

Alagar

Business Statistics

K Alagar



Business Statistics



TATA
McGRAW
HILL

business statistics

The McGraw-Hill Companies

BUSINESS STATISTICS

ABOUT THE AUTHOR

Dr K Alagar, Reader in Commerce, is a dedicated and devoted faculty member of Yadava College, an autonomous institution affiliated to Madurai Kamaraj University, Madurai, Tamil Nadu. During his teaching career spanning over 25 years, Dr. K. Alagar has been teaching subjects like statistics, banking, operations research and costing. Dr. Alagar has been actively involved in research for over 15 years and in guiding students for degrees leading for M.Phil and Ph.D. He has published research papers in national journals such as Co-operative Sugar Industrial Herald etc., He has presented several research papers in national and international level conferences, of which three have been published in the proceeding of international conference Dr. K. Alagar has edited two books, namely *Rural Industrialisation* and *Rural Developments*.

BUSINESS STATISTICS

K. Alagar

Reader in Commerce
Yadava College
Madurai



Tata McGraw-Hill Education Private Limited
NEW DELHI

McGraw-Hill Offices

New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto



Tata McGraw-Hill

Published by Tata McGraw-Hill Education Private Limited,
7 West Patel Nagar, New Delhi 110 008.

Business Statistics

Copyright © 2009, by Tata McGraw-Hill Education Private Limited. No part of this publication may be reproduced or distributed in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise or stored in a database or retrieval system without the prior written permission of the publishers. The program listings (if any) may be entered, stored and executed in a computer system, but they may not be reproduced for publication.

This edition can be exported from India only by the publishers,
Tata McGraw-Hill Education Private Limited.

ISBN (13): 978-0-07-014503-0
ISBN (10): 0-07-014503-2

Managing Director: *Ajay Shukla*
General Manager—Publishing (B&E/HSSL & School): *V Biju Kumar*
Publishing Manager—B&E: *Tapas K Maji*
Assistant Sponsoring Editor: *Piyali Ganguly*
Editorial Executive: *Hemant K Jha*
Development Editor: *Shalini Negi*
Assistant Manager (Editorial Services): *Anubha Srivastava*
Senior Production Manager: *Manohar Lal*
Production Executive: *Atul Gupta*
General Manager—Marketing (Higher Education & School): *Michael J Cruz*
Product Manager: *Vijay Sarathi Jagannathan*
General Manager—Production: *Rajender P Ghansela*
Assistant General Manager—Production: *B L Dogra*

Information contained in this work has been obtained by Tata McGraw-Hill, from sources believed to be reliable. However, neither Tata McGraw-Hill nor its authors guarantee the accuracy or completeness of any information published herein, and neither Tata McGraw-Hill nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that Tata McGraw-Hill and its authors are supplying information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

Typeset at Script Makers, 19, A1-B, DDA Market, Paschim Vihar, New Delhi 110 063
and printed at Avon Printers, Plot No. 16, Main Loni Road, Jawahar Nagar, Industrial Area, Shahdara, Delhi 110094

Cover Designer: K Anoop
Cover Printed: SDR Printers
RQDBCRCFRQCQY

Dedications to

My Students

Present

Past

Future

The McGraw-Hill Companies

PREFACE

It gives me a great pleasure to reach out to the readers through this new text book on ***Business Statistics***. Reasons may be many for authoring a book, the reason for my writing this book is my passion for my subject.

This book covers the content of the syllabi of Business Statistics introduced in various Indian universities, designed to meet the requirements for the students of B.Com, BBA, BBE, B.Sc (I.T) and BA (Co-op), M.Com, CA (PE II), ICWA (Inter), MBA and other management courses.

The main aim of writing this book is to present a clear, simple, systematic and comprehensive exposition of the principles, methods and techniques of statistics in various disciplines with special reference to Economics, Commerce and Business Administration. All the methods are illustrated with a good number of problems followed by a collection of graded exercises.

My sincere gratitude to Tata McGraw-Hill Publishing Company (P) Limited, New Delhi for their keen interest and encouragement in bringing out this book.

Suggestions for the improvement of this book shall be gratefully acknowledged.

K. Alagar
March 31, 2009

Dept. of Commerce
Yadava College
Madurai

The McGraw-Hill Companies

ACKNOWLEDGEMENTS

I am immensely indebted and thankful to Thiru K P Navaneetha Krishnan, Secretary and Correspondent of Yadava College for his motivation and encouragement.

I am thankful to Prof P Rengan, our respected Principal, for his consistent motivation and guidance.

With respect and regards, I register my thanks to Prof P Ramapandi, vice principal and Prof K Kannan, Director, Self Finance courses for their warm and friendly gestures.

I am indebted to Dr G Thiruvasagam, Vice Chancellor, Bharathiar University, Coimbatore for the vision and inspiration he ignited in me to for this venture.

I owe my thank to Dr K Alagusundram, Head, Dept of Commerce of our college and all my colleagues Dr R S Mani, Dr R Sundar Babu, Dr V Sampath, Mr T Retnakumar, Dr N Kannan, Dr G Sivaji Ganesan, Mr M Bala Subhramanian, Mr N Malaiselvam, Mr G Ayyanar for their co-operation and encouragement.

I am greatly indebted to my uncle Mr A Arumugam, G M (Retd), Pudukottai District Central Co-operative Bank, for his persuasion throughout my career.

I thank my wife Dr R Poovazhakai, Head of the Maths Dept, EMG Yadava Women's College, Madurai for her support in all my walks of life.

K. Alagar

The McGraw-Hill Companies

CONTENTS

<i>Preface</i>	vii
<i>Acknowledgements</i>	ix
Chapter 1 Statistics—An Overview	1
1.1 Introduction	1
1.2 Origin and Growth of Statistics	1
1.3 Definition of Statistics	2
1.4 Is Statistics a Science or an Art?	4
1.5 Statistical Investigation (Statistical Enquiry)	4
1.6 Functions of Statistics	5
1.7 Scope of Statistics or Application of Statistics	6
1.8 Limitations of Statistics	7
1.9 Statistical Method vs. Experimental Method	8
1.10 Distrust of Statistics	9
Summary	9
Exercises	11
Chapter 2 Collection of Data	14
2.1 Introduction	14
2.2 Methods of Collecting Primary Data	14
2.3 Sources of Secondary Data	22
2.4 Precautions to be Taken Before Using Secondary Data	24
Summary	24
Exercises	26
Chapter 3 Classification and Tabulation	29
3.1 Introduction	29
3.2 Definition	29
3.3 Chief Characteristics of Classification	29
3.4 Objectives of Classification	30
3.5 Rules of Classification	30
3.6 Types of Classification	31
3.7 Statistical Series	33
3.8 Frequency Distribution	34
3.9 Tabulation of Data	43
Summary	51
Exercises	53
Chapter 4 Diagrammatic and Graphical Representation	57
4.1 Introduction	57
4.2 Diagram	57
4.3 Rules for Preparing Diagrams	57
4.4 Types of Diagrams	58
4.5 Graphical Representation	68

xii *Contents*

4.6 Miscellaneous	77
Summary	85
Exercises	86
Chapter 5 Sampling	97
5.1 Introduction	97
5.2 Population Method or Census Method	97
5.3 Methods of Sampling	98
5.4 Sampling	99
5.5 Theoretical Basis of Sampling	103
5.6 Sampling and Non-sampling Errors	105
Summary	107
Exercises	109
Chapter 6 Measures of Central Tendency	112
6.1 Introduction	112
6.2 Definitions	112
6.3 Types of Averages	114
6.4 Miscellaneous Illustrations	147
Summary	171
Exercises	178
Chapter 7 Measures of Dispersion	184
7.1 Meaning	184
7.2 Definition of Dispersion	184
7.3 Objectives or Importance of the Measures of Dispersion	184
7.4 Methods of Measuring Dispersion	185
7.5 Miscellaneous Illustrations	212
Summary	238
Exercises	241
Chapter 8 Skewness, Kurtosis and Moments	252
8.1 Introduction	252
8.2 Definitions	252
8.3 Measures of Skewness	254
8.4 Moments	266
8.5 Kurtosis	271
8.6 Miscellaneous Illustrations	274
Summary	305
Exercises	308
Chapter 9 Correlation Analysis	314
9.1 Meaning	314
9.2 Definitions	314
9.3 Utility of Measuring Correlation (Usefulness)	314
9.4 Correlation and Causation	315
9.5 Types of Correlation	315
9.6 Methods of Studying Correlation	316

9.7 Lag and Lead in Correlation	337
9.8 Miscellaneous Illustrations	343
Summary	363
Exercises	366
Chapter 10 Regression Analysis	374
10.1 Meaning	374
10.2 Definitions	374
10.3 Uses of Regression Analysis	375
10.4 Regression Lines	375
10.5 Regression Equations	375
10.6 Standard Error of Regression	383
10.7 Miscellaneous Illustrations	394
Summary	427
Exercises	429
Chapter 11 Index Numbers	436
11.1 Introduction	436
11.2 Definitions	436
11.3 Methods of Index Numbers	440
11.4 Miscellaneous Illustrations	472
Summary	493
Exercises	496
Chapter 12 Time Series Analysis	504
12.1 Introduction	504
12.2 Definitions	504
12.3 Uses of Time Series Analysis	504
12.4 Components of Time Series	505
12.5 Preliminary Adjustments before Analysing Time Series	508
12.6 Measurement of Trend	509
12.7 Measurement of Seasonal Variation	525
12.8 Miscellaneous Illustrations	535
Summary	552
Exercises	554
Chapter 13 Interpolation and Extrapolation	560
13.1 Introduction	560
13.2 Definitions	560
13.3 Uses	561
13.4 Methods of Interpolation and Extrapolation	561
13.5 Miscellaneous Illustrations	580
Summary	600
Exercises	602
<i>Appendix</i>	607

1

CHAPTER

STATISTICS—AN OVERVIEW

1.1 INTRODUCTION

The word ‘statistics’ is derived from the Latin word ‘status’ or Italian word ‘statista’ or the German word ‘statistik’ or French word ‘statistique’ each of which means a political state. Statistics is not a new discipline but is as old as the human activity itself.

Statistics is the study of facts in numbers or quantity. The science of statistics is said to have originated from two main sources namely government records and mathematics. The government collected statistics for administrative purposes and thus it was recorded as science of kings or the science of state graft. Statistical methods are now widely used in various diversified fields such as agriculture, economics, sociology, business management, computer, finance. Statistics is an important tool for taking decisions and all the branches of science make use of it.

1.2 ORIGIN AND GROWTH OF STATISTICS

Census of population and wealth were taken even in ancient times. In 1400 BC, a census of all lands in Egypt was done. Similar reports on the ancient Chinese, Greeks and Romans are also available. People and land were the earliest objects of statistical enquiry.

Many great men like Tycko Brave (1554–1601), John Graunt (1620–1674), William Petty (1623–1687), Sussimikts (1707–1767), J. Bernoulli (1654–1705), Laplace (1749–1827), Lagrange, Gauss, Lexis, Charlier, Francis Galton (1822–1921), S. Jevon (1835–1882), Karl Pearson (1857–1936), Ronald A. Fisher (1890–1962) have contributed to the development of statistics.

The word statistics is used as singular or plural. Statistics as a singular noun refers to the various methods adopted for collection, classification, analysis

2 Business Statistics

and interpretation. As a plural noun, statistics refers to data or facts. It means data relating to income, expenditure, population, production, sales, profit, employment, literacy. Statistics means collection, classification, analysis and interpretation of data.

Reasons for the growth of statistics are:

1. The increasing demand for statistics and,
2. The decreasing cost of statistics.

1.3 DEFINITION OF STATISTICS

Various authors have defined statistics in different ways. But, it can be divided into two distinct forms: (a) Statistics as data and (b) Statistics as methods.

1.3.1 Statistics as Data

In this sense, the term statistics means a set of numerical or quantitative statement of facts.

(i) Webster defines

“Statistics are the classified facts representing the conditions of the people in a state.....specially those facts which can be stated in numbers or any tabular or classified assignment”.

(ii) Yule and Kendall define that

“By statistics we mean quantitative data affected to a marked extent by multiplicity of causes”.

(iii) A.L. Bowley defines

“Statistics are numerical statements of facts in any department of enquiry placed in relation to each other”.

(iv) Prof. Horace Secrist defines

“Statistics are aggregates 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 pre-determined purpose and placed in relation to each other.”

It may be emphasised that this definition highlights a few major characteristics or features of statistics.

1.3.2 Characteristics or Features of Statistics

(i) Statistics are Aggregate of Facts Only data related to facts are termed as statistics. This data should be plural one. Single or unrelated figures are not statistics. Since single figure could not be compared, it cannot be termed as statistics—for example, a single figure relating to the height or marks of the student.

The figures without any relationship is also not called as statistics—for example, figures like 15, 35, 50, 70, etc. If these figures are related with production

or sales etc., then it is called statistics—for example, products – 15 tonnes, 35 tonnes, 50 tonnes (or) sales – 35 units, 50 units, 70 units etc.

(ii) Statistics are Affected to a Marked Extent of Multiplicity of Causes Only those facts, which are consequential to multiple causation are statistics—for example, the price of a particular commodity is affected by number of factors such as its demand, supply, money circulation, exports, imports etc.

(iii) Statistics must be Numerically Expressed Statistics is the study of only those facts which are capable of being stated in numbers or quantity. For example, it is only the numerical data which constitutes statistics. Any qualitative expression such as intelligence, honesty, character etc., will not be called statistics unless some numerical equivalence could be established.

(iv) Statistics must be Enumerated or Estimated According to Reasonable Standard of Accuracy Enumeration means collecting data by actual counting or measurement complete accuracy may not be required on all occasions hence, the data may have to be estimated. This estimation should be done with the reasonable standard of accuracy. This reasonable standard of accuracy varies according to the nature of data, the importance of data and the purpose of the data.

(v) Statistics should be Collected in a Systematic Manner The data collected would be a systematic one. Proper arrangement in a systematic manner would minimise the time, as well as provide accurate information.

1.3.3 Statistics as Methods

In this sense, the term statistics is used in singular. By analytical statistics, we mean the application of different statistical tools in analysing and interpreting the collected data.

Some of the popular definitions of statistics as statistical methods are given below.

(i) A.L. Bowley defines statistics as

“Statistics may be called the science of counting.”

“Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations.”

“Statistics may rightly be called the science of average”.

(ii) According to Croxton and Cowden

“Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data.”

(iii) Loving defines statistics as

“Statistics is the science which deals with classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon.”

4 Business Statistics

(iv) W.I. King defines statistics as

“The science of statistics is the method of judging collective, natural or social phenomena from the results obtained by the analysis of an enumeration or collecting of estimates.”

1.4 IS STATISTICS A SCIENCE OR AN ART?

Science is the body of systematic knowledge. The characteristics of any science is the observation of certain facts. As a science, the statistical method is a part of the general scientific method and is based on the same fundamental idea and process.

Art is the skill or the power of performing certain actions. It teaches as how to do things. It is the practical application of set-up rules or principles to practice. Statistics possesses the important characteristics of an art; so, it may be concluded that it is an art.

Thus, we can conclude that statistics is both science and art. It is a science, as its methods are basically systematic. It is an art too, as its successful application depends to a considerable degree on the skill and experience of the statistician.

1.5 STATISTICAL INVESTIGATION (STATISTICAL ENQUIRY)

Statistical investigation is concerned with investigation of some problem with the help of statistical methods. Data related to the problem is the basis for investigation. Hence, the data related to the problem should be collected. The investigator is the person who conducts statistical enquiry. The respondent are the persons from whom the information is collected.

The data so collected should be analysed and processed with suitable statistical devices. The analysed data gives valuable information about the problem under investigation.

1.5.1 Stages in a Statistical Investigation

The important stages in statistical investigation are:

1. Planning a statistical enquiry
2. Collection of data
3. Organisation of data
4. Presentation of data
5. Analysis of data and
6. Interpretation of data

I. Planning a Statistical Enquiry The first step in statistical investigation is planning. Planning should be done as per the objectives and scope of the investigation. The investigator should decide in advance about the type of enquiry to be conducted, source of information, period of study and the unit of study.

2. Collection of Data The next step after planning is collection of data. The data may be collected from primary source and from secondary source. The data collected should be accurate. There should not be any bias in collecting the data.

3. Organisation of Data Organisation of data means arrangement of the data in a systematic manner. The data collected should be edited to eliminate unnecessary data. Then it should be classified on the basis of some common characteristics.

4. Presentation of Data Organised data may be presented in the form of diagrams and graphs. It helps in easy understanding of the characteristics and nature of the data.

5. Analysis of Data The data presented in table form should be analysed to draw conclusion. Various statistical methods like average, dispersion, correlation, regression, index number, time series, interpolation and extrapolation etc., are available for analysing the data.

6. Interpretation of Data The last stage in the statistical investigation is the interpretation of data. Interpretation means drawing conclusions. If interpretation is not properly done, it may give wrong conclusion. Hence, much care should be taken for interpretation of data to get valid conclusion.

