sigma I}{n}.$$

If $W_1, W_2,......, W_n$ be the weights of index numbers $I_1, I_2, ..., I_n$ respectively, then

$$I = \frac{W_1I_1 + W_2I_2 + ... + W_nI_n}{W_1 + W_2 + ... + W_n} \quad \text{or} \quad I = \frac{\Sigma WI}{\Sigma W}.$$

If G.M. is to be used for finding index number of combined group, then

$$I = AL\left(\frac{W_1 \log I_1 + W_2 \log I_2 + ... + W_n \log I_n}{W_1 + W_2 + ... + W_n}\right) \quad \text{or} \quad I = AL\left(\frac{\Sigma W \log I}{SW}\right).$$

Example 15. *Construct the index number of business activity in India for the following data :*

Item	*Weightage*	*Index*
(*i*) *Industrial Production*	*36*	*250*
(*ii*) *Mineral Production*	*7*	*135*
(*iii*) *Internal Trade*	*24*	*200*
(*iv*) *Financial Activity*	*20*	*135*
(*v*) *Exports and Imports*	*7*	*325*
(*vi*) *Shipping Activity*	*6*	*300*

NOTES

Solution. Calculation of Index No. of Business Activity

Item	*Weightage W*	*Index I*	*WI*
(*i*) Industrial Production	36	250	9000
(*ii*) Mineral Production	7	135	945
(*iii*) Internal Trade	24	200	4800
(*iv*) Financial Activity	20	135	2700
(*v*) Exports and Imports	7	325	2275
(*vi*) Shipping Activity	6	300	1800
Total	100		21520

$$\text{Index No. of combined group} = \frac{\Sigma WI}{\Sigma W} = \frac{21520}{100} = \mathbf{215.2.}$$

Example 16. *Calculate the index number of crime for 1994 with 1993 as base.*

Crime group	*1993*	*1994*	*Weight*
Robberies	*13*	*8*	*6*
Car thefts	*15*	*22*	*5*
Cycle thefts	*249*	*185*	*4*
Pocket picking	*328*	*259*	*1*
Thefts by servants	*497*	*448*	*2*

Solution. Calculation of Index No. of Crime

Crime group	C_0	C_1	*W*	$I = \frac{C_1}{C_0} \times 100$	*WI*
Robberies	13	8	6	61.54	369.24
Car thefts	15	22	5	146.67	733.35
Cycle thefts	249	185	4	74.30	297.20
Pocket picking	328	259	1	78.96	78.96
Thefts by servants	497	448	2	90.14	180.28
Total			18		1659.03

$$\text{Index No. of combined group} = \frac{\Sigma WI}{\Sigma W} = \frac{1659.03}{18} = \mathbf{92.17.}$$

EXERCISE 8.5

NOTES

1. Construct index number of combined group for the following data :

Group	*A*	*B*	*C*	*D*	*E*
Index No.	110	95	160	170	200
Weight	4	2	1	1	2

2. The following are the group index numbers and group weights of an average working class family's budget. Construct the cost of living index.

Group	*Food*	*Fuel*	*Clothing*	*Rent*	*Misc.*
Index Nos.	352	220	230	160	190
Weight	48	10	8	12	15

3. The combined index number for the following data is 138.858. You are required to find the group index for the group 'D'.

Group	A	B	C	D	E	F
Index No.	135	152	124	?	107	139
Weight	61	73	19	41	26	82

Answers

1. 136 **2.** 276.409 **3.** 148

IV TESTS OF ADEQUACY OF INDEX NUMBER FORMULAE

8.17. MEANING OF ADEQUENCY OF INDEX NUMBER

We have studied a large number of methods of constructing index numbers. Statisticians have developed certain mathematical criterion for deciding the superiority of one method over others. The following are the tests for judging the adequacy of a particular index number method :

(*i*) Unit Test. (*ii*) Time Reversal Test.
(*iii*) Factor Reversal Test. (*iv*) Circular Test.

8.17.1. Unit Test (U.T.)

An index number method is said to satisfy **unit test** if it is not changed by a change in the measuring units of some items, under consideration. All methods, except simple aggregative method, satisfies this test.

8.17.2. Time Reversal Test (T.R.T.)

NOTES

An index numbers method is said to satisfy **time reversal test,** if

$$I_{01} \times I_{10} = 1$$

where I_{01} and I_{10} are the index numbers for two periods with base period and current period reversed. Here the index numbers I_{01} and I_{10} are not expressed as percentages.

The following methods of constructing index numbers satisfies this test :

(*i*) Simple Aggregative Method.

(*ii*) Simple G.M. of Price (or Quanity) Relatives Method.

(*iii*) Fisher's Method.

(*iv*) Marshall Edgeworth's Method.

(*v*) Kelly's Method.

Now, we shall illustrate this test by verifying its validity for Fisher's price index number method.

We have $$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \quad \text{and} \quad P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

where P_{01} and P_{10} are the price index numbers for the periods t_1 and t_0 with base periods t_0 and t_1 respectively.

Now $$P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$= \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}} = \sqrt{1} = 1.$$

$$\therefore \quad P_{01} \times P_{10} = 1.$$

Example 17. *Calculate price index number for the year 1996 from the following data. Use geometric mean of price relatives. Also reverse the base (1996 as base) and show whether the two results are consistent or not.*

Commodity	*Average price 1990 (₹)*	*Average Price 1996 (₹)*
A	*16.1*	*14.2*
B	*9.2*	*8.7*
C	*15.1*	*12.5*
D	*5.6*	*4.8*
E	*11.7*	*13.4*
F	*100*	*117*

NOTES

Solution.

Index No. for 1996

Commodity	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$	$\log P$
A	16.1	14.2	$\frac{14.2}{16.1} \times 100 = 80.20$	1.9455
B	9.2	8.7	$\frac{8.7}{9.2} \times 100 = 94.57$	1.9757
C	15.1	12.5	$\frac{12.5}{15.1} \times 100 = 82.78$	1.9179
D	5.6	4.8	$\frac{4.8}{5.6} \times 100 = 85.71$	1.9331
E	11.7	13.4	$\frac{13.4}{11.7} \times 100 = 114.53$	2.0589
F	100	117	$\frac{117}{100} \times 100 = 117$	1.0682
$n = 6$				$\Sigma \log P = 11.8993$

$\therefore$ Price index no. for 1996 $= \text{AL}\left(\frac{\Sigma \log P}{n}\right) = \text{AL}\left(\frac{11.8993}{6}\right) = \text{AL } 1.9832 = \mathbf{96.20.}$

Index No. for 1990

Commodity	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$	$\log P$
A	14.2	16.1	$\frac{16.1}{14.2} \times 100 = 113.38$	2.0547
B	8.7	9.2	$\frac{9.2}{8.7} \times 100 = 105.75$	2.0244
C	12.5	15.1	$\frac{15.1}{12.5} \times 100 = 120.80$	2.0820
D	4.8	5.6	$\frac{5.6}{4.8} \times 100 = 116.67$	2.0671
E	13.4	11.7	$\frac{11.7}{13.4} \times 100 = 87.31$	1.9410
F	117	100	$\frac{100}{117} \times 100 = 85.47$	1.9319
$n = 6$				$\Sigma \log P = 12.1011$

$\therefore$ Price index no. for 1990 $= \text{AL}\left(\frac{\Sigma \log \text{P}}{n}\right) = \text{AL}\left(\frac{12.1011}{6}\right) = \text{AL } 2.0169 = 104.$

Product of index numbers = 96.20 × 104 = 10004.8 = 10000 (nearly)

Since the index numbers are expressed as percentages, the T.R.T. is satisfied if their products is $(100)^2$, which is 10000.

$\therefore$ The index numbers are consistent.

8.17.3. Factor Reversal Test (F.R.T.)

NOTES

An index number method is said to satisfy **factor reversal test** if the product of price index number and quantity index number, as calculated by the same method, is equal to the value index number.

In other words, if P_{01} and Q_{01} are the price index number and quantity index number for the period t_1 corresponding to base period t_0, then we must have

$$P_{01} \times Q_{01} = V_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Fisher's index number method is *the only method* which satisfies this test.

Let P_{01} and Q_{01} be the Fisher's price index number and quantity index numbers respectively, then

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \quad \text{and} \quad Q_{01} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}.$$

$$\text{Now} \quad P_{01} \times Q_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}}$$

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

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

= Value index number.

∴ Fisher's method satisfies this test.

8.17.4. Circular Test (C.T.)

An index number method is said to satisfy the **circular test** if I_{01}, I_{12}, I_{23}, ..., I_{n-1n} and I_{n0} are the index numbers for the periods t_1, t_2, t_3, ..., t_n, t_0 corresponding to base periods t_0, t_1, t_2, ..., t_{n-1}, t_n respectively, then

$$I_{01} \times I_{12} \times I_{23} \times \text{......} \times I_{n-1n} \times I_{n0} = 1.$$

Here, also, the index numbers have not been expressed as percentages by multiplying by 100.

If $n = I$, we have $I_{01} \times I_{10} = 1$.

This is nothing but the condition of T.R.T. Thus, we see that the circular test is an extension of T.R.T.

If $n = 2$, we have

$$I_{01} \times I_{12} \times I_{20} = 1 \quad \text{or} \quad I_{01} \times I_{12} = I_{02}. \qquad (\because \quad I_{02} \times I_{20} = 1)$$

The following methods satisfies circular test :

(*i*) Simple Aggregative Method.

(*ii*) Simple G.M. of Price (or Quantity) Relatives Method.

(*iii*) Kelly's Method.

NOTES

Now, we shall illustrate this test by verifying its validity for simple aggregative method for price index numbers.

Here $$P_{01} = \frac{\Sigma p_1}{\Sigma p_0}, P_{12} = \frac{\Sigma p_2}{\Sigma p_1}, P_{20} = \frac{\Sigma p_0}{\Sigma p_2}$$

$$\therefore \quad P_{01} \times P_{12} \times P_{20} = \frac{\Sigma p_1}{\Sigma p_0} \times \frac{\Sigma p_2}{\Sigma p_1} \times \frac{\Sigma p_0}{\Sigma p_2} = 1.$$

∴ Simple aggregative method satisfies this test.

Example 18. *With the help of following data show that Fisher's ideal index satisfies time and factor reversal tests.*

Commodity	*1993*		*1994*	
	Price	*Value*	*Price*	*Value*
A	*8*	*80*	*10*	*100*
B	*10*	*20*	*12*	*36*
C	*5*	*25*	*5*	*30*
D	*4*	*16*	*8*	*40*

Solution. Let suffixes '0' and '1' refers to data for periods 1993 and 1994 respectively.

Verification of T.R.T. and F.R.T. for Fisher's Method

Commodity	p_0	p_0q_0	p_1	p_1q_1	q_0	q_1	p_1q_0	p_0q_1
A	8	80	10	100	10	10	100	80
B	10	20	12	36	2	3	24	30
C	5	25	5	30	5	6	25	30
D	4	16	8	40	4	5	32	20
Total		141		206			181	160

Verification of T.R.T.