1.6 FUNCTIONS OF STATISTICS

The important functions of statistics are as follows:

1. It presents facts in a proper form
2. It simplifies unwidely and complex data
3. It facilitates for comparison
4. It enlarges individual experience
5. It facilitates for formulating policies
6. It tests hypotheses
7. It measures the trend behaviour

1. It Presents Facts in a Proper Form Statistics presents facts in a proper form. Facts presented in quantitative term gives more meaning than it is presented in a statement form.

2. It Simplifies Unwidely and Complex Data Data presented in raw form are unwidely and complex. If there is huge data, it leads to have complication for getting an idea about the data. The important function of statistics is to simplify the data.

6 Business Statistics

3. It Facilitates for Comparison Comparison is also one of the important functions of statistics. It enables us to understand the behaviour of data over a time period or at a point of time. Statistical devices like averages, ratios, coefficient, graphs and diagrams offer the best way of comparison between two phenomena.

4. It Enlarges Individual Experience The proper function of statistics is to enlarge individual experience. Statistics is a science which provides opportunities to individuals to enrich their knowledge and experience.

5. It Facilitates for Formulating Policies Statistics also helps in the formulation of various economic, business and other policies at state, national or global level. For example, framing of government policies on education, taxation, pollution, law and order, import and export, social welfare, wages etc., are formulated on the basis of statistical data and inferences drawn from their analysis.

Business organisations also make use of statistics to design their policies in areas of finance, marketing and personnel.

6. It Tests Hypothesis Hypothesis is an important concept in research studies. Statistics provides various methods for testing the hypothesis. The important methods for hypothesis test are chi-square, Z-test, t-test and F-Test.

7. It Measures the Trend Behaviour Statistics helps for projecting the future with the help of present and past data. Hence, plans, programmes and policies are formulated in advance with the help of statistical techniques. The important techniques adopted for measuring the trend are time series and regression analysis.

1.7 SCOPE OF STATISTICS OR APPLICATION OF STATISTICS

In olden days, statistics was used to collect information. Now, its scope was not only in collection but also in analysing the data collected, drawing inferences or in formulating a theory. Hence, it helps for planning and formulation of policies. The methods and techniques in statistics help people to solve problems presented in statistical data. Hence, it has a good application in the field of commerce, economics, physics, chemistry, botany, zoology, psychology etc.

Statistics and Business Statistics is most commonly used in business. The statistical data regarding the demand and supply of products can be collected and analysed to take a decision regarding the new business. Thus it helps to take decision regarding whether a company can start a new business. The existing companies can also make a comparative study about their performance with the performance of other companies through statistical analysis.

The existing companies can also project their future with regression and correlation analysis.

Statistics and Economics Some of the uses of statistics in economics are as follows:

- Measures of gross national product and input–output analysis have greatly advanced overall economic knowledge and opened up entirely new fields of study.
- Financial statistics are basic in the field of money and banking, short-term credit, consumer finance and public finance.
- Statistical studies of business cycles, long-term growth and seasonal fluctuations serve to expand our knowledge of economic instability and to modify older theories.

Statistics and Physical Sciences The physical sciences are making increasing use of statistics, especially astronomy, chemistry, geology, meteorology and physics.

Statistics and Natural Sciences Statistical techniques have proved to be extremely useful in the study of all natural sciences like biology, medicine, zoology, botany.

Statistics and Research Statistics is indispensable in research work. Most of the research findings in various disciplines of knowledge have great importance along with subject of statistics.

Statistics and Computer Statistical tools like SPSS package, multiple discriminant analysis, multiple regression analysis will help all the fields of the management with the help of computer.

Statistics and Management Most of the managerial decisions are taken with the help of statistics. The data regarding the performance of a company will facilitate to take decision regarding future. Statistical techniques like correlation analysis, regression analysis and time series technique can be used in this regard. Statistical techniques can also be used for the payment of wages to the employees of the company.

Statistics in Banking and Finance Statistics are mostly used in banking and finance. In banks, statistical data regarding loan, the customer deposit etc. are represented in statistical data. Financial institutions like Industrial Development Bank of India, state financial corporation of India also use statistics in projecting the future and to solve various statistical problems.

1.8 LIMITATIONS OF STATISTICS

Though statistics has a wide scope and utility, it has certain limitations. They are as follows:

8 Business Statistics

- 1. Statistics does not Study Qualitative Aspect** Statistics deals with facts which are expressed in numerical terms, that is, statistics deals only with quantitative data and not the qualitative such as honesty, efficiency, intelligence etc.
- 2. Statistics does not Deal with Individuals** Statistics deals with aggregate of facts. It requires a series of figures for calculating averages and for analysis. The individual measurement has no recognition.
- 3. Statistical Results are not Perfectly Accurate** Statistical theories won't give accurate results. The results would be only approximate value. First of all, the data collected for analysis may not be accurate also.
- 4. Data must be Uniform in Statistics** Data used for statistical analysis must be uniform. For example, the data related with the income of people could not be mixed with the data related to expenditure of the people. These two aspects should be studied separately.
- 5. Statistics can be Misused** Only experienced and efficient persons can handle statistics in a proper way. Untrained and inefficient persons may not produce accurate result with the help of statistical techniques.

1.9 STATISTICAL METHOD vs. EXPERIMENTAL METHOD

In early period, human beings acquired knowledge by perception and intuition. But now, knowledge is acquired through scientific methods rather than as a matter of chance. There are two suitable methods for enlarging their knowledge. They are:

- (i) Experimental method and
- (ii) Statistical method

1.9.1 Experimental Method

In this method, people acquire knowledge by observation or experiments. Various experiences are conducted to study or prove a theory in natural science like physics, chemistry. It is also applied in social science like commerce, economics, psychology, political science etc.

In this method, two groups namely controlled group and experimental group should be selected. The cause and effect relationship can be studied in this method.

1.9.2 Statistical Method

In this method, the cause and effect relationship is not measured. The variations among all the variables is measured. It leads to have a comparative study which gives more knowledge to the individuals.

1.10 DISTRUST OF STATISTICS

Distrust of statistics means lack of confidence (or belief) in statistical data, statistical methods and the conclusions drawn. The statistical statement or science of statistics is always subject to doubt and suspicious to the public because of its misuse by unscrupulous elements for their motive. Statistics has been looked at doubtful eyes.

Some of the common beliefs about statistics are as follows:

- (i) An ounce of truth will produce tonnes of statistics.
- (ii) Do not believe statistics.
- (iii) Statistics are the lies of the first order.
- (iv) Figures do not lie – liars figure.
- (v) There are three types of lies – lies, damned lies and statistics.

Some of the reasons for distrust of statistics:

- (i) Failure to present the complete data.
- (ii) Statistical data does not bear on the qualities.
- (iii) In appropriate comparison.
- (iv) Selection of non-representations statistical units.
- (v) Statistical tools have their own limitations.
- (vi) Data collected by inexperience enumerations.

SUMMARY

Statistics

Derived from Latin word ‘status’ or the Italian word ‘statista’ or German word ‘statistik’ which means political state. It is used as singular or plural. Statistics is the study of facts in numbers in quantity.

Statistics as a Singular noun

Refers to the various methods adopted for collection, classification, analysis and interpretation.

Statistics as a Plural noun

Refers to data or facts. It means data relating to income, expenditure, population, production, sales, profit, employment, literacy.

Statistical Investigation (or) Statistical enquiry

Concerned with investigation of some problem with the help of statistical methods.

Stages in Statistical Investigation

- Planning a statistical enquiry
- Collection of data
- Organisation of data

10 Business Statistics

- Presentation of data
- Analysis of data and
- Interpretation of data.

Functions of Statistics

- It present facts in a proper form.
- It simplifies unwidely and complex data.
- It facilitates for comparison.
- It enlarges Individual experiences.
- It facilitates for formulating policies.
- It tests hypotheses.
- It measures the trend behaviour.

Scope of Statistics

- Statistics in Business
- Statistics in Management
- Statistics in Banking and Finance
- Statistics and economics
- Statistics and physical science
- Statistics and natural science
- Statistics and research
- Statistics and computer

Limitations of Statistics

- It studies only quantitative characteristics.
- Deals with aggregates and not with individual measurement.
- It is not perfectly accurate.
- Data must be uniform.
- It can be misused.

Methods for Enlarging the Knowledge

- Experimental method
- Statistical method

Experimental Method

People acquire knowledge by observations or experiments. The cause and effect relationship can be studied.

Statistical Method

People acquire knowledge through comparison of data (or) comparative study. The cause and effect relationship is not measured.

EXERCISES

(a) Choose the best option.

1. The word ‘statistics’ seems to have been derived from the Latin word
(a) statistik (b) status (c) statista
2. Statistics is most commonly used in
(a) maths (b) science (c) social sciences
3. Statistics is the _____ of estimates and probabilities.
(a) science (b) economics (c) sociology
4. Statistics is essential for a
(a) city (b) state (c) country
5. Laws of _____ science are perfect.
(a) physical (b) moral (c) social

Answers

1. b 2. c 3. a 4. c 5. a

(b) Fill in the blanks.

1. Statistics is _____ statement.
2. _____ or _____ data are influenced by a number of factors.
3. Numerical data alone constitute _____.
4. A reasonable standard of accuracy is needed in both _____ and _____.
5. The first step of an investigation is _____.
6. Statistics is widely used in _____.
7. Economics is a _____.
8. _____ is one of the main function of statistics.
9. Statistical methods are helpful to develop _____.

Answers

- | | |
|---------------------------|------------------------------|
| 1. Numerical | 2. Quantitative, statistical |
| 3. Statistics | 4. Enumeration, Estimation |
| 5. The collection of data | 6. Education |
| 7. Social science | 8. Comparison |
| 9. New theories | |

(c) Theoretical Questions

1. Define Statistics.
2. Define Statistics. Explain the scope of statistics.

12 *Business Statistics*

3. What are the functions of statistics?
4. State the important steps in statistical investigation.
5. Distinguish between statistical methods and experimental methods.
6. Mention four uses of statistical methods.
7. Explain the importance of statistics in arriving at policy decision in any economy.
8. What do you understand by statistical methods?
9. What are the limitations of statistics?
10. What is a statistical investigation? Describe the steps in statistical investigation.
11. What are the causes of distrust of statistics?
12. Define statistics. Why some people look at this science with an eye of distrust?
13. Statistics is said to be both science and art. Why?
14. “Statistics is a body of methods for making wise decisions in the face of uncertainty.” Comment on the statement bring out clearly how does statistics help in business decision-making.

(B.Com., MKU, MSU, BDU)

15. “Statistics is the science of estimates and probabilities.” Elucidate this statement and give a more comprehensive definition of the science of statistics.
16. “Statistics affects everybody and touches life at many points. It is both a science and an art.” Explain this above statement with suitable examples.

(B.Com., MKU, CHU, BDU)

17. “Statistics are numerical statements of facts but all facts numerically stated are not statistics.” Comment upon the statement and state briefly which numerical statements of facts are not statistics.

(B.Com., CHU, BDU, BU)

18. “There is hardly any field which does not fall within the scope of statistics.” Elaborate this statement.
19. Discuss the following statements:
 - (i) ‘Statistics are numerical statements of facts, but all facts numerically stated are not statistics’.
 - (ii) ‘The science of statistics is the most useful servant but only of great value to those who understand its proper use.’
 - (iii) ‘Statistics are like clay of which you can make a God or Devil as you please.’

- 20.** “Science without statistics bears no fruit, statistics without science have no root”. Comment.
(B.Com., MKU, BU, MSU)
- 21.** “The proper function of statistics is to enlarge individual experience.” Explain.
- 22.** What is a statistical enquiry? State the preliminary steps necessary for planning statistical enquiry.
(B.Com., CHU, BDU, BU)
- 23.** What is statistical enquiry? Describe the main stages in a statistical enquiry.
- 24.** Define a statistical unit. State in brief the precautions you would take in the selection of a statistical unit for conducting an enquiry.
- 25.** Describe the various steps to be considered while planning a statistical investigation.

(B.Com., CHU, BDU)

2

CHAPTER

COLLECTION OF DATA

2.1 INTRODUCTION

Collection of data is the process of enumeration of together with the proper recording of results. For any statistical enquiry or investigation whether it is related to business, management, economics or natural sciences, the basic issue is to collect the facts and figures, relating to a particular problem. The success of an enquiry is based upon the proper collection of data.

Statistical data may be collected from primary sources and secondary sources. Primary data are those data which are collected from the original source. It may be collected by individuals or institutions or group for research purpose or business purposes or policy decisions. Data, which have already collected, is called secondary data.

Secondary data are obtained from published or unpublished sources. This data can be obtained from websites, trade journals and government statistical department. For example, data collected by Indian Council of Social Science Research (ICSSR) about the literacy is called primary data and the secondary data used by a scholar for some study are secondary data.

2.2 METHODS OF COLLECTING PRIMARY DATA

Following are the important methods of collecting primary data:

1. Direct Interview Method
2. Indirect Interview Method
3. Information collected from local agencies
4. Questionnaire Method
5. Schedule sent through enumerators

2.2.1 Direct Interview Method

Under this method, data are collected by the investigator himself through interviews. The information or data so collected is original in nature. This method facilitates for personal contact with the respondents and thereby to know the reaction of the respondents immediately.

Merits

1. In this method, highly accurate original and reliable data are collected.
2. As the data are collected by one person, there is uniformity in the collection of data.
3. Due to personal presence of the investigator, there is flexibility in the enquiry and necessary adjustments can easily be done.
4. Qualitative Information can be determined very accurately.
5. Promptness is assured as the response of the person is more encouraging.

Demerits

1. Under this method, the data collected may be influenced by subjected attitude of investigators.
2. It is a costly and time-consuming method.
3. It leads to have personal bias.
4. This method is lengthy and complex.

2.2.2 Indirect Interview Method

It is otherwise called as indirect oral interview method. Under this method, the investigator collects data by contacting third party such as friends, relatives and neighbours.

When compared to the direct interview method, information could not be accurate.

Merits

1. Under this method, a wide area can be brought under investigation.
2. This method saves time, money and labour.
3. This method is free from bias of the investigator as well as of the informant.
4. This method is most suitable to explore the truths about the social evils like theft, murder, consumption of liquor etc.

Demerits

1. The information may be a wrong one since the data are collected from the relatives of the individual.
2. It is an expensive method, since the information about an individual is to be collected from many people.

16 Business Statistics

3. Inability of the investigator may bring wrong information.
4. Sometimes the information gathered about an individual from two or more people may contradict each other.

2.2.3 Information Collected from Local Agencies

Under this method, the investigator appoints local agents in different regions of the field of enquiry. This method is usually adopted by newspaper agencies who require periodical information in areas like sports, economic trends, share markets, law and order etc.

The government also uses this method, specially for the collection of information about prices, agricultural production etc.

Merits

1. This method is very cheap.
2. This method yield results easily and promptly.
3. Under this method, wide area can be covered.
4. The cost of collecting the information is very low.
5. The data can be collected within short period.

Demerits

1. This method would be a costly one when full-time agents are appointed in various places.
2. The information may not be uniform one. Information may contradict each other.
3. Under this method, more accurate information could not be collected.

2.2.4 Questionnaire Method

Collection of data through questionnaire is the most popular method for collecting primary data. A questionnaire is a list of questions pertaining to the enquiry. In this method, a well-designed questionnaire is mailed to the respondents with a request to fill it up and return the same within the specific time schedule. It should contain indirect questions rather than direct one and should be framed in a satisfying way.

Merits

1. It is the most economical method.
2. This method is more reliable and dependent results can be obtained.
3. Data can be collected within a very reasonable short period.
4. The data collected are unbiased since personal contact is not required.

Demerits

1. The chances of receiving questionnaires posted are very rare.
2. The respondents may delay in posting the filled-up questionnaires.
3. The respondents may give wrong information.
4. Personal motivation is not possible under this method.

Drafting the Questionnaire Framing the questionnaire is essential to collect the data. The success of an investigation depends on the construction of the questionnaire. Drafting questionnaire requires a great deal of skill and experience. It is very difficult to lay down any hard and fast rules to be followed in this construction.

General Principles of Framing the Questionnaire The following are the general principles which may be helpful in framing a questionnaire.

- (i) **Covering Letter** The person conducting the survey must introduce himself and state the objective of the survey. It covers matter such as purpose of the survey, assurance about the confidentiality.
- (ii) **Number of questions should be small** The number of questions should be kept to the minimum. If the number of questions are very large, the respondents may hesitate to answer all the questions. So the number of questions should be less.
- (iii) **Questions should be arrange logically** When the questions are arranged in a logical order, a natural and spontaneous reply to each is induced. Questions applying identification and description of the respondent so come first followed by major information questions.
- (iv) **Questions should be short and simple** The question should be short and simple to understand. Technical terms should be avoided. Respondents will give correct answer only when the questions are simple and in brief.
- (v) **Ambiguous questions should be avoided** Ambiguous questions should be avoided. It will not be possible from the respondents to take a question to mean different things.