P_{01} = Fisher's price index number for 1994 with base 1993 (= 1)

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} = \sqrt{\frac{181}{141} \times \frac{206}{160}} = 1.28559$$

P_{10} = Fisher's price index number for 1993 with base 1994 (= 1)

$$= \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_1q_1} \times \frac{\Sigma p_0q_0}{\Sigma p_1q_0}} = \sqrt{\frac{160}{206} \times \frac{141}{181}} = 0.77785.$$

Now $P_{01} \times P_{10} = 1.28559 \times 0.77785 = 0.9999961 = 1$ (nearly).

∴ T.R.T. is verified.

Verification of F.R.T.

P_{01} = Fisher's price index number for 1994 with base 1993 (= 1)

$$= \sqrt{\frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_0q_1}} = \sqrt{\frac{181}{141} \times \frac{206}{160}} = 1.28559$$

Q_{01} = Fisher's quantity index number for 1994 with base 1993 (= 1)

$$= \sqrt{\frac{\Sigma q_1p_0}{\Sigma q_0p_0} \times \frac{\Sigma q_1p_1}{\Sigma q_0p_1}} = \sqrt{\frac{\Sigma p_0q_1}{\Sigma p_0q_0} \times \frac{\Sigma p_1q_1}{\Sigma p_1q_0}} = \sqrt{\frac{160}{141} \times \frac{206}{181}} = 1.13643.$$

V_{01} = Value index number for 1994 with base 1993 (= 1)

$$= \frac{\Sigma V_1}{\Sigma V_0} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0} = \frac{206}{141} = 1.46099.$$

Now, $P_{01} \times Q_{01} = 1.28559 \times 1.13643 = 1.46098 = V_{01}$ (nearly).

∴ F.R.T. is verified.

NOTES

EXERCISE 8.6

1. Compute Fisher's ideal index number for the following data and show that it satisfies time reversal test and factor reversal test.

Commodity	*1989*		*1990*	
	Price	*Quantity*	*Price*	*Quantity*
A	4	40	5	50
B	8	64	9	80
C	10	70	10	70
D	2	10	4	16

2. Prove using the following data that time reversal test and factor reversal test are satisfied by Fisher's Ideal Formula for Index Numbers :

Commodity	*Base year*		*Current year*	
	Price	*Quantity*	*Price*	*Quantity*
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

3. Verify the 'factor reversal test' by using the following data :

Item	*1993*		*1994*	
	Price per unit	*Expenditure*	*Price per unit*	*Expenditure*
A	5	125	6	180
B	10	50	15	90
C	2	30	3	60
D	3	36	5	75

4. With the help of data given below, compute Fisher's Ideal Index and show that it satisfies the time reversal and factor reversal tests.

Commodity	*1996 (Base year)*		*2002 (Current year)*	
	Price (₹)	*Qty.*	*Price (₹)*	*Qty.*
A	40	30	60	20
B	50	40	60	40
C	70	20	90	20
D	20	30	10	50

V. CONSUMER PRICE INDEX NUMBERS (C.P.I.)

NOTES

8.18. MEANING OF CONSUMER PRICE INDEX

There is no denying the fact that the rise or fall in the prices of commodities affect every family. But, this effect is not same for every family because different families consume different commodities and in different quantities. Car is not found is every house. Milk is used in almost every family but there are very few families who can afford to purchase even more than 5 litres of it, daily.

The index numbers which measures the effect of rise or fall in the prices of various goods and services, consumed by a particular group of people are called **consumer price index numbers** for that particular group of people. The consumer price index numbers help in estimating the average change in the cost of maintaining particular standard of living by a particular class of people.

8.18.1. Procedure

The first step in computing consumer price index number is to decide the category of people for whom the index is to be computed. While fixing the domain of the index, the income and occupation of families must be taken in to consideration. Different families consume different commodities and that too in different quantities. For a particular category of people, it can be expected that their expenditure on different commodities will be almost same.

For computing index, enquiry is made about the expenditure of families on various commodities. The commodities are generally classified in the following heads:

(*a*) Food (*b*) Clothing

(*c*) Fuel and lighting (*d*) House rent

(*e*) Miscellaneous.

After the decision about commodities is taken, the next step is to collect prices of these commodities. The price quotations must be obtained from that market, from where the concerned class of people purchase commodities. The price quotations must be absolutely free from the personal bias of the agent obtaining price quotations. The price quotations must preferably be cross checked in order to eliminate any possibility of personal bias.

All the commodities which are used by a particular class of people cannot be expected to have equal importance. For example, entertainment and house rent can not be given equal weightage. Weights are taken in accordance with the consumption in the base period. Either base period quantities or base period expenditure on different items are generally used,as weights for constructing C.P.I. The base period selected for this purpose must also be normal.

8.18.2. Methods

There are two methods of computing consumer price index numbers.

(*i*) Aggregate expenditure method.

(*ii*) Family budget method..

NOTES

8.18.3. Aggregate Expenditure Method

In this method, generally base period quantities are used as weights.

$$\textbf{Consumer Price Index No.} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

where '0' and '1' suffixes stand for base period and current period respectively.

$\Sigma p_1 q_0$ = sum of the products of the prices of commodities in the current period with their corresponding quantities used in the base period.

$\Sigma p_0 q_0$ = sum of the products of the prices of commodities in the base period with their corresponding quantities used in the base period.

Sometimes, current period quantities are also used for finding consumer price index numbers.

Example 19. *Calculate weighted average of price relative index number from the following data :*

Item	*Unit*	*Base year quantity*	*Base year Price (₹)*	*Current year Price (₹)*
Wheat	*Per Qtl*	*4 Qtl*	*200*	*250*
Sugar	*Per kg*	*50 kg*	*5*	*7*
Milk	*Per litre*	*50 litres*	*5*	*6*
Cloth	*Per meter*	*20 metres.*	*10*	*15*
House	*Per house*	*1*	*50*	*80*

Solution. **Calculation of Cost of Living Index Number**

Item	q_0	p_0	p_1	p_0q_0	p_1q_0
Wheat	4	200	250	800	1000
Sugar	50	5	7	250	350
Milk	50	5	6	250	300
Cloth	20	10	15	200	300
House	1	50	80	50	80
Total				1550	2030

Cost of living index number $= \dfrac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100 = \dfrac{2030}{1550} \times 100 = \mathbf{130.97.}$

8.18.4. Family Budget Method

In this method, the expenditure on different commodities in the base period, are used as weights.

NOTES

$$\textbf{Consumer Price Index No.} = \frac{\Sigma PW}{\Sigma W}$$

where $P = \text{Price relative} = \frac{p_1}{p_0} \times 100.$

p_0, p_1 refers to prices of commodities in the base period and current period respectively.

$W = p_0q_0.$

We have $$\text{C.P.I.} = \frac{\Sigma PW}{\Sigma W} = \frac{\Sigma\left(\frac{p_1}{p_0} \times 100\right)p_0q_0}{p_0q_0} = \frac{\Sigma(p_1 \times 100)q_0}{\Sigma p_0q_0} = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} \times 100.$$

Therefore, the C.P.I. calculated by using both methods would be same. Family budget method is particularly used when the expenditures on various items used in the base period are given on percentage basis.

Example 20. *An enquiry into the budgets of middle class families in a certain city gave the following information :*

Item	*% of total Expenditure*	*Price in 2000 (in ₹)*	*Price in 2002 (in ₹)*
Food	*35%*	*150*	*145*
Fuel	*10%*	*25*	*23*
Clothing	*20%*	*75*	*65*
Rent	*15%*	*30*	*30*
Miscellaneous	*20%*	*40*	*45*

What is the cost of living index number of 2002 as compared with 2000 ?

Solution. Calculation of C.P.I. by Family Budget Method

Item	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$	*W*	*PW*
Food	150	145	96.67	35	3383.45
Fuel	25	23	92	10	920
Clothing	75	65	86.67	20	1733.4
Rent	30	30	100	15	1500
Miscellaneous	40	45	112.5	20	2250
Total				100	9786.85

$$\text{Consumer Price Index No.} = \frac{\Sigma PW}{\Sigma W} = \frac{9786.85}{100} = \mathbf{97.865.}$$

NOTES

Example 21. *From the following data relating to working class consumer price index of a city, calculate index numbers for 1999 and 2001.*

Group	*Food*	*Clothing*	*Fuel*	*House rent*	*Misc.*
Weight	*48*	*18*	*7*	*13*	*14*
Index number 1999	*110*	*120*	*110*	*100*	*110*
Index number 2001	*130*	*125*	*120*	*100*	*135*

The wages were increased by 8% in 2001. Is this increase sufficient ?

Solution. **Calculation of C.P.I. for 1999 and 2001**

Group	*Weight (W)*	*I. No. 1999 (I_1)*	*I. No. 2001 (I_2)*	*WI_1*	*WI_2*
Food	48	110	130	5280	6240
Clothing	18	120	125	2160	2250
Fuel	7	110	120	770	840
House rent	13	100	100	1300	1300
Misc.	14	110	135	1540	1890
Total	100			11050	12520

$$\text{C.P.I. for 1999} = \frac{\Sigma WI_1}{\Sigma W} = \frac{11050}{100} = \mathbf{110.5}$$

$$\text{C.P.I. for 2001} = \frac{\Sigma WI_2}{\Sigma W} = \frac{12520}{100} = \mathbf{125.2}$$

$$\text{\% increase in C.P.I. in 2001} = \frac{125.2 - 110.5}{110.5} \times 100 = \frac{14.7}{110.5} \times 100 = 13.3\%$$

∴ An increase of 8% in wages is insufficient to maintain the standard of living as in 1999.

Example 22. *From the information given below, calculate the cost of living index number for 1975, with 1974 as base year by the Family Budget method:*

Item	*Quantity consumed*	*Unit*	*Price*	
			1974	*1975*
Wheat	*2 Quintals*	*Quintal*	*75*	*125*
Rice	*25 kilograms*	*kg*	*12*	*16*
Sugar	*10 kilograms*	*kg*	*12*	*16*
Ghee	*5 kilograms*	*kg*	*10*	*15*
Clothing	*25 metres*	*metre*	*4.5*	*5*
Fuel	*40 litres*	*litre*	*10*	*12*
Rent	*one house*	*one*	*25*	*40*

NOTES

Solution. **Calculation of C.P.I. by Family Budget Method**

Item	*q*	p_0	p_1	$P = \frac{p_1}{p_0} \times 100$	$W = p_0 q$	*WP*
Wheat	2	75	125	166.67	150	25000.50
Rice	25	12	16	133.33	300	39999.00
Sugar	10	12	16	133.33	120	15999.60
Ghee	5	10	15	150.00	50	7500.00
Clothing	25	4.5	5	111.11	112.5	12499.87
Fuel	40	10	12	120.00	400	48000.00
Rent	1	25	40	160.00	25	4000.00
Total					1157.5	152998.97

Now, consumer price index number $= \frac{\Sigma WP}{\Sigma W} = \frac{152998.97}{1157.5} = \mathbf{132.18.}$

EXERCISE 8.7

1. Construct the cost of living index number from the following data:

Group	*Index for 1992*	*% of Expenditure*
Food	550	46
Clothing	215	10
Fuel and Lighting	220	7
House rent	160	12
Miscellaneous	275	25

2. Calculate the cost of living index for the following data:

Group	*Price in Base year*	*Price in Current year*	*Weight*
Food	39	47	4
Fuel	8	12	1
Clothing	14	18	3
House rent	12	15	2
Miscellaneous	25	30	1

3. Construct a cost of living index number from the following price relatives for the year 1985 and 1986 with 1982 as base giving weightage to the following groups in the proportion of 30, 8, 6, 4 and 2 respectively.