For example, do you have plants?

Yes No

If yes, how many plants do you have?

- 1
- 2
- 3

18 Business Statistics

- (vi) **Personal questions should be avoided** Questions of personal nature should be avoided. For example, questions about income, sales tax paid may not be willingly answered in writing.
- (vii) **Questions should be designed for objective answers** Avoid questions of descriptive nature opinion. It is highly desirable that questions are so designed in objective type.
For example, How do you normally come to work place?
(i) by bus (ii) by taxi (iii) by scooter
(iv) on foot (v) any other mode
- (viii) **Yes or No question** Question should be answered in yes or no type. It is easy for the respondent to answer the question.
- (ix) **Questionnaire should look attractive** A questionnaire should be made to look as attractive as possible. The printing and the paper used and spaces should be left for answers depending upon the type of questions.
- (x) **Questions requiring calculation should be avoided** Questions involving calculations can be avoided.
For example, questions like ratio, percentage, proportions, average monthly income percentage should not be asked.
- (xi) **Cross checks** One or more cross checks should be incorporated into the questionnaire to determine whether the respondent is answering correctly.
- (xii) **Pre-testing questionnaire** The questionnaire should be pre-tested with a few selected respondents. It will help to find out the defects in the questionnaire.

Specimen Questionnaire

A study on Job Satisfaction in METTUR THERMAL POWER STATION

Respected Sir

Data collected are only for research purpose and it will be kept confidential, so please give your genuine answer.

Name :

Educational Qualification :

Age(in years) :

SSLC HSC

Department :

Diploma Graduation

Size of the family :

Post Graduation

Others

Experience :

Income (in Rs):

Tick the appropriate box by referring the choices given below:				
SA Strongly Agree	A Agree	N Neutral	DA Disagree	SDA Strongly Disagree
SA	A	N	DA	SDA
1. The flow of work related information in the organisation occurs relevant, timely and accurate.				
2. The rules and regulations of the organisation dominate your work.				
3. The policies and procedures and style of the management dominate the way in which you work.				
4. Each one in the organisation perceives one's role both at an individual and at a team level.				
5. Each and every employer in your organisation individual team and you are in the process of achieving the goal.				
6. The level of work freedom is high in your job.				
7. The organisation has got the capacity to absorb input and ideas from you and other employers, as well as reflecting capacity to deal with employees inputs.				
8. Management regard you as a valuable resource.				
9. At the work place, the quality and quantity of contact between managers/ supervisors and their people are good.				
10. Your work effort is generally acknowledged and appreciated by management.				
11. Working relationship provides a measure of the social cohesion between people across the organisation.				
12. More opportunity for person development is prevalent.				
13. Managerial support for equal opportunities and equitable rewards for effort is high.				

No. Factors	Highly satisfied	Satisfied	Neutral	Dis-satisfied	Highly Dis-satisfied
1. Ventilation					
2. Illumination					
3. Freedom from Noise					
4. Layout and Working space					
5. First Aid					
6. Drinking water facilities					
7. Canteen					
8. Rest Room					

No. Factors	Highly satisfied	Satisfied	Neutral	Dis-satisfied	Highly Dis-satisfied
9. Drainage					
10. Washing Facilities					
24. Grievances are usually related to					
(a) Promotion (b) Training (c) Work Burden					
(d) No Grievances (e) Others (Specify)					
25. Your commitment towards your organisation—a strong belief in and acceptance of the goal					
(a) Yes (b) No					
26. Your commitment towards your organisation a willingness to exert considerable effort on behalf of the organisation.					
(a) Yes (b) No					
27. Mention your satisfaction level for employee welfare service provided by your organisation.					
No. Employee Welfare Services	Highly satisfied	Satisfied	Neutral	Dis-satisfied	Highly Dis-satisfied
1. Housing					
2. Transports					
3. Children Education					
4. Medical Aid					
5. Recreation					
28. The intention to quit your organisation is					
(a) Very High (b) High (c) Moderate					
(d) Low (e) Very Low					
29. Your suggestions for further improvement					
1. _____					
2. _____					
3. _____					
4. _____					
5. _____					

2.2.5 Schedule Sent through Enumerators

This method of data collection is similar to that of the questionnaire. The schedule is also a proforma containing a set of questions. The difference between the questionnaire and the schedule is that the schedule is being filled in by the enumerators who are specially appointed for the purpose. But the questionnaire is filled by the respondent himself. The enumerators are otherwise called as agents or interviewers. They act as an agent of the investigators.

22 Business Statistics

Normally, this method is adopted by Research Departments, Federation of Indian Chambers of Commerce and Industries (FICCI), central government, state government etc.

Merits

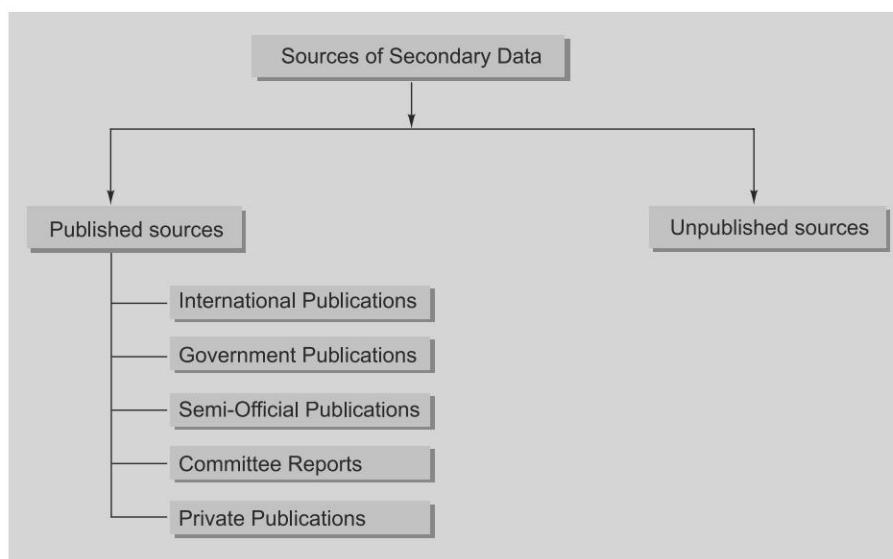
1. The information can be gathered from a wide area.
2. True and reliable data can be obtained as the personal contact between enumerators and informants is established.
3. The collected information is reasonably accurate because it is collected by trained and experienced enumerators.
4. Personal bias does not affect materially the results.

Demerits

1. This method is an expensive one since enumerators should be given remuneration for their work.
2. This method became ineffective when the enumerators are inefficient and not sincere.
3. The information collected may contradict each other.
4. If the schedule does not contain all the relevant questions, the information collected would be incomplete one.

2.3 SOURCES OF SECONDARY DATA

Sources of secondary data may be classified in the following two categories:



2.3.1 Published Sources

Governmental and non-governmental organisations publish statistics on different subjects. Such data are very reliable and useful for research purposes. The chief sources of published statistics are:

1. Publication of International Bodies Official publication of various international bodies like IMF, ILO, UNO, WTO, WHO etc., contain valuable international statistics.

2. Government Publications Various ministries and departments of the union and state governments publish regularly statistics on a number of subjects. The main publications are Labour Gazette, Indian Trade Journal Agricultural Statistics of India, Abstract of Agricultural Statistics, Statistical Abstract of India, Monthly Abstract of Statistics, Annual Survey of Industries etc.

3. Publications of Semi-Government Organisation Useful information is also published by various semi-government statistical organisations namely:

- (i) The Institute of Foreign Trade, New Delhi.
- (ii) The Institute of Economic Growth, New Delhi.
- (iii) Gokhale Institute of Politics and Economics, Pune.

Further, the statistical materials published by various other institutions like municipal and district boards, corporations, panchayat samitis provides fairly useful and reliable statistical information.

4. Publications of Research Institutions Individual research scholars, the different departments in the universities and various research organisations and institutes like Indian Statistical Institute (I.S.I) Kolkata and Delhi, Indian Council of Agricultural Research (I.C.A.R), New Delhi, Indian Agricultural Statistics Research Institute (I.A.S.R.I), New Delhi, National Council of Educational Research and Training (N.C.E.R.T) New Delhi, National Council of Applied Economic Research New Delhi, The Institute of Applied ManPower Research, New Delhi, Indian Standards Institute, New Delhi, publish the findings of their research programmes in the form of research papers or monographs and journals which are a continuous source of secondary data on the subjects concerned.

5. Publications of Commercial and Financial Institutions A number of private commercial and trade associations like Sugar Mills Association, Indian Cotton Mills Federation, Federation of Indian Chamber of Commerce and Industry (FICCI), Institute of Chartered Accountants of India, Institute of Cost and Works Accountant of India, Institute of Company Secretaries of India, trade unions, stock exchanges, bank bodies, cooperative societies etc., publish reports and Statistical material on current economic, business and other phenomena. Private concerns like Tata Consultancy Service also collect and publish statistics.

24 Business Statistics

6. Newspapers and Periodicals Newspapers like ‘The Economic Times’ ‘The Financial Express’ etc., and journals like Business Line, Commerce, Eastern Economist, Indian Labour Gazette, Journal of Industry and Trade, Monthly Statistics of Trade etc., collect and publish statistics of their subjects.

2.3.2 Unpublished Sources

It covers all those sources of secondary data where records are maintained by private agencies or business firms for their own use and are restrictedly available for use of general public. Data collected by research institutions, universities, trade associations, individuals etc., are also included in the category of unpublished sources of secondary data.

2.4 PRECAUTIONS TO BE TAKEN BEFORE USING SECONDARY DATA

The following are the precautions taken by the investigator before using the secondary data:

1. Reliability of Data Reliability of data can be tested through the source, methods of data collection, sample size, technique adopted for analysing data and the integrity of enumerators.

2. Adequacy of Data Data should be adequate for the purpose of investigation. If the data is not adequate, it will not satisfy the scope of investigation.

3. Suitability of Data Data collected from the secondary source should be suitable for the present study. For example, for a study about brand preference of typewriter are not suitable for the purpose of the study.

4. Accuracy of Data Data should be more accurate. It should not be a mere assumption. The analysis based on mere assumption will not give an accurate result.

SUMMARY

Sources of Collecting the Data

- Primary Sources
- Secondary Sources

Primary Data

If the data are collected from the original sources, they are called primary sources. They are obtained from the individual or institutions for which the data is related.

Secondary Data

Data collected not from the original source. They are obtained from published and/or unpublished sources.

Methods of Collecting Primary Data

- Direct Interview method
- Indirect Interview method
- Information collected from local agencies
- Questionnaire method
- Schedule sent through enumerators.

Qualities of a Good Questionnaire

- Covering letter
- Should be as small as possible
- Should be arranged in logical order
- Should be simple, clear and brief
- Proper words should be used
- Questions of personal nature should be avoided
- Questions of descriptive in nature should be avoided
- It should be an attractive one. Avoid the questions for which the correct answers will never be got
- Ambiguous questions should be avoided
- Questions should be of Yes or No type
- Cross-check questions should be framed
- Various type of question should be selected
- Pre-testing of questionnaire

Sources of Secondary Data

- Published sources
- Unpublished sources

Published Sources: Types

- Government publications
- Semi-Government publications
- Reports of committees and commissions
- Publication of Research Institutions
- Newspapers and Journals
- Publication of Trade associations and chamber of commerce
- International publications
- Websites

Precautions to be taken before using secondary data

- Reliability of data
- Adequacy of data
- Suitability of data
- Accuracy of data

EXERCISES

(a) Choose the best option.

1. Survey is a process of _____ of data.
(a) Classification (b) Collection (c) Tabulation
2. _____ object of a survey helps the statistician.
(a) Post-determined (b) Pre-determined
3. The data collected for the first time is called as _____.
(a) Primary data (b) Secondary data
4. The data that have already been collected is called as _____.
(a) Secondary data (b) Primary data
5. The enumerators must first understand the _____ of study.
(a) Period (b) Area (c) Purpose
6. In _____ questions, many answers are written.
(a) Simple alternative
(b) Multiple choice
(c) Specific information
7. _____ questions must be answered in Yes or No.
(a) Multiple choice
(b) Specific information
(c) Single alternative
8. _____ question is used when the investigator needs some specific information.
(a) Specific information
(b) Simple alternatives
(c) Multiple choice
9. _____ of data are suitable for the purpose of investigation.
(a) Adequacy (b) Suitability (c) Reliability
10. _____ of data can be tested.
(a) Suitability (b) Adequacy (c) Reliability

Answers

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

(b) Fill in the blanks.

1. A list of all units under study is known as _____.
2. _____ accuracy in enquiry is not possible.
3. The unit used for enumeration is called _____.
4. _____ and _____ data are two sources of information.
5. The enumerators work under a _____.
6. A constant check by the supervisor ensures _____.
7. The drafting of a schedule or a questionnaire is an _____.
8. The success of survey depends upon the work of the _____.
9. Most of the work is _____ by computers.
10. _____ and _____ are important processes in statistical investigation.
11. _____ the data are collected by the investigator personally.
12. _____ is the person who conducts the statistical enquiry.
13. The police department generally adopts _____ method.
14. _____ method local agents and correspondents will be appointed.
15. _____ method is appropriate in cases where informants are spread over a wide area.
16. _____ method is quite popularly used in practice.
17. The questionnaire is the _____ of communication.
18. The success of an investigation depends on the _____.

Answers

- | | |
|--|--------------------------------|
| 1. Frame | 2. Perfect |
| 3. Sample unit | 4. Primary, Secondary |
| 5. Supervisor | 6. Accuracy |
| 7. Art | 8. Enumerators |
| 9. Tabulated | 10. Classification, Tabulation |
| 11. Direct personal observation method | |
| 12. The investigator | 13. Indirect oral interview |
| 14. Information through agencies | 15. Mailed questionnaire |
| 16. Interview Schedule | 17. Media |
| 18. Construction of the questionnaire | |

(c) Theoretical Questions

1. (a) Define primary and secondary data. Explain their role in surveys with suitable examples.

(B.Com., MKU, BU, CHU)

- (b) What precautions would you take before using secondary data?
2. Explain the various methods that are used in the collection of primary data pointing out their merits and demerits.

(B.Com., MKU, MSU, BDU)

3. Describe the points you would consider in drafting the questionnaire.
4. Discuss the validity of the statement. “A secondary source is not as reliable as a primary source”.
5. What is a questionnaire? How does it differ from a blank form? What are the essential characteristics of a good questionnaire?

(B.Com., MKU, BU, BDU)

6. Explain in detail the method of collecting primary data through questionnaire. Distinguish between a schedule and a questionnaire.
7. What are the reasons for sampling? Under what conditions can cluster sampling be more efficient than other types of random sampling?

(B.Com., MKU, BU, CHU)

8. How will you conduct sample survey? What special points should be kept in mind in the selection of sample and collection of data?
9. Distinguish between primary data and secondary data.
10. What are the important aspects to be considered by an investigator before using secondary data?

(B.Com., MKU, BDU, MSU)

11. Distinguish between direct interview method and indirect interview method of collecting primary data.

(B.Com., MSU, BU, CHU)

12. State any two methods of collecting statistical data.
13. State the main sources from which secondary data are collected.
14. Explain the sources of secondary data.
15. Define direct interview method and indirect interview method.

3

CHAPTER

CLASSIFICATION AND TABULATION

3.1 INTRODUCTION

During every statistical investigation, the collected data, also known as raw data or ungrouped data, are always in an unorganised form and need to be organised and presented in meaningful form in order to facilitate further statistical analysis.

The first step in the analysis and interpretation of data is classification and tabulation. Classification means arranging the data into different groups on the basis of their similarities. The next step is tabulation which is concerned with the systematic arrangement and presentation of classified data.

3.2 DEFINITION

Classification is the process of arranging the collected data into classes and sub-classes according to their common characteristics. It can be defined as follows:

Classification is the process of arranging things (either actually or notionally) in the groups according to their resemblances and affinities and given expression to the unity of attributes that may subsist amongst a diversity of individuals.

— Prof. Cornor

Classification is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts.

— Secrist

3.3 CHIEF CHARACTERISTICS OF CLASSIFICATION

1. The basis of classification is unity in diversity.
2. Classification may be either real or imaginary.

30 Business Statistics

3. In the classification process, all the facts are classified into homogeneous groups.
4. Classification may be according to either similarities or dissimilarities.
5. It should be flexible to accommodate adjustments.

3.4 OBJECTIVES OF CLASSIFICATION

The chief objectives of classification are:

1. To present a mass of data in a condensed form.
2. To highlight the points of similarity and dissimilarity
3. To facilitate comparison, study relationships between several characteristics
4. To bring out a relationship
5. To prepare the basis for tabulation and analysis
6. To condense raw data into a form suitable for further statistical analysis
7. To eliminate unnecessary details

3.5 RULES OF CLASSIFICATION

Classification should adhere to the following rules:

3.5.1 Exhaustive

The system of classification must be exhaustive. There must be a class for each item of data. If classification is exhaustive enough, there will be no place for ambiguity.

3.5.2 Mutually Exclusive

Classes must not overlap that is, each item of data must find its place in one class only.

3.5.3 Suitability

Classification should be suitable for the object of enquiry.

3.5.4 Stability

Only one principle must be maintained throughout the analysis. Then only it will lead to have meaningful comparison.

3.5.5 Flexibility

A good classification should be flexible and capable of being adjusted according to changed situations and circumstances.

3.5.6 Homogeneity

Items included in one class should be homogeneous.

3.5.7 Arithmetical Accuracy

Items included in total and sub-totals of each class and sub-class must be the same.

3.5.8 Unambiguous

Classification should not lead to any ambiguity or confusion.

3.6 TYPES OF CLASSIFICATION

The collected data are classified on the basis of the purpose and objectives of the investigation or enquiry. Generally, the data can be classified on the basis of the following criteria:

1. Geographical Classification
2. Chronological Classification
3. Qualitative Classification
4. Quantitative Classification
5. Conditional Classification

3.6.1 Geographical Classification

It is also known as spatial classification. Here the data are classified on the basis of geographical or vocational differences such as state, cities, districts, zones or villages between various items of the data set. This is illustrated in the table below.

City	Bangalore	Chennai	Delhi	Kolkata	Mumbai
Population Density (Per sq.km)	178	205	423	685	654

3.6.2 Chronological Classification

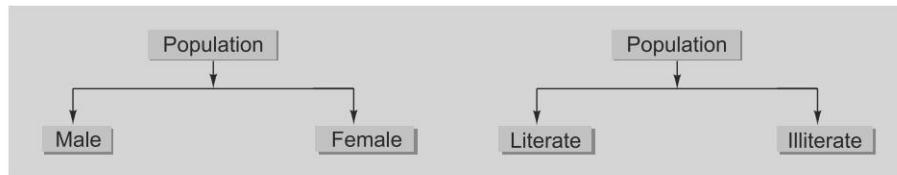
When data are classified on the basis of differences in time such as years, months, weeks, days, hours etc., the classification is known as chronological classification. The following is an example:

Year	1901	1911	1921	1931	1941	1951	1961	1971	1981	1991	2001
Population of India (in crores)	23.8	25	25.2	27.9	31.8	36.1	43.9	54.8	68.4	85.8	98.6

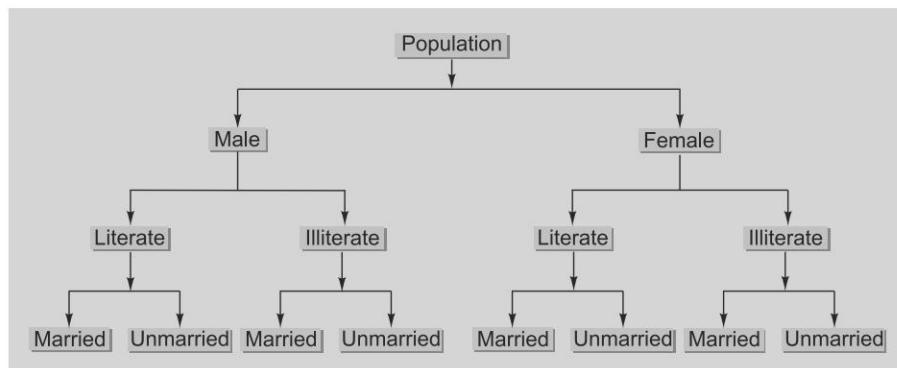
3.6.3 Qualitative Classification

When data are classified according to some qualitative phenomena like honesty, employment, intelligence, literacy, beauty, caste, etc., the classification is termed as qualitative or descriptive or by attributes. Here the data are classified according to the presence or absence of the attributes.

(a) Simple Classification When classification is done with respect to one attribute, two classes are formed. One possessing the attribute and the other not possessing the attribute. This type of classification is called simple or dichotomous classification.



(b) Manifold Classification Here, classification is done simultaneously with respect to two attributes for example, sex and literacy. The population is first classified with respect to ‘sex’ into ‘males’ and ‘females’. Each of these classes may further be classified into ‘literate’ and ‘illiterate’. This type of classification is called ‘manifold classification’.



3.6.4 Quantitative Classification

If data are classified on the basis of phenomenon which is capable of quantitative measurement like height, weight, income, expenditure, sales, profits etc., is termed as quantitative classification.

Variable Characteristics of data which is quantitative in nature and which can be measured in numerical terms is called a ‘variable’ or ‘variate’. Quantitative variables can be divided into the following two types:

(i) Discrete Variable It is the one whose values change by steps and can only assume an integral values depending upon the variable under study.

Discrete variable is one where the variates differ from each other by definite amounts. — **Boddington**

For example, number of employees in a company, number of children in a family.

(ii) Continuous Variable It is the one that can take any value within the given specified range of numbers. For example, the age of students, distance (in kms).

3.6.5 Conditional Classification

When the data are classified according to certain conditions, other than geographical or chronological, it is called a conditional classification.

3.7 STATISTICAL SERIES

Statistical series are prepared to present the collected and classified data in a properly arranged way.

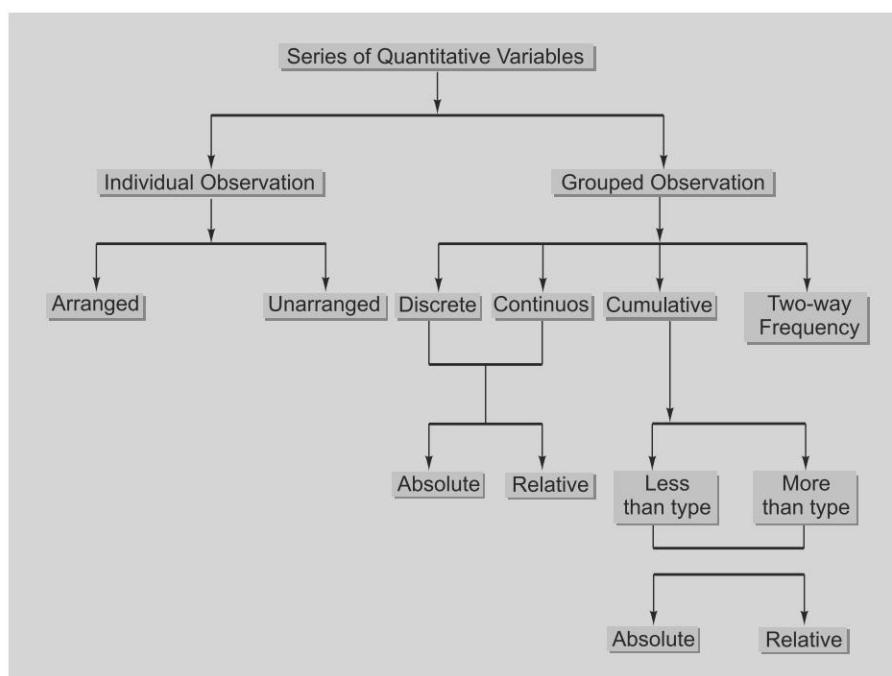
For example, if the data pertaining to the height of 10 students are put in a systematic way, it can be called a statistical series.

According to Secrist, “A series as used statically may be defined as things or attributes of things arranged according to some logical order”.

According to L.R. Conner, “If two variable quantities can be arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other, the result is said to form a statistical series”.

3.7.1 Types of Series

The following chart brings out the various types of series based on numerical or quantitative value.



3.7.2 Array

Raw data can be arranged in descending or ascending order of magnitude and is called an array.

3.8 FREQUENCY DISTRIBUTION

Frequency Distribution is simply a table in which the data are grouped into classes and the number of cases which fall in each class is recorded.

A classification according to the number possessing the same values of the variables.
—Erricker

Frequency Distribution can be divided into two types:

1. Univariate Frequency Distribution
 2. Bi-variate Frequency Distribution (Two-way Frequency Distribution)
- Again, univariate frequency distribution can be divided into three ways:
1. Individual Observation
 2. Discrete Frequency Distribution
 3. Continuous Frequency Distribution

3.8.1 Individual Observation

Individual observation is a series where items are listed singly after observation, as distinguished from listing them in groups. The following table is an example.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks	40	33	27	38	41	48	44	51	39	55

The data in the above form is called raw or disorganised data. The above presentation does not give any useful information. A better presentation of the above raw data would be to arrange in an ascending or descending order of magnitude which is called ‘arraying’ of the data.

3.8.2 Discrete (ungrouped) Frequency Distribution

In a discrete series, data are presented in a way that the exact measurement of the units are clearly indicated. There is definite difference between the variables of different groups of items. Each class is distinct and separate from the other classes.

Observed Values	Arraying	
	Ascending Order	Descending Order
40	27-Lowest	55-Highest
33	33	51
27	38	48
38	39	44
41	40	41
48	41	40
44	44	39
51	48	38
39	51	33
55	55-Highest	27-Lowest

Making a Frequency Table Data may be given in the form of individual observation. They are to be converted into discrete frequency distribution.

Steps

Form a table with three headings—‘Variable, Tally Marks, Frequency’.

- In the first column, place all possible values of the variable.
- In the second column, vertical bar (I) called Tally mark is put against the number.
- After a particular value has occurred four times, for the fifth occurrence put a cross tally mark (||||) cutting the first four diagonally and this gives us a block of 5.
- For the sixth item, put another tally mark leaving some space.
- Leaving space after a block of five, facilitates easy and correct counting.
- Finally, count the number of bars corresponding to each value of the variable and place it in the column entitled Frequency.

Illustration 3.1

Consider the marks scored by 25 students:

7	5	3	5	6	4	6	2	8
1	7	2	8	5	5	6	3	
4	5	9	6	10	5	5	10	

We are unable to understand the significance of marks scored by the 15 students as it is in row form. We have to form discrete series out of the above data. First, we note down the lowest and highest values. In the first column, we place all possible values of the variables. In the second column, a vertical bar (I), called tally mark is put against the number (variable) whenever it occurs. After a particular value has occurred four times, for the fifth occurrence, we put a cross tally mark (||||) or (||||), cutting the first four tally marks and this gives us a block of five.

Marks	Tally-sheet	Number of students (Frequency)
1	I	1
2		2
3		2
4		2
5		7
6		4
7		2
8		2
9	I	1
10		2
Total		25

Marks	Frequency
1	1
2	2
3	2
4	2
5	7
6	4
7	2
8	2
9	1
10	2
Total	25

3.8.3 Continuous or Grouped Frequency Distribution

Continuous series is one where measurement are only approximations and are expressed in class intervals, that is, within certain limits.

“The variable which can take any intermediate value between the smallest and longest value in the distribution”.

In a continuous frequency distribution, the class intervals theoretically continue from the beginning to the end of the frequency distribution without break. The continuous frequency distribution consists of two limits—upper limit and lower limit of each class interval while the discrete frequency distribution will possess only one list of classification of values.

Continuous Series A collection of items which cannot be exactly measured, but placed within certain limit is called continuous series.

Class Limits The class limits are the smallest and highest value in the class. Class limit is also known as class boundaries.

Class Intervals The difference between the lower limit and the upper limit of the class is known as class interval. The formula for class interval is

$$i = \frac{L - S}{K}$$

where

L = Largest item

S = Smallest item

K = The number of class

There are two methods of class interval (a) Exclusive Method (b) Inclusive Method.

(a) Exclusive Method (Overlapping) Under this method, the upper limit of one class is the lower limit of the next class. For example,

Marks	10–20	20–30	30–40	Total
No. of Students	15	20	10	45

This method ensures continuity of data. A student whose mark is between 10 and 19.9 would be included in the 10–20 class. A student whose mark is 20 would be included in the class 20–30.

(b) Inclusive Method (Non-Overlapping) In this method, the upper limit of one class is included in that class itself.

Marks	10–19	20–29	30–39	40–49	Total
No. of Students	17	15	12	10	54

Here a student getting 29 marks is included in 20–29 class interval. Here the confusion is avoided because the upper limit of class is not the lower limit of the next class.

Class Frequency The number of observations falling within a class interval is called its class frequency or frequency.

EXAMPLE

The frequency of the class 10–20 is 8; this means that there are 8 students having the marks between 10 and 20. If we add the frequency of all individual classes, we obtain the total frequency. Thus the total frequency of all the classes is 50, which means that in all, there are 50 students whose marks are studied.

Magnitude of Class Intervals The magnitude of class intervals depends on the range of the data and the number of classes. The range is the difference between the largest and smallest observation in the given data.

EXAMPLE

If the range of the marks of a group of students is 50 and if we desire to have 10 classes then the magnitude of each class interval would be $50 \div 10 = 5$ marks.

Formula The following is the formula for determining the number of classes:

$$K = 1 + 3.322 \log N$$

K = number of classes

$\log N$ = Logarithm of the total number of observations

If the number of observation is 100, number of classes would be

$$1 + 3.322 \times \log 100$$

or

$$1 + 3.322 \times 2$$

$$1 + 6.644 = 7.644 = 8$$

(Fractions are always rounded up)

Struges has also given a formula for determining the magnitude of class interval.

$$i = \frac{\text{Range}}{1 + 3.322} \log N$$

i = Magnitude of the class interval

$\log N$ = Logarithm of the total number of observations

Class Mid-point or Mid-value The central value of the class interval is called mid-point. It lies half way between the lower limit and the upper limit of the class interval.

$$\text{Mid-point of a class} = \frac{\text{upper limit (U.L)} + \text{lower limit (L.L)} \text{ of the class}}{2}$$

or

$$K \frac{\text{U.L} + \text{L.L}}{2}$$

Illustration 3.2

Marks scored by 15 students are given below.

21	35	28	27	33	28	13	22	40	21	33
27	28	35	10							

- (a) Arrange the marks in ascending order
- (b) Arrange the marks in descending order
- (c) Convert the marks into a continuous series of a class-interval of 10.

Solutions

(a) Marks arranged in ascending order			(b) Marks arranged in descending order		
10	27	33	40	28	22
13	27	33	35	28	21
21	28	35	35	28	21
21	28	35	33	27	13
22	28	40	33	27	10

(c) Formation of continuous series

Marks	Tally Marks	Frequency
10–20		2
20–30	, III	8
30–40		5
Total		15

Continuous class frequency

Marks	Frequency
10–20	2
20–30	8
30–40	5
Total	15

Illustration 3.3

Using sturges rule $n = 1 + 3.322 \log N$, classify in equal intervals, the following data of hours spent in working by 20 workers for period of a month in a certain factory.

114	71	161	175
30	149	94	133
104	157	146	156
93	151	113	204
203	124	87	144

Solutions

$$\begin{aligned} h &= 1 + (3.322 \log N) \\ &= 1 + (3.322 \times 1.6990) \\ &= 1 + 5.6441 = 6.6441 = 7 \end{aligned}$$

$$i = \frac{\text{Range}}{1 + 3.322 \log N} = \frac{204 - 30}{7} = 24.85 \text{ or } 25$$

Class	Tally Marks	Frequency
30–55		1
55–80		1
80–105		4
105–130		3
130–155		5
155–180		4
180–205		2
Total		20

Cumulative Frequency Distribution (Cf) Cumulative frequencies are derived by the cumulation of the frequencies of successive values. It represents the total frequency of all previous variables including the variable or the class.

The cumulation is started from the lowest size to the highest size.

Less than cumulative frequency is obtained by adding successively the frequencies of all the previous variable including the variable against which it is written. More than cumulative frequency distribution is obtained by finding the cumulation total of frequencies starting from the highest to the lowest variable.

Illustration 3.4

Marks	Frequency	Cumulative frequency less than	Cumulative Frequency more than
20–30	3	3	100
30–40	8	(3 + 8)11	(100 – 3)

Contd.

Marks	Frequency	Cumulative frequency less than	Cumulative Frequency more than
40–50	10	(11 + 10)21	(97–8)
50–60	5	(21 + 5)26	(89–10)
60–70	14	(26 + 14)40	(79–5)
70–80	20	(40 + 20)60	(74–14)
80–90	28	(60 + 28)88	(60–20)
90–100	12	(84 + 12)100	(40–28)

The above less than cumulative frequency distribution and more than cumulative frequency distribution can also be expressed in the following forms:

Less than c.f.d		More than c.f.d	
End values (upper limit)	c.f (less than)	End values (lower limit)	c.f (more than)
Less than 30	3	More than 20	100
Less than 40	11	More than 30	97
Less than 50	21	More than 40	89
Less than 60	26	More than 50	79
Less than 70	40	More than 60	74
Less than 80	60	More than 70	60
Less than 90	84	More than 80	40
Less than 100	100	More than 90	12

From the above table, it is easy and quick to find out the number of students who have scored marks less than or more than a particular mark; for example, if we want to know the number of students who scored marks more than 50, it can be known by looking at the cumulative frequency table. In the above illustration, the number of students who scored marks more than 50 is 79 and less than 50 is 21.