Group	*1982*	*1985*	*1986*
Food	100	114	116
Rent	100	115	125
Clothing	100	108	110
Fuel	100	105	104
Misc.	100	102	104

NOTES

4. From the following data, find the cost of living index number of 1980 on the basis of 1970 by the Family budget method:

Item	*Quantity consumed*	*Unit*	*Prices*	
			1970	*1980*
Wheat	2 quintals	Qtl.	50	75
Rice	25 kilograms	Qtl.	100	120
Sugar	10 kilograms	Qtl.	80	120
Ghee (Desi)	5 kilograms	kg.	10	10
Ghee (Dalda)	5 kilograms	kg.	3	5
Oil	25 kilograms	Qtl.	200	200
Clothing	25 metres	metre	4	5
Fuel	4 quintals	Qtl.	8	10
Rent	One House	House	20	25

5. Taking 1996 as base, construct a consumer index for the year 1998 from the following data:

Item	*Unit*	*Price (1996)*	*Price (1998)*	*Weight*
A.	Kg.	0.50	0.75	10%
B	Litre	0.60	0.75	25%
C	Dozen	2.00	2.40	20%
D	Kg.	0.80	1.00	40%
E	One pair	8.00	10.00	5%

6. An enquiry into the budgets of the middle class families of a certain city revealed that on an average the percentage expenses on the different groups were:

Food 45, Rent 15, Clothing 12, Fuel 8, Miscellaneous 20.

The group index numbers for the current year as compared with a fixed base year were respectively 410, 150, 343, 248 and 285. Calculate the cost of living index number for the current year.

Mr. X was getting 7200 p.m. in the base year and ₹ 12900 p.m in the current year. State how much he ought to have received as extra allowance to maintain his former standard of living.

Answers

1. 377.85	**2.** 126.16	**3.** 112.24, 115.28	**4.** 126.75
5. 126.50	**6.** ₹ 10500.		

EXERCISE 8.8

1. "Index Numbers are economic barometers". Discuss this statement. What precautions will you take while construction an index number ?
2. Distinguish between fixed base and chain base index numbers.
3. State the different uses of index numbers.
4. Explain the use of index numbers. What are the difficulties in the construction of index number?
5. Explain different types of index numbers. Examine the various problems involved in the construction of an index. Discuss in brief the use of an index number.

6. State and explain the fisher's ideal formula for price index number. Show how it satisfies the time reversal and factor reversal test. Why is it used little in practice?
7. Discuss the problems involved in the construction of an index number.

NOTES

8.19. SUMMARY

- The **index numbers** are defined as specialized averages used to measure change in a variable or a group of related variables with respect to time or geographical location or some other characteristic.
- The index numbers are used to measure the change in production, prices, values etc., in related variables over time or geographical location. The barometers are used to study changes in whether conditions, similarly the index numbers are used to study the changes in economic and business activities. That is, why, the index numbers are also called **'Economic Barometers'.**
- Index numbers are used for computing real incomes from money incomes. The wages, dearness allowances etc., are fixed on the basis of real income. The money income is divided by an appropriate consumer's price index number to get real income.
- Index numbers are constructed to compare the changes in related variables over time. Index numbers of industrial production can be used to see the change in the production that has occurred in the current period.
- Index numbers are used to study the changes occurred in the past. This knowledge help in forecasting.
- Index numbers are used to study the changes in prices, industrial production, purchasing powers of money, agricultural production etc., of different countries. With the use of index numbers, the comparative study is also made possible for such variables.
- **Quantity index numbers** are used to show the average change in the quantities of related goods with respect to time. These index numbers are also used to measure the level of production.
- An index number method is said to satisfy **unit test** if it is not changed by a change in the measuring units of some items, under consideration. All methods, except simple aggregative method, satifies this test.
- The index numbers which measures the effect of rise or fall in the prices of various goods and services, consumed by a particular group of people and called **consumer price index numbers** for that particular group of people. The consumer price index numbers help in estimating the average change in the cost of maintaing particular standard of living by a particular class of people.

NOTES

9. REGRESSION ANALYSIS

STRUCTURE

9.1. INTRODUCTION

The literal meaning of the word 'regression' is 'stepping back towards the average'. British biometrician *Sir Francis Galton* (1822–1911) studies the heights of many persons and concluded that the offspring of abnormally tall or short parents tend to *regress* to the average population height. In statistics, *regression analysis* is concerned with the measure of average relationship between variables. Here we shall deal with the derivation of appropriate functional relationships between variables. Regression explains the nature of relationship between variables.

There are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable (or regressed variable or predicted variable or explained variable)*. The variable which influences the value of dependent variable is called *independent variable (or regressor or predictor or explanator)*. Prediction is possible in regression analysis, because here we study the average relationship between related variables.

9.2. USES OF REGRESSION ANALYSIS

The tools of regression analysis are definitely more important and useful than those of correlation analysis. Some of the important uses of regression analysis are as follows:

(*i*) Regression analysis helps in establishing relationship between dependent variable and independent variables. The independent variables may be more than one. Such relationships are very useful in further studies of the variables, under consideration.

NOTES

(*ii*) Regression analysis is very useful for prediction. Once a relation is established between dependent variable and independent variables, the value of dependent variable can be predicted for given values of the independent variables. This is very useful for predicting sale, profit, investment, income, population etc.

(*iii*) Regression analysis is specially used in Economics for estimating demand function, production function, consumption function, supply function etc. A very important branch of Economics, called *Econometrics,* is based on the techniques of regression analysis.

(*iv*) The coefficient of correlation between two variables can be found easily by using the regression lines between the variables.

9.3. TYPES OF REGRESSION

If there are only two variables under consideration, then the regression is called **simple regression.** For example, the study of regression between 'income' and 'expenditure' for a group of family would be termed as simple regression. If there are more than two variables under consideration then the regression is called **multiple regression.** In this text, we shall restrict ourselves to the study of only simple regression. The regression is called **partial regression** if there are more than two variables under consideration and relation between only two variables is established after excluding the effect of other variables. The simple regression is called **linear regression** if the point on the scatter diagram of variables lies almost along a line otherwise it is termed as **non-linear regression** or **curvilinear regression.**

9.4. REGRESSION LINES

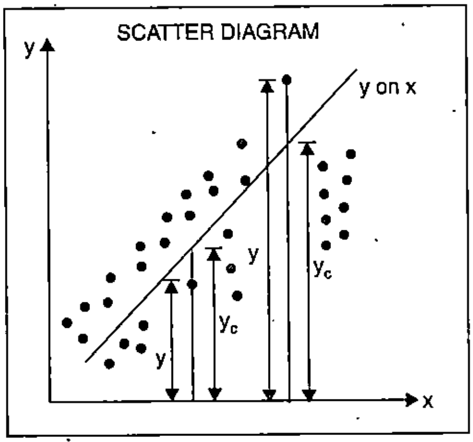
Let the variables under consideration be denoted by 'x' and 'y'. The line used to estimate the value of y for a given value of x is called the *regression line* of y *on* x. Similarly, the line used to estimate the value of x for a given value of y is called the *regression line of* x *on* y. In regression line of y on x (x on y), the variable y is considered as the dependent (independent) variable whereas x is considered as the independent (dependent) variable. The position of regression lines depends upon the given pairs of value of the variables. Regression lines are also known as *estimating lines.* We shall see that in case of perfect correlation between the variables, the regression lines will be coincident. The angle between the regression lines will increases for 0° to 90° as the correlation coefficient numerically decreases from 1 to 0. If for a particular pair of variables, $r = 0$, then the regression lines will be perpendicular to each other. The regression lines will be determined by using the *principle of least squares.*

9.5. REGRESSION EQUATIONS

NOTES

We have already noted that for two variables x and y, there can be two regression lines. If the intention is to depict the change in y for a given change in x, then the regression line of y on x is to be used. Similar argument also works for regression line of x on y.

(*i*) **Regression equation of y on x.** The regression equation of y on x is estimated by using the 'principle of least squares'. This principle will ensure that the sum of the squares of the *vertical* deviations of actual values of y from estimated values for all possible values of x is minimum.



Mathematically, $\Sigma(y - y_c)^2$ is least, where y and y_c are the corresponding actual and computed values of y for a particular value of x.

Let n pairs of values $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ of two variables x and y be given.

Let the regression equation of y on x be $y = a + bx$. ...(1)

By using derivatives, it can be proved that the constants a and b are found by using the *normal equations :*

$$\Sigma y = an + bSx \qquad ...(2)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2. \qquad ...(3)$$

Dividing (2) by n, we get

$$\frac{\Sigma y}{n} = a + b\frac{\Sigma x}{n}.$$

$$\Rightarrow \qquad \overline{y} = a + b\,\overline{x} \qquad ...(4)$$

Subtracting (4) from (1), we get

$$y - \overline{y} = b(x - \overline{x}) \qquad ...(5)$$

Multiplying (2) by Σx and (3) by n and subtracting, we get

$$(\Sigma x)(\Sigma y) - n\Sigma xy = b(\Sigma x)^2 - bn\Sigma x^2$$

$$\Rightarrow \qquad n\Sigma xy - (\Sigma x)(\Sigma y) = b(n\Sigma x^2 - (\Sigma x)^2)$$

$$\therefore \qquad b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}.$$

The constant b is denoted by b_{yx} and is called **regression coefficient** of y on x.

NOTES

$\therefore$ (5) $\Rightarrow$ $y-\bar{y}=b_{yx}(x-\bar{x})$, where $b_{yx}=\dfrac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma x^2-(\Sigma x)^2}$.

Remark. $b_{yx}=\dfrac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma x^2-(\Sigma x)^2}$ implies

$$b_{yx}=\frac{n\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2-(\Sigma x)^2}\sqrt{n\Sigma y^2-(\Sigma y)^2}}\times\frac{\dfrac{\sqrt{n\Sigma y^2-(\Sigma y)^2}}{n}}{\dfrac{\sqrt{n\Sigma x^2-(\Sigma x)^2}}{n}}=r\times\frac{\sqrt{\dfrac{\Sigma y^2}{n}-\left(\dfrac{\Sigma y}{n}\right)^2}}{\sqrt{\dfrac{\Sigma x^2}{n}-\left(\dfrac{\Sigma x}{n}\right)^2}}=r.\frac{\sigma_y}{\sigma_x}.$$

$\therefore$ $\mathbf{b_{yx}=r\dfrac{\sigma_y}{\sigma_x}}$.

Thus we see that the regression equation of y on x is $\mathbf{y-\bar{y}=b_{yx}(x-\bar{x})}$, where $\mathbf{\bar{x}=\dfrac{\Sigma x}{n}}$, $\mathbf{\bar{y}=\dfrac{\Sigma y}{n}}$, $\mathbf{b_{yx}=\dfrac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma x^2-(\Sigma x)^2}}$, which is also equal to $\mathbf{r\dfrac{\sigma_y}{\sigma_x}}$.