Less than cumulative frequency corresponds to the upper limit of the class and more than cumulative frequency corresponds to the lower limit.

Illustration 3.5

- (a) Make a frequency distribution with intervals of 10 from the following data.
- (b) Also prepare less than cumulative frequency distribution.
- (c) And prepare more than cumulative frequency distribution.

20 36 63 84 11 72 52 66 85 74 21 43
 57 95 15 45 88 72 74 54

Solutions

(a) Frequency Distribution

Marks	Tally marks	Frequency
10–20		2
20–30		2

Contd.

Marks	Tally marks	Frequency
30–40		1
40–50		2
50–60		3
60–70		2
70–80		4
80–90		3
90–100		1
Total		20

(b) Less than c.f.d

Marks	c.f	Marks	c.f
Less than 20	2	More than 10	20
Less than 30	(2 + 2)4	More than 20	(20–2)18
Less than 40	(4 + 1)5	More than 30	(18–2)16
Less than 50	(5 + 2)7	More than 40	(16–1)15
Less than 60	(7 + 3)10	More than 50	(15–2)13
Less than 70	(10 + 2)12	More than 60	(13–3)10
Less than 80	(12 + 4)16	More than 70	(10–2)8
Less than 90	(16 + 3)19	More than 80	(8–4)4
Less than 100	(19 + 1)20	More than 90	(4–3)1

(c) More than c.f.d

Illustration 3.6

Present the following data of the percentage marks of 20 students in the form of frequency table with ten classes of equal width, one class being 0–9.

06	17	28	63	70	60	80	75	92	11	09	32
43	58	73	55	23	82	33	88				

Solutions

Frequency distribution of the marks of 20 students.

Marks	Tally Marks	Frequency
0–9		2
10–19		2
20–29		2
30–39		2
40–49		1
50–59		2
60–69		2
70–79		3
80–89		3
90–99		1
Total		20

42 Business Statistics

Two-way Frequency Distribution (Bivariate) A frequency table where two variables have been measured in the same set of items through cross classification is known as bivariate frequency distribution or two-way frequency distribution.

For example, marks obtained by students on two subjects, ages of husbands and wives, weights and heights of students etc.

Illustration 3.7

The following data represent the marks in Statistics (x) and Commerce (y) of 25 students. Prepare a bivariate table from the following data.

x	25	18	20	8	30	35	17	26	42	28
y	46	32	30	20	50	52	8	42	64	50
x	30	28	25	25	60	7	32	36	25	25
y	50	46	42	38	70	8	48	55	38	40
x	17	22	31	36	48					
y	28	36	58	64	78					

Solutions**Steps for construction**

1. Determine the class interval of each of the variables.
2. Write one of the variables on the left hand side of the table and the other at the top.
3. The first student gets 25 in Statistics and 46 in Commerce. A tally mark has to be put in the cell where the column showing 25–35 marks in Statistics intersects the row showing 45–55 marks in Commerce.
4. Repeat the procedure for all the 25 students.
5. Total the tallies at the bottom and to the right side.
6. Totals at the right at the extreme column are for Commerce and those at the bottom row are for Statistics.

Bivariate Frequency Distribution

$x \rightarrow$	5–15	15–25	25–35	35–45	45–55	55–65	Total
$y \downarrow$							
5–15							2
15–25							1
25–35							3
35–45							6
45–55							7
55–65							4
65–75							1
75–85							1
Total	2	5	12	4	1	1	25

Illustration 3.8

15 pairs of values of two variables, x and y are given below. Form a two-way table.

x	16	22	35	43	29	24	14
y	150	244	298	516	387	440	120
x	39	41	19	28	38	39	29
y	481	453	247	415	387	451	512

Take class intervals of x as 10–20, 20–30 etc. and that of y as 100–200, 200–300 etc.

Solutions

Formation of two-way table:

$x \rightarrow$	10–20	20–30	30–40	40–50	Total
$y \downarrow$					
100–200					2
200–300					3
300–400					2
400–500					6
500–600					2
Total	3	5	4	3	15

3.9 TABULATION OF DATA

3.9.1 Meaning

Tabulation is another way of systematic summarisation and presentation of information contained in the given data in rows and columns in accordance with some salient features or characteristics. Such presentation facilitates comparison by bringing related information close to each other and helps in further statistical analysis and interpretation.

3.9.2 Definition

In the words of W.A. Spurr and C.P. Bonini, “A statistical table is a classification of related numerical facts in vertical columns and horizontal rows”.

According to A.M. Tuttle, “A statistical table is the logical listing of related quantitative data in vertical columns and horizontal rows of numbers with sufficient explanatory and qualifying words, phrases and statements in the form of titles, heading and notes to make clear the full meaning of data and their origin”.

D. Gregory and H. Ward say, “Tabulation is the process of condensing classified data in the form of a table so that it may be more easily understood and so that any comparisons involved may be more readily made”.

3.9.3 Objectives of Tabulation

Tabulation is a process which helps in understanding complex numerical facts.

According to Harry Jerome, “A good statistical table is not a mere careless grouping of columns and rows of figures; it is a triumph of ingenuity and technique, a master-piece of economy of space combined with a maximum of clearly presented information.” To prepare a first class table, one must have a clear idea of the facts to be presented, the contrasts to be stressed, the points upon which emphasis is to be placed and lastly, a familiarity with the technique of preparation:

The main objectives of tabulation are:

1. To present collected data in an orderly manner
2. To simplify the complex data
3. To economise space
4. To depict pattern in figures
5. To facilitate comparison
6. To facilitate statistical analysis
7. To help reference
8. To detect errors and omission in the data
9. To identify trend and tendencies of the given data

3.9.4 Parts of Tabulation

The following parts must be presented in all tables.

- | | |
|-----------------|----------------------|
| 1. Table Number | 2. Title |
| 3. Head Note | 4. Caption |
| 5. Stubs | 6. Body of the Table |
| 7. Foot Note | 8. Source Note |

1. Table Number A table should be numbered for identification, specially when there are a large number of tables in a study.

2. Title of the Table Every table should have a brief, clear, self-explanatory and carefully worded title. It must be written on the top of the table.

3. Head Note It is a statement given below the table and enclosed in brackets.

4. Caption Every column in a table is to indicate what it represents. A table may have column heads or even column sub-heads. The caption designates the data found in the column of the table.

5. Stubs Each row of the table must be given a heading. The designations of the rows are called ‘stubs’ or ‘stub items’ and the complete row is known as ‘stub row’.

6. Body of the Table It contains the numerical information. It is the most important part of the table. The arrangement in the body is generally from left to right in rows and from top to bottom in columns.

7. Foot Note If any explanation or elaboration regarding any item is necessary, foot notes should be given.

8. Source Note It refers to the source from where information has been taken. It is useful to the reader to check the figures and gather additional information.

STRUCTURE OF A TABLE		
	Number	Title
Sub-Heading	(Head-Note if any)	
	Caption	Total
	Col. Heading	
	BODY	
Stub entries		
Total		
Source: ...		

3.9.5 Rules for Tabulation

It may be noted that no hard and fast rule can be laid down for preparing a statistical table. In the words of A.L. Bowley, "In the tabulation of the data, common sense is the chief requisite and experience is the chief teacher".

The following rules will serve as a general guide in the construction of tables:

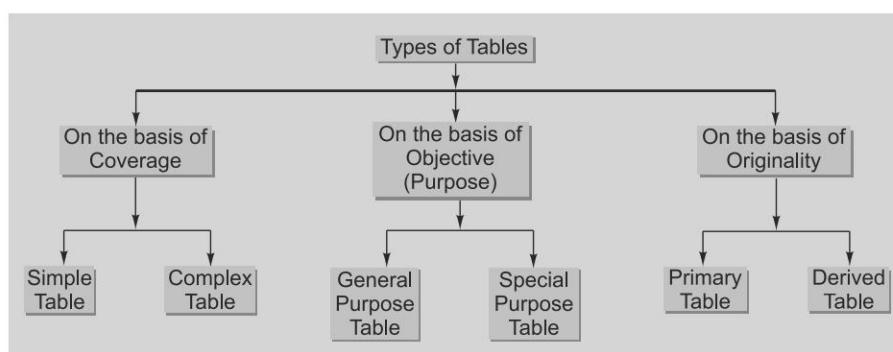
1. The table should not be overcrowded with details.
2. The number of main headings should be few.
3. Figures may be rounded to avoid unnecessary details in the table but a footnote to this effect must be added.
4. Table should be adjusted to the space available.
5. Columns should be carefully planned.
6. Size of the table should be such that all its contents are visible at a glance.
7. Important figures should be indicated in bold letters.
8. If certain data are not available or have been estimated, this fact must be given as a foot note.
9. Data which can be classified properly may be shown under the head 'miscellaneous'.
10. Totals of the columns should be put at the bottom of the table.
11. The items in the table should be logically and systematically classified.
12. Use of abbreviations should be avoided.

46 Business Statistics

13. Do not use Ditto marks that may be mistaken.
14. If it is a big table, it will lose its simplicity and understandability and in such a case, break it into two or three tables.

3.9.6 Types of Tables

The classification of tables depends on various aspects like objectives or purpose of investigation, nature of enquiry, extent of data, coverage and so on. The various types of tables are displayed in the following diagrammatic scheme.



I. On the Basis of Coverage

(a) Simple Table Simple table is also known as one-way table or table of first order. The data are presented according to one characteristic of data only.

Distribution of Marks				
Class Marks	20–30	30–40	40–50	Total
No. of Students	10	18	22	50

(b) Complex Table It is also known as manifold table in which data are presented according to two or more characteristics simultaneously. The complex tables are two-way or three-way tables according to which two or three characteristics are presented simultaneously.

(c) Two-way Table If the caption or stub is classified in two characteristics and if it gives information of two interrelated questions, then such a table is called two-way table.

EXAMPLE

Distribution of Marks (Girls and Boys)

Class Marks	Number of Students		
	Boys	Girls	Total
20–30	6	4	10
30–40	8	10	18
40–50	10	12	22
Total	24	26	50

(d) *Three-way Table* In this type of table, three characteristics are shown. It gives information regarding three interrelated characteristics of a phenomenon.

EXAMPLE

Distribution of population by Age, Sex and Literacy

Age Group (Years)	Males			Females			Total		
	Literate	Illiterate	Total	Literate	Illiterate	Total	Literate	Illiterate	Total
0–18									
19–25									
25–35									
35–45									

(e) *Manifold or Higher Order Table* Number of students in M.K. University (according to faculty, age, sex and residence)

Faculty age group (year)	Male			Female			Total		
	Host eller	Days scholar	Total	Host eller	Days scholar	Total			
Commerce									
20–25									
25–30									
Above 30									
Arts									
20–25									
25–30									
Above 30									
Science									
20–25									
25–30									
Above 30									

2. On the basis of Objective (Purpose)

(a) *General Purpose Table* It is also known as informative table and provides information for general use and usually in chronological order.

Government agencies prepare this types of tables. These are used by research workers and statisticians. These are placed in the appendix of a report for reference.

(b) *Special Purpose Table* It is also called a summary table or text table or analytical table or derivative table or derived table. It presents the data relating to a particular or a special purpose. Ratios, percentages etc., are used to facilitate comparison.

3. On the Basis of Originality

(a) *Primary Table* In this table, the statistical forms are expressed in original. It contains actual and absolute figures.

48 Business Statistics

(b) Derived Table In a derived table, figures and results are derived from primary data. It presents totals, percentages, ratios, averages, dispersion etc.

Illustration 3.9

Draw a blank table to show the candidates sex-wise, appearing for college, first year, second year and third year examinations of a college in the faculties of science, computer and arts in a certain year.

Table showing the distribution of candidates appearing in various college examinations.

Faculty	Boys			Girls			Total					
	College	I	II	III	College	I	II	III	College	I	II	III
Science												
Computer												
Arts												
Total												

Illustration 3.10

Present the following information in a suitable tabular form, supplying the figures not directly given.

In 2005, out of 3000 workers in a factory, 2,100 were members of a trade union. The number of woman workers employed was 500, out of which 400 did not belong to any trade union.

In 2006, the number of union workers was 4,000 of which 3,500 were men. The number of non-union workers was 1480, among whom 1175 were women.

Comparative study of the membership of trade union in a factory in 2005 and 2006.

Year	2005			2006		
	Males	Females	Total	Males	Females	Total
Trade Union	2000	100	2100	3500	500	4000
Non-members	500	400	900	305	1175	1480
Total	2500	500	3000	3805	1675	5480

Illustration 3.11

Present the following information in a suitable form supplying the figures not directly given.

Figures are not given in the problem. These are obtained by minor calculation of additions and subtractions.

In 2007, out of a total of 5000 workers in a company, 4300 were members of trade union. The number of women workers employed was 400 out of which 300 did not belong to any union.

In 2006, the number of workers in the union was 4,450 of which 4,200 were men. The number of non-union workers was 660 of which 230 were women.

Solutions

Table showing the genderwise distribution of membership of trade union members.

Year Trade Union	2006			2007		
	Males	Females	Total	Males	Females	Total
Members	4200	250	4450	4200	100	4300
Non-members	430	230	660	400	300	700
Total	4630	480	5110	4600	400	5000

Illustration 3.12

Present in tabular form with suitable captions the information contained in the following.

In 2005, out of total of 2650 workers of a factory, 2100 workers were members of a trade union. The number of women employed was 300 of which 195 did not belong to a trade union.

In 2006, the number of union workers increased to 2,610 of which 2,310 were men. On the other hand, the number of non-union workers fell down to 1,306 of which 1,160 were men.

In 2007, there were on the pay-rolls of the factory 2600 employees who belonged to a trade union and 150 who did not of all the employees in 2007, 500 were women of whom only 12 did not belong to a trade union.

Solutions

Comparative study of the membership of trade union in factory in 2005, 2006 and 2007.

Year	Members of trade unions			Non-members of trade unions			Total
	Men	Women	Total	Men	Women	Total	
2005	1995	105	2100	355	195	550	2650
2006	2310	300	2610	1160	146	1306	3916
2007	2112	488	2600	138	12	150	2750

Illustration 3.13

In a newspaper account describing the incidence of influenza among cancer persons living in the same family the following passage appeared:

"Exactly 4/15th of 6,000 inhabitants showed signs of cancer and no less than 600 among them had an attack of influenza but among them only 100 lived in infected houses. In contrast with this, 1/5th of the cancer persons who did not have influenza were still exposed to infection. Altogether, 2000 were attacked by influenza had 3700 were exposed to the risk of infection but the number who

50 Business Statistics

having influenza but not cancer living in houses where no other case of influenza occurred was only 400".

Redraft the information in a concise and elegant tabular form.

Solutions

Table showing the incidence of influenza among cancer persons living the same family.

	Attacked by Influenza		Not Attacked by Influenza		Total	
	Living in infected houses	Living in not infected houses	Total	Living in infected houses	Living in not infected houses	Total
Attacked by cancer	100	500	600	200	800	1000
Not attacked by cancer	1000	400	1400	2400	600	3000
Total	1100	900	2000	2600	1400	4000
						6000

Illustration 3.14

Present the following information in a suitable tabular form, supplying the figures not directly given.

In 2006, out of 8000 workers in a factory, 3800 were members of a trade union. The number of women workers employed was 1500, out of which only 1100 belong to union.

In 2007, the number of union workers were 7500 of which 4000 were men. The number of union workers were 3200, out of which 2500 were women.

Solutions

Table showing membership of trade union by sex.

Year	Trade Union	Males	Females	Total
2006	Members	2700	1100	3800
	Non-members	3800	400	4200
	Total	6500	1500	8000
2007	Members	4000	3500	7500
	Non-members	700	2500	3200
	Total	4700	6000	10700

Illustration 3.15

Present in a tabular form with suitable captions the information given below :

In 1995, out of total of 7750 workers of a factory, 6200 were members of trade union. The number of women employed was 400 of which 200 did not belong to a trade union.

In 2000, the number of union workers increased to 7580 of which 7290 were men on the other hand, the number of non-union workers fell down to 208 out of which 180 were men.

In 2005, there were 7800 employees who belonged to a trade union and 6050 did not. Of all the employees in 2005, 400 were women, of whom only 108 did not belong to a trade union.

Solutions

Table showing distribution of union and non-union members according to sex, for the years 1995, 2000 and 2005.

Year Sex	1995			2000			2005		
	M	F	T	M	F	T	M	F	T
Union members	6000	200	6200	7290	290	7580	7500	292	7800
Non-union members									
Non-union members	1350	200	1550	180	28	208	5942	108	6050
Total	7350	400	7750	7470	318	7788	13450	400	13850

SUMMARY**Classification**

Process of arranging the collected data into classes and sub-classes according to their common characteristics.

Objectives of Classification

- Simple form of presentation of data
- Facilitates comparison
- Relationship of data can easily be understood
- Facilitates for easy understanding
- Forms a basis for tabulation.

Characteristics of Classification

- | | |
|--|--|
| <ul style="list-style-type: none"> • Exhaustive • Flexibility • Suitability | <ul style="list-style-type: none"> • Mutually exclusive • Homogeneity • Arithmetical access |
|--|--|

Methods of Classification

- Geographical Classification
- Qualitative Classification
- Chronological Classification
- Quantitative Classification

Geographical Classification

The data are classified on the basis of states, regions, zones, cities, villages.

Chronological Classification

On the basis of time like years, months, week, days, hours.

Qualitative Classification

On the basis of some attributes or characteristics.

Quantitative Classification

On the basis of characteristics which can be quantitatively measured.

Conditional Classification

On the basis of some conditions.

Tabulation

Process of systematic and scientific presentation of data in a compact form for further analysis.

Rules for Tabulation

- Table must be arranged with number.
- It must have a title.
- The details about the information in each row are called stubs. It should be clear and brief.
- The details about the information in each column are called captions. It should be clear and brief.
- Body of the table must contain only relevant information.
- Proper space should be provided in between various data.
- Averages and total should be properties given, if required.
- Further explanation should be given after the table if needed.
- Special markings can be used in respective places.
- Source is an important requirement for the table.
- It indicates place or person from whom the information is collected.

Requisites for a good Tabulation

- It must be simple.
- Columns and rows should not be too narrow or too wide.
- Brief description should be given.
- It should clearly show the unit of measurement.
- It should be an optimum one.

Types of Tables

- Simple and Complex Table
- General Purpose and Special Purpose Table.

EXERCISES**(a) Choose the best option.**

1. Time series are also called _____.
(a) Qualitative (b) Chronological (c) Quantitative
2. _____ classification is the universe classified.
(a) Manifold (b) Qualitative (c) Quantitative
3. Geographical classification is otherwise called as _____ classification.
(a) Spatial (b) Manifold (c) Qualitative
4. _____ method is the upper limit of one class is the interval is the lower limit of the next class.
(a) Inclusive (b) Exclusive
5. _____ method the upper limit of one class is included in that class itself.
(a) Inclusive (b) Exclusive

Answers

1. b 2. a 3. a 4. b 5. a

(b) Fill in the blanks.

1. The collected data in any statistical investigation are known as _____.
2. Classification is the process of _____ data.
3. Classification should be _____.
4. _____ are expressed in class intervals.
5. _____ are the smallest and the largest in the class.
6. _____ depends on the range of the data and the number of classes.
7. _____ is the difference between the largest and smallest observation in the given data.
8. Central value of the class interval is called _____.
9. _____ is systematic presentation of numerical data.
10. A goods statistical table is an _____.

Answers

- | | |
|-----------------|---------------------------------|
| 1. Raw data | 2. Arranging data |
| 3. Flexible | 4. Continuous series |
| 5. Class limits | 6. Magnitude of class intervals |
| 7. Range | 8. Mid-point |
| 9. Tabulation | 10. Art |

(c) Theoretical Questions

1. Define classification.
2. What are the objectives of classification?
3. Explain the important methods of classification.
4. State the principle underlying classification of data.
5. What is chronological classification of data? Give examples.
6. What is meant by discrete series?
7. What do you understand by the term continuous series?
8. What is class interval?
9. Explain the difference between frequency distribution and cumulative frequency distribution.
10. “In the case of exclusive series true class-limit are the same as class limit”—Comment.
11. Define Tabulation.
12. What are the requisites of a good table?
13. What do you mean by tabulation? State the different types of tables.
14. Distinguish between classification and tabulation.

(d) Practical Problems

15. Construct a frequency table with equal class intervals number following data on the monthly wages (in rupees) of 28 labourers working in a factory, taking one of the class intervals as 210–230 (230 included):
220, 268, 258, 242, 210, 268, 272, 242, 311, 290, 300, 320, 319, 304,
302, 318, 306, 292, 254, 278, 210, 240, 280, 310, 306, 215, 256, 236.
16. The heights (in cm) of 30 students of a class are given below.
155, 158, 154, 158, 160, 148, 149, 150, 153, 159, 161, 148, 157, 153, 157, 162, 159, 151, 154, 156, 152, 156, 160, 152, 147, 155, 163, 155, 157, 153
Prepare a frequency distribution table with 160–164 as one of the class intervals.
17. The monthly wages of 30 workers in a factory are given below.
830, 835, 890, 810, 835, 836, 869, 845, 898, 890, 820, 860, 832, 833, 855, 845, 804, 808, 812, 840, 885, 835, 836, 878, 840, 868, 890, 806, 840, 890
Prepare a table.
18. Following are the ages of 360 patients getting medical treatment in a hospital on a day.

Age (in years) : 10–20 20–30 30–40 40–50 50–60 60–70

No. of patients : 90 50 60 80 50 30

Prepare a cumulative frequency table. (B.Com, MKU, CHU)

19. The marks scored by 55 students in a test are given below:

Marks : 0–5 5–10 10–15 15–20 20–25 25–30 30–35

No. of

Students: 2 6 13 17 11 4 2

Prepare a cumulative frequency table.

20. Here are given the following marks secured by 25 students in the examination.

23, 28, 30, 32, 35, 36, 40, 41, 43, 44, 45, 45, 48, 49, 52, 53, 54, 56, 56, 58, 61, 62, 65, 68, 55

(a) Arrange the above data on frequency distribution taking class intervals 20 – 29, 30 – 39, 40 – 49, 50 – 59, 60 – 69.

21. From the following continuous series, prepare the (i) Less than (ii) More than cumulative series

0–10 8

10–29 12

20–30 30

30–40 25

40–50 18

50–60 17 (B.Com, MKU, CHU)

22. Following is the distribution of marks of 70 students in a test:

Marks (Less than) 10 20 30 40 50

Number of students 3 8 17 20 22

From the above data, form a frequency table. Find the number of students securing more than.

23. The frequency distribution of the marks obtain by 100 students of a class is given below:

Consumption No. of Factories Consumption No. of Factories

(in kW) **(in kW)**

20 – 30 6 50 – 60 22

30 – 40 18 60 – 70 17

40 – 50 25 70 – 80 12

Prepare a cumulative frequency table. (B.Com, MKU, CHU)

24. 50 students of a class obtained the following marks (out of 100) in the Statistics paper of the B.Com examination

21 32 32 51 50 62 65 75 85 83

40 37 30 42 44 44 57 53 54 75

Contd.

56 Business Statistics

73	96	65	66	66	66	48	45	55	55
51	59	59	64	58	58	63	63	58	56
74	77	60	56	61	61	67	65	50	51

Form a frequency distribution taking the lowest class interval as 20–30. **(B.Com, MSU, BU, MKU, CHU)**

- 25.** Values of two variables X and Y are given below. Form a two-way frequency table showing the relationship between the two. Take class intervals of X as 10 to 20, 20 to 30 etc., that of Y as 100 to 200, 200 to 300 etc.

X : 12, 34, 33, 22, 44, 37, 26, 36, 55, 48

Y : 140, 266, 360, 470, 470, 380, 480, 315, 420, 390

X : 27, 52, 41

Y : 440, 312, 330

X : 37, 21, 51, 27, 42, 43, 52, 57, 44, 48

Y : 390, 590, 250, 550, 360, 570, 290, 416, 380, 492

X : 48, 69

Y : 370, 590

Also obtain:

1. The marginal frequency distributions of X and Y
2. The conditional distribution of X when Y lies in 300–400 and conditional distribution of Y when X lies in 40–50.

4

CHAPTER

DIAGRAMMATIC AND GRAPHICAL REPRESENTATION

4.1 INTRODUCTION

Classification refers to grouping of data into homogeneous class and categories. Tabulation is the process of presenting the classified data in tables. Classification and tabulation are applied in order to make the collected data understandable. Many figures may be uninteresting and even confusing. So, a better way of representing data is by diagram and graph.

4.2 DIAGRAM

A diagram is a visual form for presentation of statistical data. Diagrams refer to the various types of devices such as bars, circles, maps, pictorials, cartograms. These devices can take many attractive forms.

4.3 RULES FOR PREPARING DIAGRAMS

The important rules for the preparation of diagrams are given below.

4.3.1 Heading

Every diagram must have a suitable title. The title, in bold letters, conveys the main facts depicted by the diagram.

4.3.2 Size

The size of a diagram should neither be too big nor too small. It must match with the size of the paper. It should be in the middle of the paper.

4.3.3 Scale

The scale should be mentioned in the diagram. The scale should convey the proportional magnitude of the data.

58 *Business Statistics*

4.3.4 Drawing

Diagram should be drawn with the help of drawing instruments. In order to make the drawing attractive, colouring, dotting, crossing etc. should be done in the diagram.

4.3.5 Index

When different colours, dots or markings are given in diagram, the index should be given. The index should properly convey the meaning of the colours, dots or markings.

4.3.6 Accuracy

The diagram should be drawn with accurate measurement. Inaccurate measurement lead to wrong diagram.

4.3.7 Simplicity

Diagram should be very simple. It must be so simple that even a layman who does not have knowledge of mathematical or statistical background can understand it easily.

4.3.8 Sources

If the data presented have been acquired from some external source, that source should be indicated at the bottom.

4.4 TYPES OF DIAGRAMS

The following are the important types of diagrams:

- (i) One-dimensional diagrams
- (ii) Two-dimensional diagrams
- (iii) Three-dimensional diagrams
- (iv) Pictograms and Cartograms

4.4.1 One-dimensional Diagrams

One-dimensional diagram can also be called as bar diagram. In bar diagrams, only the length is considered. The following are the important types of bar diagram.

- (i) Simple Bar diagrams
- (ii) Sub-divided Bar diagrams
- (iii) Multiple Bar diagrams
- (iv) Percentage Bar diagrams
- (v) Deviation Bar diagrams

Simple Bar Diagrams

Simple bar diagram represents only one variable. It gives much importance to only one characteristic of the data. The figures like production, sales in factories, number of students in a college year after can be represented by such bars. The width of the bar is not given any importance.

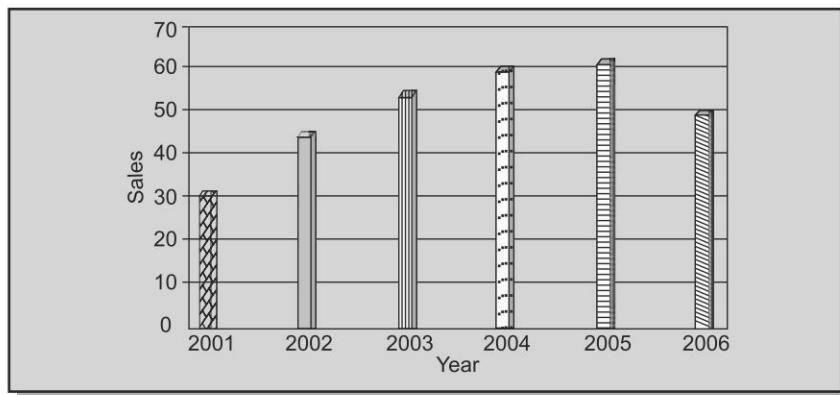
Illustration 4.1

Prepare a bar diagram from the following data:

Year :	2001	2002	2003	2004	2005	2006
Sales :	32	45	55	60	62	50

Solutions

Simple Bar Diagram



Sub-divided Bar Diagrams

Sub-divided bars are used to present such data which are to be shown in the parts or which are totals of various sub-divisions. Each part may explain different characters of the data. For example, the number of students of a college may be divided course-wise.

Illustration 4.2

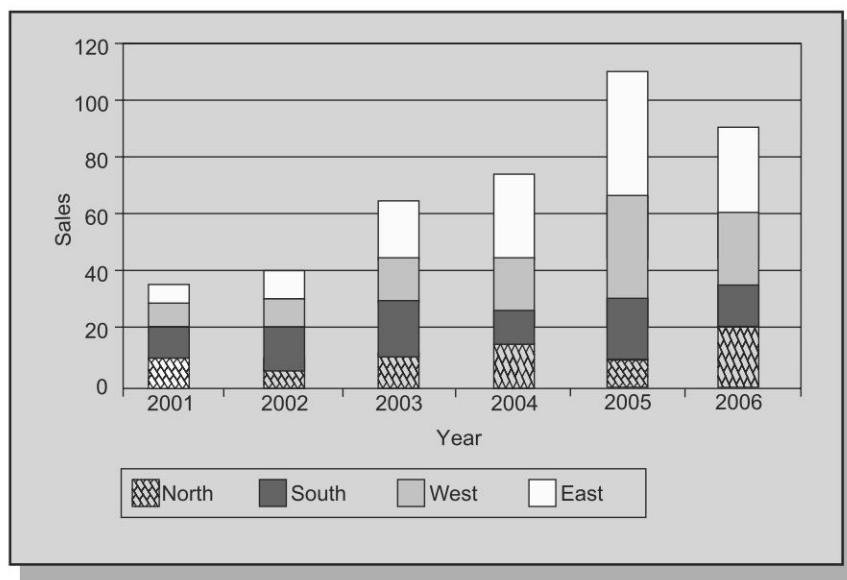
The sales of a company for 2001–06 are stated below.

Year	Regionwise Sales (Rs. '000)				Total Rs. ('000)
	North	South	West	East	
2001	10	10	9	6	35
2002	5	15	10	10	40
2003	10	20	15	20	65
2004	15	10	20	30	75
2005	10	20	35	45	110
2006	20	15	25	30	90

Represent the data in a bar diagram.

Solutions

Sub-divided Bar Diagram

**Multiple Bar Diagrams**

The techniques of simple bar diagrams can be extended to represent two or more sets of inter-related data in one diagram. It supplies information about one phenomenon.

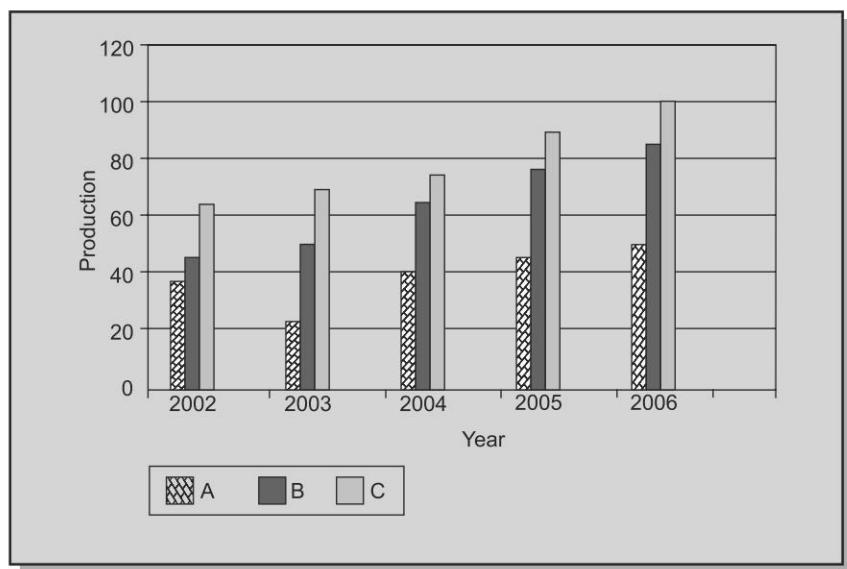
Illustration 4.3

From the following data, draw a suitable diagram:

Year	Production (Rs. '000)		
	A	B	C
2002	37	45	63
2003	22	50	69
2004	40	65	75
2005	45	77	88
2006	52	86	100

Solutions

Multiple Bar Diagrams



Percentage Bar Diagrams

In percentage bar diagram, the length of all the bars are equal. Various parts of each bar are converted into percentage.

Illustration 4.4

Prepare a sub-divided bar diagram drawn on the percentage basis.

Year	Direct Cost Rs.	Indirect Cost Rs.	Profit Rs.	Sales Rs.
1992	15	10	8	33
1993	20	15	10	45
1994	12	11	6	29
1995	5	25	15	45

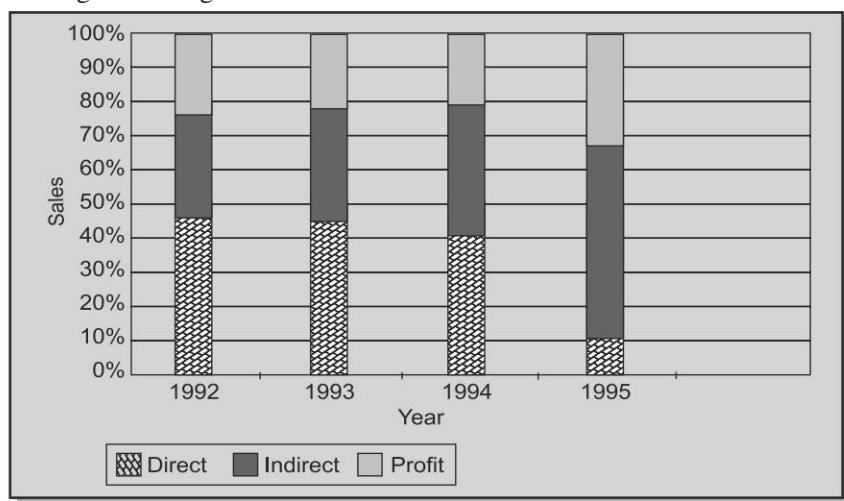
Solutions

Converted table according to percentage basis

Year	Direct Cost Rs.	Indirect Cost Rs.	Profit Rs.	Sales Rs.
1992	46	30	24	100
1993	45	33	22	100
1994	41	38	21	100
1995	11	56	33	100

62 Business Statistics

Percentage Bar Diagrams



Deviation Bar Diagrams

Deviation bar diagrams represent only the difference of (deviations of) figures which is shown in the shape of bars. Bars representing positive differences are shown on one side and those representing negative difference on the other side.