Example 1. *Find b_{yx} from the following data :*

$\{(x, y)\}=\{(5, 2), (7, 4), (8, 3), (4, 2), (6, 4)\}$.

Solution. **Calculation of $\mathbf{b_{yx}}$**

S. No.	x	y	xy	x^2
1	5	2	10	25
2	7	4	28	49
3	8	3	24	64
4	4	2	8	16
5	6	4	24	36
$n=5$	$\Sigma x=30$	$\Sigma y=15$	$\Sigma xy=94$	$\Sigma x^2=190$

$$b_{yx}=\frac{n\Sigma xy-(\Sigma x)(\Sigma y)}{n\Sigma x^2-(\Sigma x)^2}=\frac{5(94)-(30)(15)}{5(190)-(30)^2}=\frac{20}{50}=\mathbf{0.4}.$$

Example 2. *Find the most likely price in Mumbai corresponding to price of ₹ 75 at Calcutta from the following data :*

	Calcutta	*Mumbai*
Average price	*₹ 65*	*₹ 68*
Standard deviation	*₹ 2.5*	*₹ 3.5*

Coefficient of correlation between two prices = 0.78.

Solution. Let the 'price in Calcutta' and 'price in Mumbai' be denoted by x and y respectively.

We have $\bar{x}$ = ₹ 65, $\bar{y}$ = ₹ 68,

σ_x = ₹ 2.5, σ_y = ₹ 3.5, $r=0.78$.

The regression line of y on x is $y-\bar{y}=b_{yx}(x-\bar{x})$.

$\Rightarrow$ $y-\bar{y}=r\dfrac{\sigma_y}{\sigma_x}(x-\bar{x})$ $\Rightarrow$ $y-68=0.78\times\dfrac{3.5}{2.5}(x-65)$

$\Rightarrow$ $y-68=1.092(x-65)$ $\Rightarrow$ $y=1.092x+68-(1.092\times 65)$

$$\Rightarrow \qquad y = 1.092x - 2.98.$$

When x is ₹ 75, the expected value of

$$y = 1.092\ (75) - 2.98 = ₹\ 78.92.$$

∴ Price at Mumbai = ₹ **78.92.**

NOTES

Example 3. *For the following data, find the regression line of y on x :*

x	1	2	3	4	5	8	10
y	9	8	10	12	14	16	15

Solution. **Regression line of y on x**

S. No.	x	y	xy	x^2
1	1	9	9	1
2	2	8	16	4
3	3	10	30	9
4	4	12	48	16
5	5	14	70	25
6	8	16	128	64
7	10	15	150	100
$n = 7$	$\Sigma x = 33$	$\Sigma y = 84$	$\Sigma xy = 451$	$\Sigma x^2 = 219$

The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{33}{7} = 4.714, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{84}{7} = 12$$

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(451) - (33)(84)}{7(219) - (33)^2} = \frac{385}{444} = \mathbf{0.867.}$$

∴ The equation of regression line of y on x is

$$y - 12 = 0.867\ (x - 4.714)$$

or $$y = 0.867x + 12 - (0.867)(4.714)$$

or $$\mathbf{y = 0.867x - 7.913.}$$

(*ii*) **Regression equation of x on y.** The regression equation of x on y is also estimated by using the 'principle of least squares'. This principle will ensure that the sum of the squares of the *horizontal* deviations of actual values of x from estimated values for all possible values of y is minimum. Mathematically, $\Sigma(x - \bar{x}_c)^2$ is least, where x and x_c are the corresponding actual and computed values of x for a particular value of y.

Let n pairs of values $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ of two variables x and y be given.

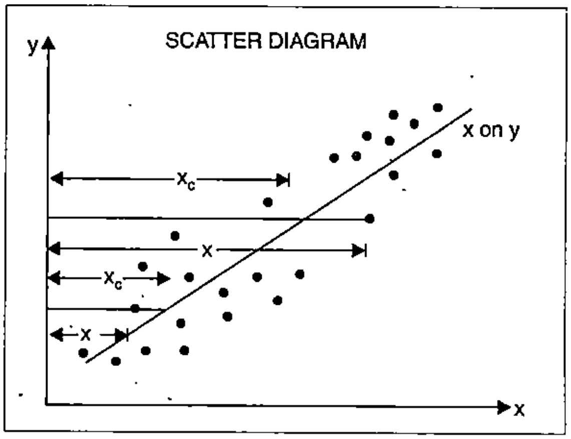
Let the regression equation of x on y be $x = a + by$ ...(1)

By using derivatives, it can be proved that the constants a and b are found by using the *normal equations*:

$$\Sigma x = ax + b\Sigma y \qquad ...(2)$$

and $$\Sigma xy = a\Sigma y + b\Sigma y^2. \qquad ...(3)$$

NOTES



Dividing (2) by n, we get

$$\frac{\Sigma x}{n} = a + b\frac{\Sigma y}{n}$$

$$\Rightarrow \qquad \bar{x} = a + b\bar{y} \qquad \text{...(4)}$$

Subtracting (4) from (1), we get

$$x - \bar{x} = b(y - \bar{y}) \qquad \text{...(5)}$$

Multiplying (2) by Σy and (3) by n and subtracting, we get

$$(\Sigma x)(\Sigma y) - n\Sigma xy = b(\Sigma y)^2 - bn\Sigma y^2$$

$$\Rightarrow \qquad n\Sigma xy - (\Sigma x)(\Sigma y) = b(n\Sigma y^2 - \Sigma y)^2$$

$$\therefore \qquad b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}.$$

The constant b is denoted by b_{xy} and is called **regression coefficient** of x on y.

$$\therefore \quad (5) \quad \Rightarrow \qquad x - \bar{x} = b_{xy}(y - \bar{y}), \text{ where } \qquad b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$$

Remark. $b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}$ implies

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \times \frac{\dfrac{\sqrt{n\Sigma x^2 - (\Sigma x)^2}}{n}}{\dfrac{\sqrt{n\Sigma y^2 - (\Sigma y)^2}}{n}}$$

$$= r \times \frac{\sqrt{\dfrac{\Sigma x^2}{n} - \left(\dfrac{\Sigma x}{n}\right)^2}}{\sqrt{\dfrac{\Sigma y^2}{n} - \left(\dfrac{\Sigma y}{n}\right)^2}} = r\frac{\sigma_x}{\sigma_y}.$$

$$\therefore \qquad \mathbf{b_{xy} = r\frac{\sigma_y}{\sigma_x}}.$$

Thus we see that the regression equation of x on y is $\mathbf{x - \bar{x} = b_{xy}(y - \bar{y})}$, where $\mathbf{\bar{x} = \frac{\Sigma x}{n}}$, $\mathbf{\bar{y} = \frac{\Sigma y}{n}}$, $\mathbf{b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2}}$, which is also equal to $\mathbf{r\frac{\sigma_x}{\sigma_y}}$.

Example 4. *Find the regression coefficient b_{xy} between x and y for the following data:*

$$\Sigma x = 30,\ \Sigma y = 42,\ \Sigma xy = 199,\ \Sigma x^2 = 184,\ \Sigma y^2 = 318,\ n = 6.$$

Solution. $$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{6(199) - (30)(42)}{6(318) - (42)^2} = \frac{-66}{144} = -0.4583.$$

Example 5. *Find the two regression equations from the following data and estimate the value of X, if Y is 6 :*

x	*1*	*2*	*3*	*4*	*5*
y	*2*	*5*	*3*	*8*	*7*

Solution. **Regression Equations**

S. No.	x	y	xy	x^2	y^2
1	1	2	2	1	4
2	2	5	10	4	25
3	3	3	9	9	9
4	4	8	32	16	64
5	5	7	35	25	49
$n = 5$	$\Sigma x = 15$	$\Sigma y = 25$	$\Sigma xy = 88$	$\Sigma x^2 = 55$	$\Sigma y^2 = 151$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{15}{5} = 3,\quad \overline{Y} = \frac{\Sigma Y}{n} = \frac{25}{5} = 5$$

The regression equation of Y on X is $Y - \overline{Y} = b_{YX}(X - \overline{X})$.

$$b_{YX} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2} = \frac{5(88) - (15)(25)}{5(55) - (15)^2} = \frac{65}{50} = 1.3$$

∴ The equation is

$$Y - 5 = 1.3\ (X - 3)$$

or $$Y = 1.3X + 5 - 3.9 \quad \text{or} \quad \mathbf{Y = 1.3X + 1.1.}$$

The regression equation of X on Y is $X - \overline{X} = b_{XY}(Y - \overline{Y})$.

$$b_{XY} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{5(88) - (15)(25)}{5(151) - (25)^2} = \frac{65}{130} = 0.5$$

∴ The equation is

$$X - 3 = 0.5\ (Y - 5)$$

or $$X = 0.5\,Y + 3 - 2.5 \quad \text{or} \quad \mathbf{X = 0.5Y + 0.5.}$$

For estimating the value of X, we shall use the regression equation of X on Y.

The regression equation of X on Y is $X = 0.5Y + 0.5$.

∴ When Y = 6, we have

$$X = 0.5 \times 6 + 0.5 = \mathbf{3.5}.$$

Example 6. *The coefficient of correlation between ages of husbands and wives in a community was found to be + 0.8, the average of husband's age is 25 years and that of wive's age was 22 years. Their standard deviations were 4 years and 5 years respectively. Find with the help of regression equations:*

(i) *The expected age of husband when wife's age is 12 years.*

(ii) *The expected age of wife when husband's age is 20 years.*

Solution. Let x and y denote the variables 'age of husband' and 'age of wife' respectively.

$\therefore$ We have

$$r = 0.8, \quad \bar{x} = 25 \text{ years}, \ \bar{y} = 22 \text{ years}, \ \sigma_x = 4 \text{ years and } \ \sigma_y = 5 \text{ years.}$$

(*i*) We are to find the expected age of husband (x) for a given age of wife (y).

$\therefore$ We use regression equation of x on y which is given by

$$x - \bar{x} = b_{xy}(y - \bar{y}).$$

$$\Rightarrow \quad x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \ \Rightarrow \ x - 25 = (0.8)\frac{4}{5}(y - 22)$$

$$\Rightarrow \quad x - 25 = 0.64(y - 22) \quad \Rightarrow \quad x = 0.64y + 25 - (0.64)\,22$$

$$\Rightarrow \quad x = 0.64y + 10.92.$$

When $y = 12$ years, the expected value of

$$x = 0.64\,(12) + 10.92 = \mathbf{18.6 \text{ years.}}$$

(*ii*) We are to find the expected age of wife (y) for a given age of husband (x).