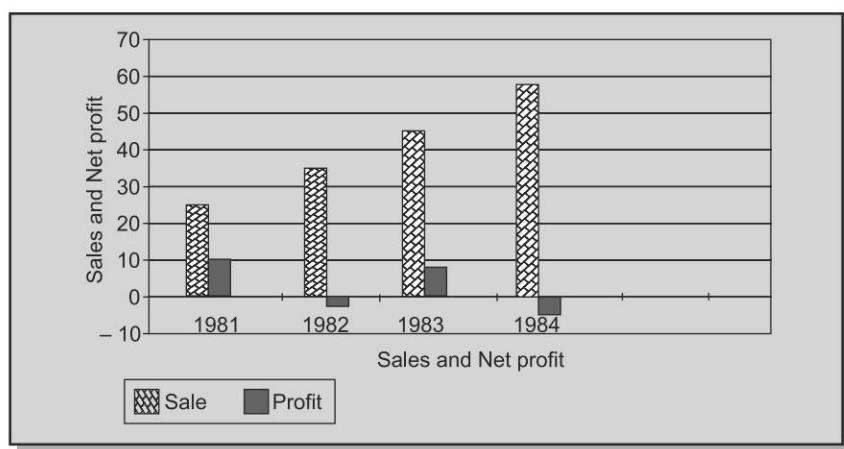
Illustration 4.5

Represent the following data in a suitable bar diagram.

Year	Sales (Rs. In '000)	Profit/Loss (Rs. In '000)
1981	24	10
1982	35	-3
1983	45	7
1984	59	-5

Deviation Bar Diagrams

Sales and Net Profit Position



4.4.2 Two-dimensional Diagrams

In one-dimensional diagram, only one dimension using heights (length) is considered. But in two-dimensional diagram, both lengths as well as width are taken into account. It is also called as area diagram or surface diagram. The important types of two-dimension diagram are:

- (i) Rectangles
- (ii) Squares
- (iii) Circles
- (iv) Pie Diagram

(i) Rectangles In a rectangle diagram, both the dimensions (length and width) of the bars are taken into account. A rectangle is a two-dimension diagram because it is based on the area principle (length and breadth).

Illustration 4.6

Represent the following data in a rectangular diagram.

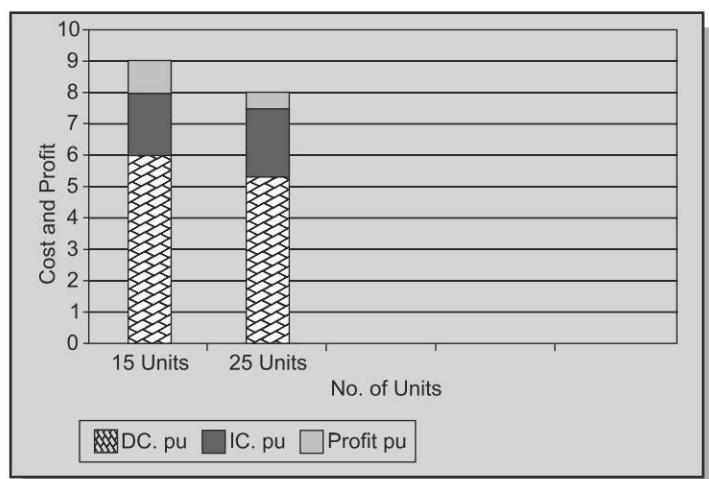
Particulars	Product X	Product Y
Direct Cost (Rs.)	95	135
Indirect Cost (Rs.)	30	50
Profit (Rs.)	15	15
No. of units sold	15	25
Price per unit (Rs.)	9	8

Solutions

Calculation of per unit cost of Product X and Product Y.

Cost and Profit	Product X (15 Units)		Product Y (25 Units)	
	Total Cost (Rs.)	Per unit cost (Rs.)	Total Cost (Rs.)	Per unit cost (Rs.)
Direct Cost (Rs.)	95	6	135	5.4
Indirect Cost (Rs.)	30	2	50	2
Profit (Rs.)	15	1	15	0.6

Rectangular diagram



64 Business Statistics

(ii) **Squares** Under this method, each bar diagram is represented in the form of squares. First we convert the square root of each value of the variable. Then the value should be represented in the form of bar diagram.

Illustration 4.7

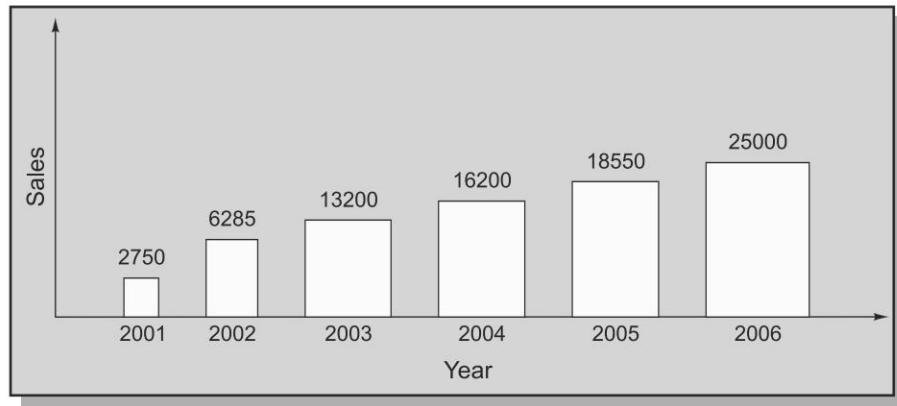
Represent the following data in a square form of two dimensional diagrams.

Year	Sales (Rs.)
2001	2750
2002	6285
2003	13200
2004	16200
2005	18550
2006	25000

Solutions

Calculation of square roots and side of squares for square diagrams.

Year	Sales (Rs.)	Square root in (cms)	Side of the Square
2001	2750	52.44	13.11
2002	6285	79.28	19.82
2003	13200	114.89	28.72
2004	16200	127.28	31.82
2005	18550	136.20	34.05
2006	25000	158.11	39.52



(iii) **Circles** Circle diagrams are more attractive and appealing than square diagrams. The area of the circle is directly proportional to its radius. Each value is taken as area of a circle. The radius is found for each circle, by dividing with π (22/7) and then taking the circle, based upon the radius, circle diagram can be drawn.

Illustration 4.8

Represent the following data in circles in two dimensional diagrams.

Year	Profit (Rs. '000)
1981	8
1982	14
1983	65
1984	100
1985	140
1986	160

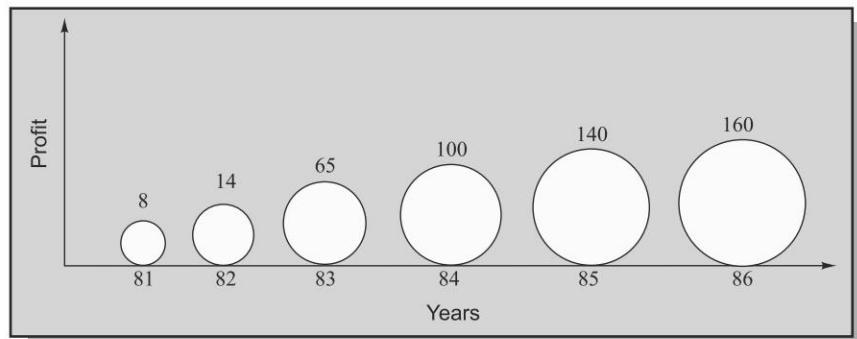
Square root of such value.

Solutions

Calculation of Radius

Year	Profit (Rs. '000)	Dividing by π (22/7)	Square root
1981	8	2.55	1.6
1982	14	4.46	2.11
1983	65	20.68	4.55
1984	100	31.82	5.64
1985	140	44.55	6.67
1986	160	50.91	7.14

Circle Diagram



(iv) **Pie Diagram** Pie diagram means sub-divided circle diagram. It is a representative of various data on the basis of different segments or sections. It gives a clear idea about the percentage of the component part to the total. The percentage/value of any component part is calculated by applying the following formula, because the angle at the centre of the circle is 360° .

$$\frac{\text{Component value}}{\text{Total value}} \times 360$$

It is also known as angular diagram.

Illustration 4.9

Draw a pie diagram from the given data. The selling price of a product contains the following elements of costs and profit

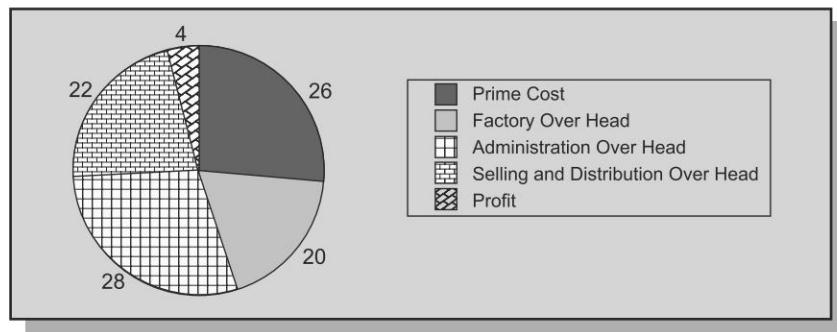
Prime Cost	26%
Factory Overhead	20%
Administrative Overhead	28%
Selling and Distribution Overhead	22%
Profit	4%

Solutions

Calculations for the preparation of pie diagram

Elements of Cost and Profit	Percentage	Angles
Prime Cost	$26/100 \times 360 = 26$	94°
Factory Overhead	$20/100 \times 360 = 20$	72°
Administrative Overhead	$28/100 \times 360 = 28$	101°
Selling and Distribution Overhead	$22/100 \times 360 = 22$	79°
Profit	$4/100 \times 360 = 4$	14°
	<u>100</u>	<u>360</u>

Pie Diagram



4.4.3 Three-dimensional Diagrams

Three-dimensional diagrams are those in which three dimensions wing, length, breadth and height are taken into account. They are constructed in the form of cubes, spheres, cylinders and blocks.

4.4.4 Pictograms and Cartograms

(i) **Pictograms** Pictograms is the technique of presenting statistical data through appropriate pictures. Pictures are more attractive and appealing to the eye. The

number of pictures drawn or the size of the pictures being proportional to the values of the different magnitude to be presented.

Present the following data in a pictogram

Animals	Forest I	Forest II
Elephants	16,000	12,000
Horses	8,000	9,000
Deer	4,000	3,500
Monkeys	15,000	11,000
Giraffes	2,800	3,200
Lions	900	1,100
Bears	800	900

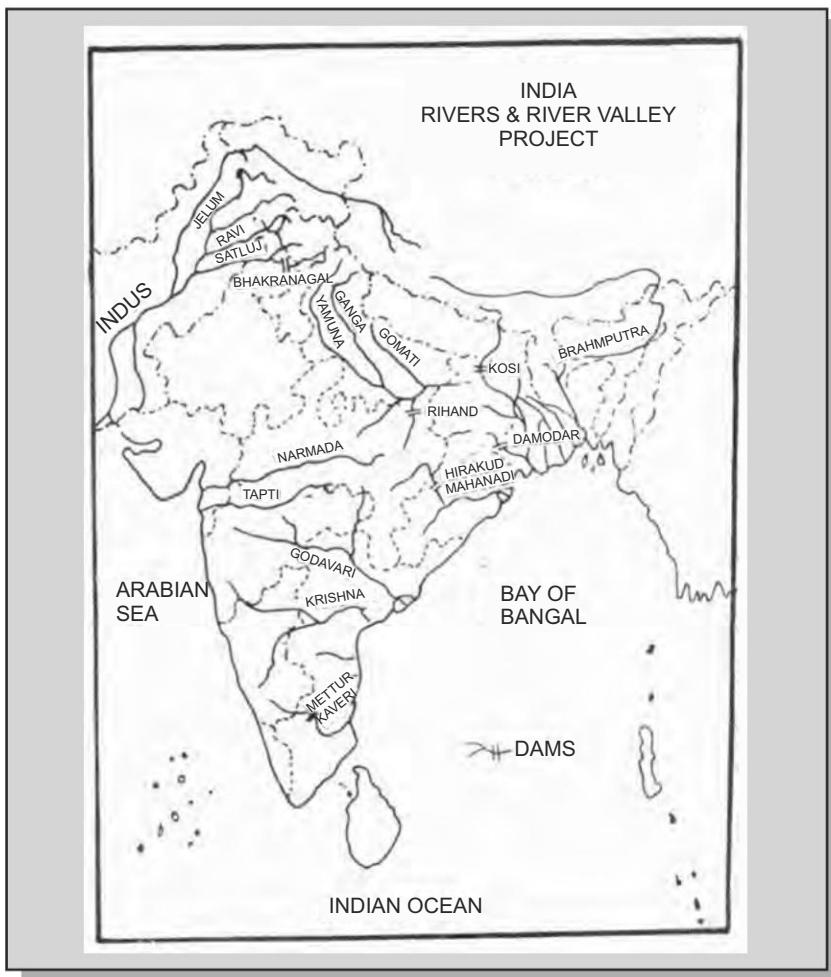
		Forest A	Forest B
Elephant		16,000	12,000
Camel		8,000	9,000
Deer		4,000	3,500
Monkey		15,000	11,000
Giraffes		2,800	3,200
Lions		900	1,100
Bears		800	900

Illustration 4.10

Represent the following data in a pictogram.

(ii) **Cartograms** In Cartograms, statistical facts are presented through maps accompanied by various types of diagrammatic representative. It is the presentation of data in geographical basis. It is also called as statistical maps. The quantities or magnitudes in the regions or geographical zones may be shown by dots, clouds, different shades.

For example, a cartogram representing rivers and river valleys for the map of India is presented below.



4.5 GRAPHICAL REPRESENTATION

Frequency distribution related to discrete and continuous series can be well drawn in a graph. A graph is a visual form of presentation. Graphical presentation of statistical data gives a pictorial effect. Graphs are very useful for studying time series. Graphs are drawn on a special type of paper known as graph sheet. The special feature of the graphs is that they are more obvious, accurate and precise diagram.

4.5.1 Classification of Graphs

It is classified into two major heads:

- (i) Graphs of frequency distribution
 - (a) Histogram
 - (b) Frequency Polygon

- (c) Frequency Curve
- (d) Ogives or cumulative frequency curve
- (ii) Graphs of time series
 - (a) Nature Scale Method
 - (i) Line Graph or Line Chart for one variable
 - (ii) Line Graph or Line chart for two or more variables
 - (b) Ratio Scale Method

(i) Graphs of Frequency Distribution

(a) Histogram It is one of the major popular and commonly used devices for drawing continuous frequency distribution. It is a set of vertical bars. Frequencies representing the variables should be drawn in a graph in the form of vertical bars. It is also called as graphs of time series.

Differences between Histogram and Bar Diagram.

1. A histogram is a two-dimensional figure where both the width and the length are important whereas a bar diagram is one-dimensional diagram in which only length is to be considered.
2. In a histogram, the rectangles are adjacent to each other whereas in bar diagram proper spacing is given between two rectangles.
3. The class frequencies are represented by the area of the rectangles.

Types of Histogram

- (a) Histogram for frequency distributions having equal class-intervals.
- (b) Histogram for frequency distributions having unequal class-intervals.

(a) Histogram for frequency distributions having equal class-intervals

Construction of Histogram The following are steps for the construction of histogram when class-intervals are equal.