$\therefore$ We use regression equation of y on x which is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}).$$

$$\Rightarrow \quad y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \ \Rightarrow \quad y - 22 = (0.8)\frac{5}{4}(x - 25)$$

$$\Rightarrow \quad y - 22 = x - 25. \quad \Rightarrow \quad y = x - 3.$$

When $x = 20$ years, the expected value of $y = 20 - 3 =$ **17 years.**

Example 7. *For the following data, find the regression line of x on y. Also show the regression line on a graph paper:*

x	*1*	*2*	*3*	*4*	*5*	*6*	*7*
y	*9*	*8*	*10*	*12*	*11*	*13*	*14*

Solution. **Regression line of x on y**

S. No.	*x*	*y*	*xy*	*y*^{2}
1	1	9	9	81
2	2	8	16	64
3	3	10	30	100
4	4	12	48	144
5	5	11	55	121
6	6	13	78	169
7	7	14	98	196
$n = 7$	$\Sigma x = 28$	$\Sigma y = 77$	$\Sigma xy = 334$	$\Sigma y^2 = 875$

The regression line of x on y is $\quad x - \bar{x} = b_{xy}(y - \bar{y})$.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{28}{7} = 4, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{77}{7} = 11.$$

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{7(334) - (28)(77)}{7(875) - (77)^2} = \frac{182}{196} = 0.929.$$

∴ The equation of regression line of x on y is

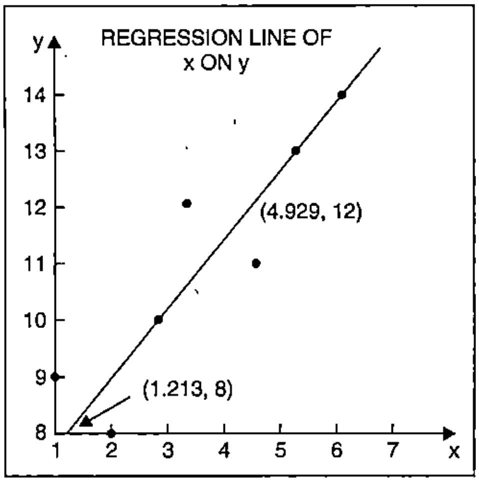
$$x - 4 = 0.929\,(y - 11) \quad \text{or} \quad x = 0.929y + 4 - 11\,(0.929)$$

or $$\mathbf{x = 0.929y - 6.219.}$$

To draw this line on the graph paper, we take two points on it.

$$y = 8 \quad \Rightarrow \quad x = 0.929(8) - 6.219 = 1.213$$

$$y = 12 \quad \Rightarrow \quad x = 0.929(12) - 6.219 = 4.929.$$



∴ The points (1.213, 8) and (4.929, 12) are on the regression line of x on y. The line joining these points is the required regression line of x on y.

Example 8. *Show that regression coefficients are independent of the change of origin but not of scale.*

Solution. Let x and y be any two variables. Let A, B, h and k be any constant.

Let $$u = \frac{x - A}{h} \quad \text{and} \quad v = \frac{y - B}{k}.$$

∴ u and v are variables obtained by changing origin and scale of given variables x and y respectively.

∴ $$x = A + hu \quad \text{and} \quad y = B + kv$$

Summing both sides and dividing by the number of values, we get

$$\bar{x} = A + h\bar{u} \quad \text{and} \quad \bar{y} = B + k\bar{v}.$$

∴ $$x - \bar{x} = (A + hu) - (A + h\bar{u}) = h(u - \bar{u})$$

and $$y - \bar{y} = (B + kv) - (B + k\bar{v}) = k(v - \bar{v}).$$

Now $$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2}\sqrt{\Sigma(y - \bar{y})^2}} \cdot \frac{\sqrt{\dfrac{\Sigma(y - \bar{y})^2}{n}}}{\sqrt{\dfrac{\Sigma(x - \bar{x})^2}{n}}}$$

$$= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma[h(u - \bar{u}) \times k(v - \bar{v})]}{\Sigma[h(u - \bar{u})]^2}$$

$$= \frac{hk}{h^2} \cdot \frac{\Sigma(u - \bar{u})(v - \bar{v})}{\Sigma(u - \bar{u})^2} = \frac{k}{h} \cdot b_{vu}.$$

NOTES

Also
$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \cdot \frac{\sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n}}}{\sqrt{\dfrac{\Sigma(y-\bar{y})^2}{n}}}$$
$$= \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{\Sigma[h(u-\bar{u}) \times k(v-\bar{v})]}{\Sigma[k(v-\bar{v})]^2}$$
$$= \frac{hk}{k^2} \cdot \frac{\Sigma(u-\bar{u})(v-\bar{v})}{\Sigma(v-\bar{v})^2} = \frac{h}{k} . b_{uv} .$$

∴ Regression coefficients are independent of change of origin but not of scale.

EXERCISE 9.1

1. Find b_{yx} from the following data:
$\Sigma x = 30, \Sigma y = 42, \Sigma xy = 199, \Sigma x^2 = 184, \Sigma y^2 = 318, n = 6.$
2. Find b_{yx} from the following data:
$\Sigma x = 24, \Sigma xy = 306, \Sigma x^2 = 164, \Sigma y = 44, \Sigma y^2 = 574, n = 4.$
3. Find the regression coefficient b_{xy} between x and y for the following data:
$\Sigma x = 55, \Sigma y = 88, \Sigma x^2 = 385, \Sigma y^2 = 1114, \Sigma xy = 586, n = 10.$
4. Find the regression coefficient b_{xy} between x and y for the following data:
$\Sigma x = 24, \Sigma y = 44, \Sigma xy = 306, \Sigma x^2 = 164, \Sigma y^2 = 574, n = 4.$
5. Find b_{yx} from the following data:

x	1	2	3	4	5
y	6	8	7	6	8

6. Find the regression line of y on x, where:
$\Sigma x = 55, \Sigma y = 88, \Sigma x^2 = 385, \Sigma y^2 = 1114, \Sigma xy = 586, n = 10.$
7. x and y are correlated variables. Eight observations of (x, y) have the following results:
$\Sigma x = 55, \Sigma y = 55, \Sigma xy = 350, \Sigma x^2 = 385.$
Predict the value of y when the value of x is 8.
8. For observations of pairs (x, y) of variables x and y, the following results are obtained:
$\Sigma x = 125, \Sigma y = 100, \Sigma x^2 = 1650, \Sigma y^2 = 1500, \Sigma xy = 50$ and $n = 25.$
Find the equation of the line of regression of x and y. Estimate the value of x if $y = 5$.
9. Find the value of X when Y = 60, and the value of Y when X = 50 from the following information:

	Variable X	*Variable Y*
Mean	24	140
S.D.	16	48

Also, $r = 0.6$.
10. Given the following data, find what will be : (*a*) the height of a policeman whose weight is 200 pounds, (*b*) the weight of a policeman who is 5 ft tall.
Average height = 68 inches, average weight = 150 pounds, coefficient of correlation between height and weight = 0.6, S.D. of height = 2.5 inches, S.D. of weight = 20 pounds.

11. Find the equations of regression lines for the following data:

x	1	2	3	4	5
y	7	8	10	12	13

12. Find the equations of two regression lines from the following data:

x	1	2	3	4	5
y	7	6	5	4	3

Hence find the estimated value of y for $x = 3.5$ from the appropriate line of regression.

13. The data of sales and promotion expenditure on a product for 10 years are given below:

Sales (₹ lakh)	8	10	9	12	10	11	12	13	14	15
Promotion expenditure (₹ thousand)	2	2	3	4	5	5	5	6	7	8

Use two-variable regression model to estimate the promotion expenditure for a given sale of ₹ 20 lakh. Forecast the sales when the company wants to spends ₹ 10 thousand on promotion.

14. A computer while calculating the correlation coefficient between two variables x and y obtained the following constants:

$$n = 25,\ \Sigma x = 125,\ \Sigma y = 100,\ \Sigma x^2 = 650,\ \Sigma y^2 = 460,\ \Sigma xy = 508$$

It was however, later discovered at the time of checking that it had copied down two pairs of observations as: while the correct value were : . After making the necessary corrections, find the:

x	y
6	14
8	6

x	y
8	12
6	8

(*i*) regression coefficients (*ii*) regression equations and

(*iii*) correlation coefficient.

Answers

1. – 0.3235 **2.** 2.1 **3.** 0.3004 **4.** 0.4667

5. 0.2 **6.** $y = 1.2364x + 1.9998$ **7.** 2.2727

8. $x = -0.4091y + 6.6364$, 4.5909

9. 8, 186.8

10. (*a*) 71.75 inches (*b*) 111.6 pounds **11.** $y = 1.6x + 5.2$, $x = 0.615y - 3.15$

12. Regression line of y on $x : y = -x + 8$;

Regression line of x on $y : x = -y + 8$; $y = 4.5$ when $x = 3.5$

13. $y = 0.815x - 4.591$, when $x = 20$, $y = 11.709$; $x = 1.003y + 6.686$, when $y = 10$, $x = 16.716$.

14. (*i*) $b_{yx} = 0.8$, $b_{xy} = 0.555$

(*ii*) Regression equations of y on $x : y = 0.8x$

(*iii*) Regression equation of x on $y : x = 0.555y + 2.78$

(*iv*) $r = 0.667$.

9.6. STEP DEVIATION METHOD

When the values of x and y are numerically high, the step deviation method is used.

Deviations of values of variables x and y are calculated from some chosen arbitrary numbers, called A and B. Let h be a positive common factor of all deviations $(x - \text{A})$ of items in the x-series. Similarly let k be a positive factor of all deviations $(y - \text{B})$ of items in the y-series. The step deviations are :

$$u = \frac{x - \text{A}}{h}, \quad v = \frac{x - \text{B}}{k}.$$

In practical problems, if we do not bother to divide the deviations by common factors, then these deviations would be thought of as step deviations of items of given series with '1' as the common factor for both series.

The equation of regression line of y on x in terms of step deviations is

$$y - \bar{y} = b_{yx}(x - \bar{x}),$$

where
$$\bar{x} = \text{A} + \left(\frac{\Sigma u}{n}\right)h, \quad \bar{y} = \text{B} + \left(\frac{\Sigma v}{n}\right)k$$

and
$$b_{yx} = b_{vu} \cdot \frac{k}{h} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \cdot \frac{k}{h}.$$

The equation of regression line of x on y in terms of step deviations is

$$x - \bar{x} = b_{xy}(y - \bar{y}),$$

where
$$\bar{x} = \text{A} + \left(\frac{\Sigma u}{n}\right)h, \quad \bar{y} = \text{B} + \left(\frac{\Sigma v}{n}\right)k$$

and
$$b_{xy} = b_{uv} \cdot \frac{h}{k} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} \cdot \frac{h}{k}.$$

Remark. In particular if $u = x - A$ and $v = y - \text{B}$ *i.e.*, when $h = 1$, $k = 1$, we have

$$\bar{x} = \text{A} + \frac{\Sigma u}{n}, \quad \bar{y} = \text{B} + \frac{\Sigma v}{n},$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \quad \text{and} \quad b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2}.$$

Example 9. *For a bivariate data, you are given the following information :*

$$\Sigma(x - 44) = -5,\ \Sigma(y - 26) = -6,\ \Sigma(x - 44)^2 = 255,$$
$$\Sigma(y - 26)^2 = 704,\ \Sigma(x - 44)(y - 26) = -306.$$

Number of pairs of observations = 12.

Find the regression equations.

Solution. Let $u = x - 44$ and $v = y - 26$.

$\therefore$ $\Sigma u = -5$, $\Sigma v = -6$, $\Sigma u^2 = 255$, $\Sigma v^2 = 704$, $\Sigma uv = -306$, $n = 12$.

$$\bar{x} = 44 + \frac{\Sigma u}{n} = 44 + \frac{(-5)}{12} = 43.58$$

$$\bar{y} = 26 + \frac{\Sigma v}{n} = 26 + \frac{(-6)}{12} = 25.5$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{12(-306) - (-5)(-6)}{12(255) - (-5)^2} = \frac{-3702}{3035} = -1.22$$

$$b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} = \frac{12(-306) - (-5)(-6)}{12(704) - (-6)^2} = \frac{-3702}{8412} = -0.44.$$

Regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$\Rightarrow \quad y - 25.5 = -1.22\ (x - 43.58)$

$\Rightarrow \quad y = -1.22x + 25.5 + (1.22)\ (43.58)$

$\Rightarrow \quad \mathbf{y = -1.22x + 78.67.}$

Regression equation of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

$\Rightarrow \quad x - 43.58 = -0.44(y - 25.5)$

$\Rightarrow \quad x = -0.44y + 43.58 + (0.44)(25.5)$

$\Rightarrow \quad \mathbf{x = -0.44y + 54.8.}$

Example 10. *Obtain the regression equations of 'x on y' and 'y on x' taking origin as 2 and 200 for x and y respectively :*

x	*1*	*2*	*3*	*4*	*5*
y	*166*	*184*	*142*	*180*	*338*

Solution. **Computation of Regression Equations**

S. No.	*x*	*y*	*u = x – A* *A = 2*	*v = y – B* *B = 200*	*uv*	*u²*	*v²*
1	1	166	– 1	– 34	34	1	1156
2	2	184	0	– 16	0	0	256
3	3	142	1	– 58	– 58	1	3364
4	4	180	2	– 20	– 40	4	400
5	5	338	3	138	414	9	19044
$n = 5$			$\Sigma u = 5$	$\Sigma v = 10$	$\Sigma uv = 350$	$\Sigma u^2 = 15$	$\Sigma v^2 = 24220$

Regression equation of 'x on y'

The regression equation of x on y is $x - \bar{x} = b_{xy}\ (y - \bar{y})$.

We have $\bar{x} = \text{A} + \frac{\Sigma u}{n} = 2 + \frac{5}{5} = 3$

$\bar{y} = \text{B} + \frac{\Sigma v}{n} = 200 + \frac{10}{5} = 202$

$$b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} = \frac{5(350) - (5)(10)}{5(24220) - (10)^2} = \frac{1700}{121000} = 0.014.$$

$\therefore$ The required equation is $x - 3 = 0.014\ (y - 202)$

or $\quad x = 0.014y + 3 - (0.014)(202)$

or $\quad \mathbf{x = 0.014y + 0.172.}$

Regression equation of 'y on x'

The regression equation of y on x is $y - \bar{y} = b_{yx}\ (x - \bar{x})$.

$\bar{x} = 3,\ \bar{y} = 202$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{5(350) - (5)(10)}{5(15) - (5)^2} = \frac{1700}{50} = 34.$$

$\therefore$ The regression equation is $y - 202 = 34(x - 3)$

or $\quad y = 34x + 202 - 34(3) \quad$ or $\quad \mathbf{y = 34x + 100.}$

NOTES

Example 11. *Find (i) r and (ii) regression equations for the following data :*

x	*75*	*89*	*97*	*69*	*59*	*79*	*68*	*61*
y	*125*	*137*	*156*	*112*	*107*	*136*	*123*	*108*

Solution. **Computation of 'r' and Regression Equations**

S. No.	*x*	*y*	$u = x - A$ $A = 100$	$v = y - B$ $B = 100$	*uv*	u^2	v^2
1	75	125	−25	25	−625	625	625
2	89	137	−11	37	−407	121	1369
3	97	156	−3	56	−168	9	3136
4	69	112	−31	12	−372	961	144
5	59	107	−41	7	−287	1681	49
6	79	136	−21	36	−756	441	1296
7	68	123	−32	23	−736	1024	529
8	61	108	−39	8	−312	1521	64
$n = 8$			$\Sigma u = -203$	$\Sigma v = 204$	$\Sigma uv =$ -3663	$\Sigma u^2 =$ 6383	$\Sigma v^2 =$ 7212

(*i*)

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

$$= \frac{8(-3663) - (-203)(204)}{\sqrt{8(6383) - (-203)^2}\sqrt{8(7212) - (204)^2}} = \frac{12108}{\sqrt{9855}\sqrt{16080}} = \mathbf{0.9619.}$$

(*ii*) **Regression equation of 'y on x'**

The regression equation of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$\bar{x} = \text{A} + \frac{\Sigma u}{n} = 100 + \frac{(-203)}{8} = 74.625$$

$$\bar{y} = \text{B} + \frac{\Sigma v}{n} = 100 + \frac{204}{8} = 125.5$$

$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{8(-3663) - (-203)(204)}{8(6383) - (-203)^2} = \frac{12108}{9855} = 1.2286$$

∴ The required equation is $y - 125.5 = 1.2286\ (x - 74.625)$

or $y = 1.2286x + 125.5 - (1.2286)(74.625)$

or $\mathbf{y = 1.2286x + 33.8157.}$

Regression equation of x on y

The regression equation of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

$$\bar{x} = 74.625,\ \bar{y} = 125.5$$

$$b_{xy} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma v^2 - (\Sigma v)^2} = \frac{8(-3663) - (-203)(204)}{8(7212) - (-204)^2} = \frac{12108}{16080} = 0.753$$

∴ The required equation is $x - 74.625 = 0.753\ (y - 125.5)$

or $x = 0.753y + 74.625 - (0.753)(125.5)$

or $\mathbf{x = 0.753y - 19.8765.}$

Example 12. *A panel of judges A and B graded seven debtors and independently awarded the following marks :*

Debtors	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Marks by A	*40*	*34*	*28*	*30*	*44*	*38*	*31*
Marks by B	*32*	*39*	*26*	*30*	*38*	*34*	*28*

The eighth debtor was awarded 36 marks by judge A while judge B was not present. If judge B was also present, how many marks would you expect him to award to the eighth debtor assuming that the same degree of relationship exists in their judgement.

Solution. Let the variables 'marks by A' and 'marks by B' be denoted by 'x' and 'y' respectively. We shall estimate the marks given by judge B to the eighth debtor by using the fact that he has been awarded 36 marks by judge A. In other words, we shall estimate the value of y, when $x = 36$. For this, we shall need the regression equation of y on x.

Computation of Regression Line of y on x

Debtor	x	y	$u = x - A$ $A = 35$	$v = y - B$ $B = 35$	uv	u^2
1	40	32	5	−3	−15	25
2	34	39	−1	4	−4	1
3	28	26	−7	−9	63	49
4	30	30	−5	−5	25	25
5	44	38	9	3	27	81
6	38	34	3	−1	−3	9
7	31	28	−4	−7	28	16
$n = 7$			$\Sigma u = 0$	$\Sigma v = -18$	$\Sigma uv = 121$	$\Sigma u^2 = 206$

The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

We have
$$\bar{x} = A + \frac{\Sigma u}{n} = 35 + \frac{0}{7} = 35$$
$$\bar{y} = B + \frac{\Sigma v}{n} = 35 + \frac{(-18)}{7} = 32.429$$
$$b_{yx} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} = \frac{7(121) - 0(-18)}{7(206) - (0)^2} = \frac{121}{206} = 0.5874$$

∴ The required equation is $y - 32.429 = 0.5874(x - 35)$

or $$y = 0.5874x + 32.429 - (0.5874)35$$

or $$\mathbf{y = 0.5874x + 11.87.}$$

∴ When $x = 36$, the estimated value of
$$y = (0.5874)36 + 11.87 = 33.0164 \simeq 33.$$

∴ The judge B would have awarded **33 marks** to the eighth debtor.

EXERCISE 9.2

1. For a bivariate data, you are given the following information:

$$n = 10,\ \Sigma x = 320,\ \Sigma y = 380,\ \Sigma(x-32)^2 = 140,$$
$$\Sigma(y-38)^2 = 398,\ \Sigma(x-32)(y-38) = -93.$$

Find the:

(*i*) regression coefficients (*ii*) regression equations

(*iii*) correlation coefficient.

(**Hint.** Let $u = x - 32$ and $v = y - 38$.

$$\therefore \quad \Sigma u^2 = 140,\quad \Sigma v^2 = 398,\quad \Sigma uv = -93$$
$$\Sigma u = \Sigma(x-32) = \Sigma x - 32n = 320 - 32 \times 10 = 0$$
$$\Sigma v = \Sigma(y-38) = \Sigma y - 38n = 380 - 38 \times 10 = 0$$
$$\bar{x} = 32 + \frac{\Sigma u}{n} = 32 + \frac{0}{12} = 32$$
$$\bar{y} = 38 + \frac{\Sigma v}{n} = 38 + \frac{0}{12} = 38.$$

2. Find the regression equations for the following data:

Age of husband, x	36	23	27	28	28	29	30	31	33	35
Age of wife, y	29	18	20	22	27	21	29	27	29	28

3. From the following data, obtain the two regression equations:

Sales	91	97	108	121	67	124	51
Purchase	71	75	69	97	70	91	39

4. The following data gives the experience of machine operators and their performance rating as given by the number of good pieces turned out per 100 pieces:

Operator	1	2	3	4	5	6	7	8
Experience (in years)	16	12	18	4	3	10	5	12
Performance rating	87	88	89	68	78	80	75	83

Calculate the regression line of performance rating on experience and estimate the probable performance rating if an operator has seven years experience.

5. The following table gives the aptitude test scores and productivity indices of 10 workers selected at random:

Aptitude score, x	60	62	65	70	72	48	53	73	65	82
Productivity index, y	68	60	62	80	85	40	52	62	60	81

Calculate the two regression equations and estimate the productivity index of a worker whose test score is 78 and the test score of a worker whose productivity index is 88.

Answers

1. (*i*) $b_{yx} = -0.6643$, $b_{xy} = -0.2337$

(*ii*) Regression equation of y on x : $y = -0.6643x + 59.2576$

Regression equation of x on y : $x = -0.2337y + 40.881$

(*iii*) $r = -0.394$

2. Regression equation of y on x : $y = 0.8913x - 1.739$

Regression equation of x on y : $x = 0.75y + 11.25$

3. If x and y represent the variables 'sales' and 'purchases' respectively, then
   Regression equation of y on x : $y = 0.607x + 15.998$
   Regression equation of x on y : $x = 1.286y + 0.081$
4. $y = 1.133x + 69.67$, $y = 78$ when $x = 7$
5. Regression equation of x on y : $x = 0.596y + 26.26$
   Regression equation of y on x : $y = 1.168x - 10.92$
   $y = 80.184$ when $x = 78$, $x = 78.708$ when $y = 88$

9.7. REGRESSION LINES FOR GROUPED DATA

In case of grouped data if either x or y or both variables represent classes, then their respective mid-points are taken as their representatives.

In this case, if $u = \frac{x - A}{h}$, $v = \frac{y - B}{k}$,

then the regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$,

where
$$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h,$$

$$\bar{y} = B + \left(\frac{\Sigma fv}{N}\right)k$$

and
$$b_{yx} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fu^2 - (\Sigma fu)^2} \cdot \frac{k}{h}$$

The regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}),$$

where
$$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h,$$

$$\bar{y} = B + \left(\frac{\Sigma fv}{N}\right)k$$

and
$$b_{xy} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fv^2 - (\Sigma fv)^2} \cdot \frac{h}{k}.$$

Example 13. *Calculate regression lines for the following data :*

		x					*Total*
		18	*19*	*20*	*21*	*22*	
y	*0—5*	*0*	*0*	*0*	*3*	*1*	*4*
	5—10	*0*	*0*	*0*	*3*	*2*	*5*
	10—15	*0*	*0*	*7*	*10*	*0*	*17*
	15—20	*0*	*5*	*4*	*0*	*0*	*9*
	20—25	*3*	*2*	*0*	*0*	*0*	*5*
	Total	*3*	*7*	*11*	*16*	*3*	*40*

NOTES

Solution.

Values of x	:	18	19	20	21	22
Deviation (u) from A = 20	:	− 2	− 1	0	1	2
Class of y	:	0—5	5—10	10—15	15—20	20—25
Mid-point (y)	:	2.5	7.5	12.5	17.5	22.5
Deviation from B = 12.5	:	− 10	− 5	0	5	10
Step deviation by k = 5 $\left(v = \frac{y - 12.5}{5}\right)$	:	− 2	− 1	0	1	2

Regression Table

	x	18	19	20	21	22	f	fv	fv^2	fuv
y	v \ u	−2	− 1	0	1	2				
0–5	− 2	$\boxed{0}$ 0	$\boxed{0}$ 0	$\boxed{0}$ 0	$\boxed{-6}$ 3	$\boxed{-4}$ 1	4	− 8	16	− 10
5–10	− 1	$\boxed{0}$ 0	$\boxed{0}$ 0	$\boxed{0}$ 0	$\boxed{-3}$ 3	$\boxed{-4}$ 2	5	− 5	5	− 7
10–15	0	$\boxed{0}$ 0	$\boxed{0}$ 0	$\boxed{0}$ 7	$\boxed{0}$ 10	$\boxed{0}$ 0	17	0	0	0
15–20	1	$\boxed{0}$ 0	$\boxed{-5}$ 5	$\boxed{0}$ 4	$\boxed{0}$ 0	$\boxed{0}$ 0	9	9	9	− 5
20–25	2	$\boxed{-12}$ 3	$\boxed{-4}$ 2	$\boxed{0}$ 0	$\boxed{0}$ 0	$\boxed{0}$ 0	5	10	20	− 16
f		3	7	11	16	3	N = 40	Σfv = 6	$\Sigma fv^2 = 50$	Σfuv = − 38
fu		− 6	− 7	0	16	6	Σfu = 9			
fu^2		12	7	0	16	12	$\Sigma fu^2 = 47$			
fuv		− 12	− 9	0	− 9	− 8	Σfuv = − 38			

Now
$$\bar{x} = \text{A} + \left(\frac{\Sigma fu}{\text{N}}\right) h = 20 + \left(\frac{9}{40}\right) 1 = 20.225$$

$$\bar{y} = \text{B} + \left(\frac{\Sigma fv}{\text{N}}\right) k = 12.5 + \left(\frac{6}{40}\right) 5 = 13.25$$

$$b_{yx} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fu^2 - (\Sigma fu)^2} \cdot \frac{k}{h} = \frac{40(-38) - (9)(6)}{40(47) - (9)^2} \cdot \frac{5}{1}$$

$$= \frac{-7870}{1799} = -4.375$$

$$b_{xy} = \frac{N\Sigma fuv - (\Sigma fu)(\Sigma fv)}{N\Sigma fv^2 - (\Sigma fv)^2} \cdot \frac{h}{k} = \frac{40(-38) - (9)(6)}{40(50) - (6)^2} \cdot \frac{1}{5}$$

$$= \frac{-1574}{9820} = -0.160.$$

The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$

or $y - 13.25 = -4.375\ (x - 20.225)$

or $y = -4.375x + 13.25 + (4.375)(20.225)$

or $\mathbf{y = -4.375x + 101.734.}$

The regression line of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$

or $x - 20.225 = -0.16\ (y - 13.25)$

or $x = -0.16y + 20.225 + (0.16)(13.25)$

or $\mathbf{x = -0.16y + 22.345.}$

EXERCISE 9.3

1. The following table shows the ages of daughters and mothers. Calculate the coefficients of regression and the regression equations :

	Age of daughter (x)				
Age of mother (y)	*5—10*	*10—15*	*15—20*	*20—25*	*25—30*
15—25	6	3	0	0	0
25—35	3	16	10	0	0
35—45	0	10	15	7	0
45—55	0	0	7	10	4
55—65	0	0	0	4	5

2. Following is the data relating 'Annual dividend (x)' and 'Security price (y)'. Compute the regression lines :

Security prices (in ₹)	*Annual Dividend (in ₹)*					
	6—8	*8—10*	*10—12*	*12—14*	*14—16*	*16—18*
130—140	0	0	1	3	4	2
120—130	0	1	8	3	3	1
110—120	0	1	2	3	2	0
100—110	0	2	3	2	0	0
90—100	2	2	1	1	0	0
80—90	3	1	1	0	0	0
70—80	2	1	0	0	0	0

Answers

1. $b_{yx} = 1.6045$, $b_{xy} = 0.4011$, $y = 1.6045x + 11.763$, $x = 0.4011y + 1.3769$
2. $y = 4.8042x + 55.8869$, $x = 0.1186y - 1.6032$

9.8. PROPERTIES OF REGRESSION COEFFICIENTS AND REGRESSION LINES

NOTES

(*i*) We have $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$.

σ_x and σ_y are always non-negative.

$\therefore$ The signs of b_{yx} and b_{xy} are same as that of r.

$\therefore$ The signs of regression coefficients and correlation coefficient are same.

Thus b_{yx}, b_{xy} and r are all either positive or negative.

(*ii*) $b_{yx} \cdot b_{xy} = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} = r^2.$

Now $0 \le r^2 \le 1$ because $-1 \le r \le 1$.

$\therefore \quad 0 \le b_{yx}\, b_{xy} \le 1.$

$\therefore$ **The product of regression coefficients is non-negative and cannot exceed one.**

(*iii*) $b_{yx} \cdot b_{xy} = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} = r^2$

$$\therefore \qquad \mathbf{r = \pm \sqrt{b_{yx}\, b_{xy}}}\,.$$

The sign of r is taken as that of regression coefficients.

(*iv*) The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$\Rightarrow \qquad y = b_{yx}x + (\bar{y} - b_{yx}\bar{x})$$

$\therefore$ **When y is kept on the left side, then the coefficient of x on the right side gives the regression coefficient of y on x.**

For example, let $4x + 7y - 9 = 0$ be the regression line of y on x.

We write this as $y = -\frac{4}{7}x + \frac{9}{7}$.

$\therefore$ Regression coefficient of y on x = coefficient of $x = -\frac{4}{7}$.

The regression line of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

$$\Rightarrow \qquad x = b_{xy}y + (\bar{x} - b_{xy}\bar{y})$$

$\therefore$ **When x is kept on the left side, then the coefficient of y on the right side gives the regression coefficient of x on y.**

For example, let $5x + 9y - 8 = 0$ be the regression line of x on y.

We write this as $x = -\frac{9}{5}y + \frac{8}{5}$.

$\therefore$ Regression coefficient of x on y = coefficient of $y = -\frac{9}{5}$.

(*v*) The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

This equation is satisfied by the point $(\bar{x}, \bar{y})$. This point also lies on the regression line of x on y :

NOTES

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$\therefore$ **The point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is common to both regression lines. In other words, if the correlation between the variables is not perfect, then the regression lines intersect at $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$.**

(*vi*) **Angle between the lines of regression**

The regression line of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

$$\Rightarrow \quad y = b_{yx}x + (\bar{y} - b_{yx}\bar{x}) \quad \therefore \quad \text{Slope} = b_{yx} = m_1 \text{ (say)}$$

The regression line of x on y is $x - \bar{x} = b_{xy}(y - \bar{y})$.

$$\Rightarrow \quad y = \frac{1}{b_{xy}}x + \left(\bar{y} - \frac{1}{b_{xy}}\bar{x}\right) \quad \therefore \quad \text{Slope} = \frac{1}{b_{xy}} = m_2 \text{ (say)}$$

Let θ be the acute angle between the regression lines.

$$\therefore \quad \tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right| = \left|\frac{b_{yx} - \dfrac{1}{b_{xy}}}{1 + b_{yx}\cdot\dfrac{1}{b_{xy}}}\right| = \left|\frac{b_{yx}b_{xy} - 1}{b_{xy} + b_{yx}}\right|$$

$$= \left|\frac{r\dfrac{\sigma_y}{\sigma_x}.r\dfrac{\sigma_x}{\sigma_y} - 1}{r\dfrac{\sigma_x}{\sigma_y} + r\dfrac{\sigma_y}{\sigma_x}}\right| = \left|\frac{r^2 - 1}{r\left(\dfrac{{\sigma_x}^2 + {\sigma_y}^2}{\sigma_x\sigma_y}\right)}\right|$$

$$= \frac{|r^2 - 1||\sigma_x\sigma_y|}{|r||{\sigma_x}^2 + {\sigma_y}^2|} = \frac{(1 - r^2)\,\sigma_x\sigma_y}{|r|({\sigma_x}^2 + {\sigma_y}^2)}$$

$$\therefore \quad \mathbf{\tan\theta = \frac{(1 - r^2)\,\sigma_x\sigma_y}{|r|({\sigma_x}^2 + {\sigma_y}^2)}}\ .$$

Particular cases :

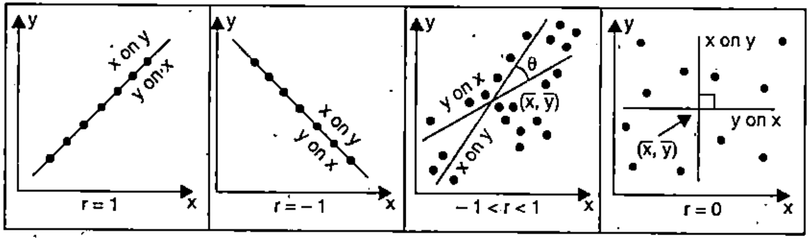
(*i*) **r = 0.** In this case, tan θ is not defined.

$\therefore$ θ = 90° *i.e.*, the regression lines are *perpendicular* to each other.

(*ii*) **r = 1 (or – 1).** In this case, tan θ = 0.

$\therefore$ The regression lines are *coincident*, because the point $(\bar{x}, \bar{y})$ is on both the regression lines.

Thus, we see that if the variables are not correlated, then the regression lines are perpendicular to each other and if the variables are perfectly correlated, then the regression lines are coincident. The closeness of regression lines measure the degree of linear correlation between the variables.



Example 14. *Are the following statements correct ? Give reasons :*

(i) The regression coefficient of y on x is 3.2 and that of x on y is 0.8.

(ii) The two regression coefficients are 0.8 and – 0.2.

(iii) The two regression coefficients are given to be 0.8 and 0.2 and the coefficient of correlation is 0.4.

(iv) 40x – 18y = 5 and 8x – 10y + 7 = 0 are respectively the regression equations of y on x and x on y.

Solution. (*i*) We have $b_{yx} = 3.2,\ b_{xy} = 0.8.$

$$b_{yx} \cdot b_{xy} = 3.2 \times 0.8 = 2.56 > 1.$$

This is impossible, because $0 \le b_{yx} \cdot b_{xy} \le 1.$

∴ The given statement is **false.**

(*ii*) We have $b_{yx} = 0.8$ and $b_{xy} = -\,0.2.$

This is impossible, because the regression coefficients are either both +ve or both –ve.

∴ The statement is **false.**

(*iii*) We have $b_{yx} = 0.8$ and $b_{xy} = 0.2.$

$$\therefore \qquad r = +\sqrt{b_{yx} \cdot b_{xy}} = +\sqrt{(0.8)(0.2)} = +\,0.4.$$

∴ The statement is **true.**

(*iv*) The regression line of y on x is $40x - 18y = 5.$

$$\Rightarrow \qquad 18y = 40x - 5 \ \Rightarrow\ y = \frac{40}{18}x - \frac{5}{18}$$

$$\therefore \qquad b_{yx} = \frac{40}{18} = \frac{20}{9}.$$

The regression line of x on y is $8x - 10y + 7 = 0.$

$$\Rightarrow \qquad 8x = 10y - 7 \ \Rightarrow\ x = \frac{10}{8}y - \frac{7}{8}.$$

$$\therefore \qquad b_{xy} = \frac{10}{8} = \frac{5}{4}.$$

$$\therefore \qquad b_{yx} \cdot b_{xy} = \frac{20}{9} \times \frac{5}{4} = \frac{100}{36} > 1.$$ This is impossible.

∴ The given statement is **false.**

Example 15. *Out of the following two regression lines, find the line of regression of x on y : 2x + 3y = 7 and 5x + 4y = 9.*

Solution. The regression lines are

$2x + 3y = 7$ ...(1) and $5x + 4y = 9.$ ...(2)

Let (1) be the regression line of x on y.

∴ (2) is the regression line of y on x.

NOTES

(1) $\Rightarrow$ $x = -\frac{3}{2}y + \frac{7}{2}$ $\therefore$ $b_{xy} = -\frac{3}{2}$

(2) $\Rightarrow$ $y = -\frac{5}{4}x + \frac{9}{4}$ $\therefore$ $b_{yx} = -\frac{5}{4}$

b_{xy} and b_{yx} are of same sign.

Also $b_{xy} \cdot b_{yx} = \left(-\frac{3}{2}\right)\left(-\frac{5}{4}\right) = \frac{15}{8} > 1$.

This is impossible because $0 \le b_{xy} \cdot b_{yx} \le 1$.

Our choice of regression line is incorrect.

(1) is not the regression line of x and y.

The regression line of x on y is **5x + 4y = 9.**

Example 16. *The equations of two regression lines obtained in a correlation analysis are $3x + 12y = 19$ and $3y + 9x = 46$. Obtain :*

(*i*) *the mean values of x and y,*

(*ii*) *the value of correlation coefficient.*

Solution. The regression equations are

$3x + 12y = 19$ ...(1) $9x + 3y = 46$ ...(2)

(*i*) We know that the regression lines passes through the point $(\bar{x}, \bar{y})$.

$\therefore$ The values of $\bar{x}$ and $\bar{y}$ can be obtained by solving the regression equations.

(1) × 3 $\Rightarrow$ $9x + 36y = 57$ ...(3)

(2) – (3) $\Rightarrow$ $0 - 33y = -11$

or $y = \frac{-11}{-33} = \frac{1}{3}$ $\Rightarrow$ $\bar{y} = \frac{1}{3}$

$\therefore$ (1) $\Rightarrow$ $3x + 12(1/3) = 19$ $\Rightarrow$ $\bar{x} = 5$.

$\therefore$ The means of x and y are 5 and 1/3 respectively.

(*ii*) We don't know exactly as to which of the above equations is regression equation of y on x. Let us suppose that (1) is regression equation of x on y and (2) is regression equation of y on x.

(1) $\Rightarrow$ $x = -4y + \frac{19}{3}$ $\therefore$ $b_{xy} = -4$

(2) $\Rightarrow$ $y = -3x + \frac{46}{3}$ $\therefore$ $b_{yx} = -3$

$\therefore$ $b_{xy} \cdot b_{yx} = (-4)(-3) = 12 > 1$. This is impossible.

$\therefore$ Our supposition is wrong.

$\therefore$ (1) is the regression equation of y on x and (2) is the regression equation of x on y

(1) $\Rightarrow$ $y = -\frac{1}{4}x + \frac{19}{12}$ $\therefore$ $b_{yx} = -\frac{1}{4}$

(2) $\Rightarrow$ $x = -\frac{1}{3}y + \frac{46}{9}$ $\therefore$ $b_{xy} = -\frac{1}{3}$

$\therefore$ $r = -\sqrt{b_{yx} \cdot b_{xy}} = -\sqrt{\left(-\frac{1}{4}\right)\left(-\frac{1}{3}\right)} = \mathbf{-0.2887}$.

NOTES

Example 17. *For the following observations, find the regression coefficients b_{yx} and b_{xy} and hence find the correlation coefficient between x and y :*

$$\{(x, y) : (4, 2), (2, 3), (3, 2), (4, 4), (2, 4)\}.$$

Solution. **Calculation of b_{yx} and b_{xy}**

S. No.	*x*	*y*	*xy*	x^2	y^2
1	4	2	8	16	4
2	2	3	6	4	9
3	3	2	6	9	4
4	4	4	16	16	16
5	2	4	8	4	16
$n = 5$	$\Sigma x = 15$	$\Sigma y = 15$	$\Sigma xy = 44$	$\Sigma x^2 = 49$	$\Sigma y^2 = 49$

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{5(44) - (15)(15)}{5(49) - (15)^2} = \frac{-5}{20} = -\frac{1}{4}$$

$$b_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{5(44) - (15)(15)}{5(49) - (15)^2} = \frac{-5}{20} = -\frac{1}{4}.$$

The regression coefficients are –ve, so the correlation coefficient is also –ve.

$$\therefore \quad r = -\sqrt{(b_{yx})(b_{xy})} = -\sqrt{\left(-\frac{1}{4}\right)\left(-\frac{1}{4}\right)} = -\frac{1}{4}.$$

Example 18. *The two regression lines obtained by a student were as given below :*

$$3X - 4Y = 5 ; \quad 8X + 16Y = 15$$

Do you agree with him ? Explain with reasons.

Sol. The regression lines are

$$3X - 4Y = 5 \quad \text{....(1)} \qquad 8X + 16Y = 15 \quad \text{...(2)}$$

Let (1) be the regression line of Y on X.

$\therefore$ (2) is the regression line of X on Y.

(1) $\Rightarrow$ $\quad Y = \frac{3}{4}X - \frac{5}{4} \qquad \therefore \quad b_{YX} = \frac{3}{4}$

(2) $\Rightarrow$ $\quad X = -2Y + \frac{15}{8} \qquad \therefore \quad b_{XY} = -2$

This is impossible because regression coefficients cannot be of different signs.

Let (1) be the regression line of X on Y.

$\therefore$ (2) is the regression line of Y on X.

(1) $\Rightarrow$ $\quad X = \frac{4}{3}Y + \frac{5}{3} \qquad \therefore \quad b_{XY} = \frac{4}{3}$

(2) $\Rightarrow$ $\quad Y = -\frac{1}{2}X + \frac{15}{16} \qquad \therefore \quad b_{YX} = -\frac{1}{2}$

This is also impossible.

$\therefore$ We do not agree with the student.

EXERCISE 9.4

1. If two regression coefficients are 2 and 0.45, what will be the coefficient of correlation ?
2. Out of the following two regression lines, find the regression line of y on x :
   (*i*) $3x + 12y = 8$, $3y + 9x = 46$
   (*ii*) $x + 2y - 5 = 0$, $2x + 3y = 8$.
3. From the following data :

x	4	7	10	12	18
y	12	15	8	13	18

   Verify that correlation coefficient is G.M. between the regression coefficients.
4. The regression lines between two variables x and y are found to be
   $$4x - 5y + 33 = 0 \text{ and } 20x - 9y = 107.$$
   Find the coefficient of correlation.
5. The equations of two regression lines obtained in a correlation analysis are as follows :
   $$2x + 3y - 10 = 0,\ 4x + y - 5 = 0.$$
   Obtain (*i*) the means of x and y
   (*ii*) the regression coefficients b_{yx} and b_{xy}
   (*iii*) the correlation coefficient.
6. A student obtained the two regression equations as :
   $$2x - 5y - 7 = 0 \text{ and } 3x + 2y - 8 = 0.$$
   Do you agree with him ?
7. The lines of regression in a bivariate distribution are $x + 2y = 5$ and $2x + 3y - 8 = 0$. Find means of x and y. Also find the correlation coefficient r_{xy} and regression coefficients b_{yx} and b_{xy}.
8. The equations of regression lines are given to be
   $$3x + 2y = 26 \text{ and } 6x + y = 31.$$
   A student obtained the mean values $\bar{x} = 7$, $\bar{y} = 4$ and coefficient of correlation $r = 0.5$. Do you agree with him ? If not, suggest your results.
9. Two regression equations are given below :
   Find out :
   (*i*) Mean values of X and Y :
   (*ii*) Standard deviation of Y.
   (*iii*) Coefficient of correlation between X and Y.
   The regression equations are $8X - 10Y + 70 = 0$. $15X - 6Y = 60$, variance of $X = 9$.

Answers

1. 0.9487   2. (*i*) $3x + 12y = 8$ (*ii*) $x + 2y - 5 = 0$
4. $r = 0.6$
5. (*i*) $\bar{x} = \frac{1}{2}$, $\bar{y} = 3$ (*ii*) $b_{yx} = -\frac{2}{3}$, $b_{xy} = -\frac{1}{4}$ (*iii*) $r = -\frac{1}{\sqrt{6}}$
6. No
7. $\bar{x} = 1$, $\bar{y} = 2$, $r_{xy} = -0.866$, $b_{yx} = -0.5$, $b_{xy} = -1.5$
8. No. $\bar{x} = 4$, $\bar{y} = 7$, $r = -0.5$
9. (*i*) $\overline{X} = 10$, $\overline{Y} = 15$ (*ii*) $\sigma_X = 3\sqrt{2}$ (*iii*) $r = 2\sqrt{2}/5$.

BBA-201

सर्वे भवन्तु सुखिनः सर्वे सन्तु निरामयाः !
सर्वे भद्राणिः पश्यन्तु माकश्चिद् दुःख भाग्भवेत् !!

Directorate of Distance Education
SWAMI VIVEKANAND
SUBHARTI
UNIVERSITY
UGC Approved
Meerut
Where Education is a Passion ...
Subharti Puram, N.H.—58, Delhi-Haridwar By Pass Road,
Meerut, Uttar Pradesh 250005
Website: www.subhartidde.com , E-mail: ddesvsu@gmail.com

ACCREDITED WITH GRADE A NAAC