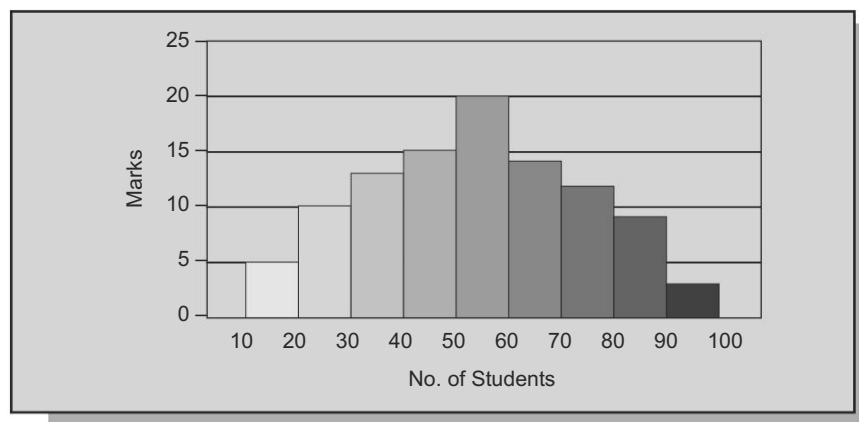
1. Variables are taken in X axis.
2. Frequencies are taken in Y axis.
3. Plot points of variables and frequencies. These two points with equal distance are taken for the lower and upper limits of each class interval.
4. Connect all the plotted points towards the X axis to get rectangular adjacent bars.

Illustration 4.11

Draw a histogram from the following information.

Marks	No. of Students
10–20	5
20–30	10
30–40	13

40–50	15
50–60	20
60–70	14
70–80	12
80–90	9
90–100	3



(b) Histogram for frequency distributions having unequal class-intervals

Steps:

For construction of Histogram when class intervals are unequal.

- The unequal class intervals must be corrected first with the help of frequency density method.

Corrected Class-Interval =

$$\text{Frequency of unequal class interval} \times \frac{\text{Width of the lowest}}{\text{Width of the unequal class interval}} \times \text{Class interval}$$

- Variables are plotted on the x axis.
- Frequencies are plotted on the y axis.
- Plot points of variables and frequencies. These two points with equal distance is taken for the lower and upper limits of each class-interval.
- Connect all the plotted points towards the x axis to get rectangles adjacent bars.
- Histogram for frequency distribution having unequal class intervals.

Illustration 4.12

Draw a histogram from the following data.

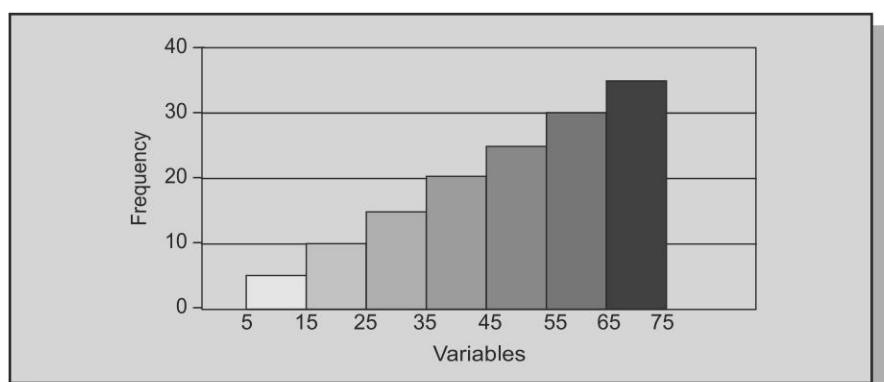
Daily Wages (Rs.)	No. of Workers
5–10	2
10–15	4
15–20	6
20–30	18
30–35	5
35–40	3
40–50	16

Solutions

Since in this problem the class intervals are not equal, the frequencies should be adjusted for the purpose of drawing the histogram.

Calculation of Frequency Density

Daily Wages (Rs.)	No. of Workers	Frequency Density
5–10	2	2
10–15	4	4
15–20	6	6
20–30	18	9
	18/10*5	
30–35	5	5
35–40	3	3
40–50	16	8
	16/10*5	



(b) **Frequency Polygon** It is another device of distribution. It gives a curve instead of bars. It is an improved method of histogram. It is drawn for both discrete series and continuous series.

In case of discrete frequency distribution, frequency polygon is obtained by plotting the frequencies on the *Y* axis against the corresponding values of the variable on the *X* axis and joining the points so obtained by straight lines.

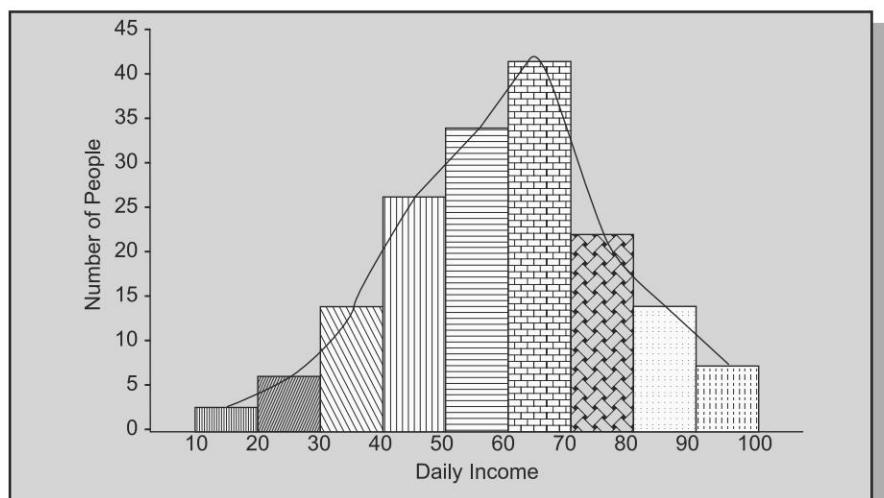
Illustration 4.13

Represent the following data in a histogram and frequency polygon.

Daily Income (Rs.)	No. of People
10–20	2
20–30	6
30–40	14
40–50	26
50–60	34
60–70	41
70–80	22
80–90	14
90–100	6

Solutions

Histogram and Frequency Polygon



(c) **Frequency Curve** A frequency curve is a smooth, free hand curve drawn through the vertices of a frequency polygon. The object of this curve is to eliminate the erratic ups and downs. For drawing the frequency curve, first of draw a frequency polygon by joining mid-points of each class interval. Then the frequency polygon should be smoothed.

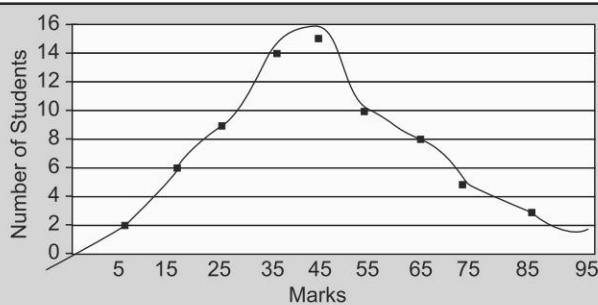
Illustration 4.14

Draw a frequency curve through frequency polygon from the data given below:

Marks	No. of Students
5–15	2
15–25	6
25–35	9

35–45	14
45–55	15
55–65	10
65–75	8
75–85	5
85–95	3

Solutions



(d) Ogives or Cumulative Frequency Curve Ogives, pronounced as olive, is a graphical presentation of the cumulative frequency of continuous series. It is drawn by connecting plots of the cumulative frequency and the class intervals.

Ogives can be constructed in two methods:

(i) Less than Ogives This consists in plotting the ‘less than’ cumulative frequencies against the upper class boundaries of the respective classes. The points so obtained are joined by a smooth, free hand curve to give less than ogive. This curve is an increasing curve, sloping upwards from left to right.

(ii) More than Ogives Similarly, more than ogives are plotted against the lower class boundaries of the respective classes. The points so obtained are joined by a smooth free hand curve to give ‘more than olives’. It is a decreasing curve and slope downwards from left to right.

Illustration 4.15

Draw less than and more than ogives from the following data:

Marks	No. of Students
0–10	2
10–20	4
20–30	8
30–40	14
40–50	20
50–60	15
60–70	8
70–80	5
80–90	4

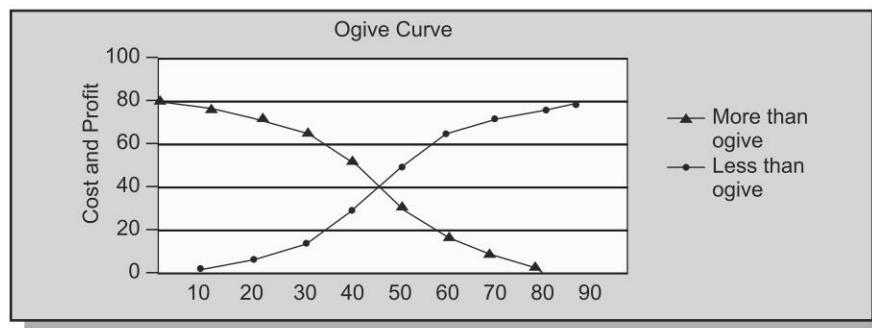
Solutions

Calculation for less than ogives

Marks	No. of Students	Cumulative Frequencies
10	2	2
20	4	6
30	8	14
40	14	28
50	20	48
60	15	63
70	8	71
80	5	76
90	4	80

Calculation for more than ogives

Marks	No. of Students	Cumulative Frequencies
0	2	80
10	4	78
20	8	74
30	14	66
40	20	52
50	15	32
60	8	17
70	5	9
80	4	4



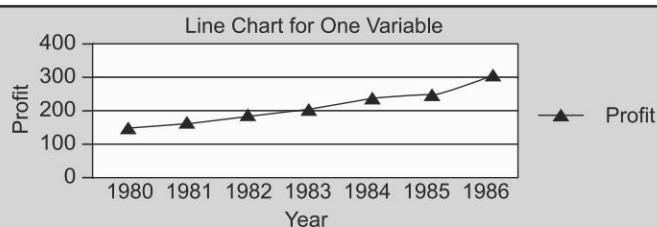
Graphs of Time Series Time series is concerned with the representation of data for different periods of time. Time (year, month and day) may be taken in x axis and variables (population, demand, production) may be taken in y axis. There are two methods for constructing graphs for Time Series. They are:

- Nature Scale Method
- Ratio Scale Method

Illustration 4.16

Construct a line chart for the following data:

Year	No. of Students
1980	145
1981	160
1982	180
1983	200
1984	230
1985	240
1986	300

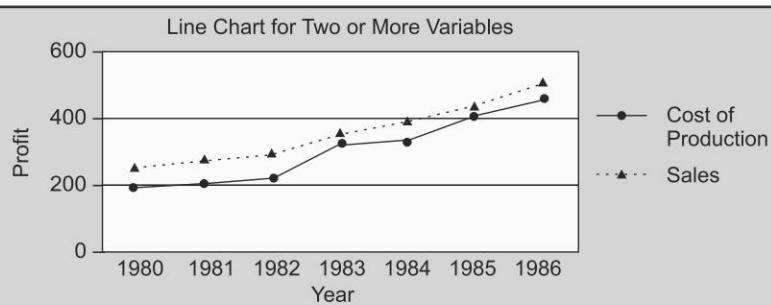
Solutions

(ii) **Line chart for two or more variables** When two or more variables are available corresponding to various periods of time, line graph shows the two or more line charts. Time should be taken in x axis and variable should be taken in y axis. First draw the line chart for one variable and draw the line chart for second variable.

Illustration 4.17

Represent the following data in the graph.

Year	Cost of Production Rs. ('000)	Sales Rs. ('000)
1980–81	185	250
1981–82	200	270
1982–83	210	290
1983–84	320	350
1984–85	325	385
1985–86	400	425
1986–87	450	500

Solutions

Band Graph Band graph is a type of line chart which shows the different components of variables in the graph. For example, the cost of production of a factor for different periods may be represented with different elements of cost, such as material, labour and overheads.

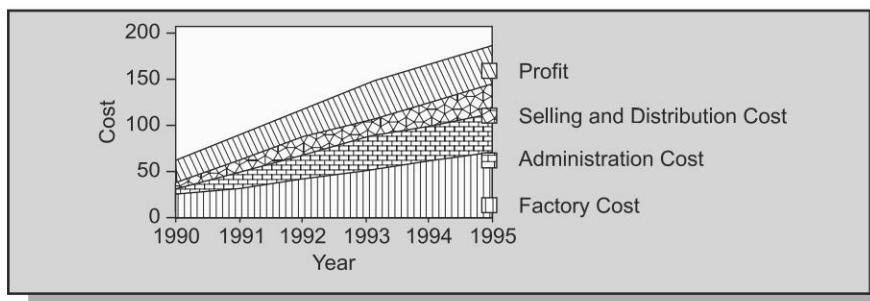
Illustration 4.18

Prepare a Band graph for the following data.

Year	Factory Cost Cost	Admn.Costs (Rs. '000)	Selling and Dist. Cost	Profit (Rs. '000)	Total (Rs. '000)
1990	25	10	5	2	42
1991	30	20	10	5	65
1992	40	30	15	10	95
1993	50	38	20	20	128
1994	60	39	25	25	149
1995	70	42	30	30	172

Solutions

Band Graph



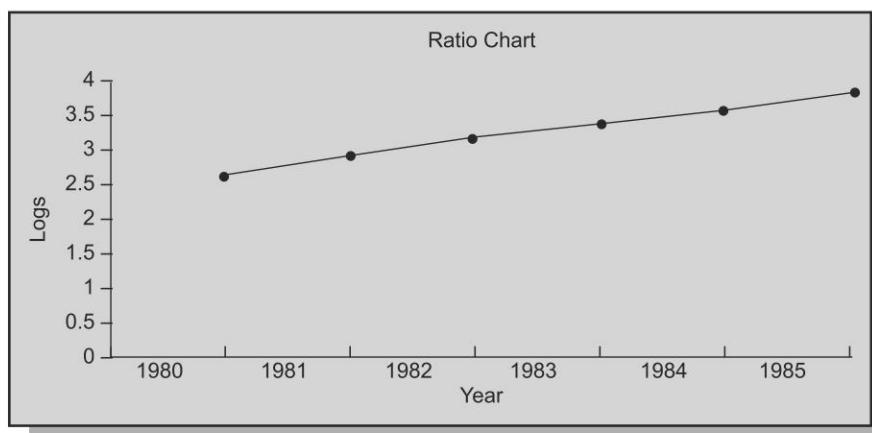
(b) Ratio Scale The graphs drawn with ratio scale are called ratio charts or logarithmic charts. It shows the related change in variables whereas the natural scale shows only the actual change in variables. In order to compare the relative change of variables over a period of time, ratio chart or logarithmic chart is prepared. The values in natural scale are taken in x axis and logarithmic of the values of variables are taken in y axis. For example, population of our country for every census and the logarithmic value are to be taken for drawing the chart.

Illustration 4.19

Construct a ratio chart for the following data.

Year	No. of Students	Logarithmic Value
1980	400	2.6021
1981	900	2.9542

1982	1600	3.2041
1983	2400	3.3802
1984	3300	3.5185
1985	4600	3.6628



Z Curve or Zee Chart Z chart is a special kind of graph. The chart consists of three curves. When drawn, the curves taken together looks like Z. The three curves are:

- 1 Curve of original data
 - 2 Curve of cumulative data
 - 3 Moving total curve
- Z chart indicates the trend of the business firm.

4.6 MISCELLANEOUS

Illustration 4.20

The sale proceeds cost and the profit or loss per wooden chair of a firm are:

Particulars	2001	2002	2003
Sale Proceeds	150	200	250
Cost per chair material	75	100	125
Wages	30	60	90
Other costs	25	50	75
Total cost	130	210	290
Profit/Loss	20	-10	-10

Draw bars to show (a) absolute value and (b) relative values of the problem.

Solutions

Table showing absolute and relative values

Particulars	2001		2002		2003	
	Absolute Values	Relative Values	Absolute Values	Relative Values	Absolute Values	Relative Values
Sale Proceeds	150	100%	200	100%	250	100%
Cost Per Chair:						
Material	75	50%	100	50%	125	50%
Wages	30	20%	60	30%	90	36%
Other costs	25	16.67%	50	25%	75	30%
Total Cost	130	86.67%	210	105%	290	116%
Profit/Loss	20	13.33%	-10	-5%	-40	-16%

Diagram showing an absolute value

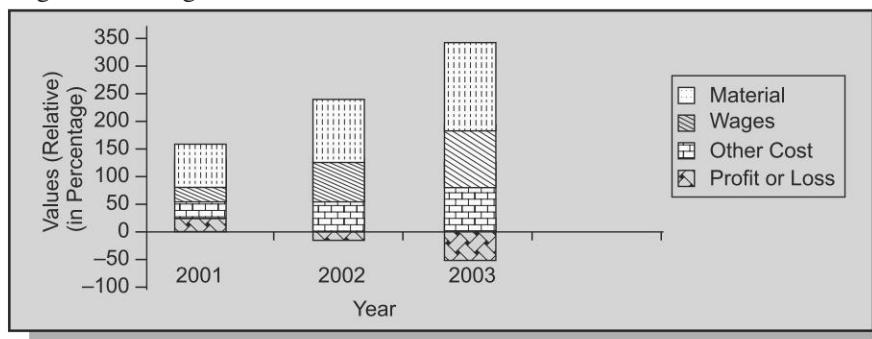


Diagram showing a relative value

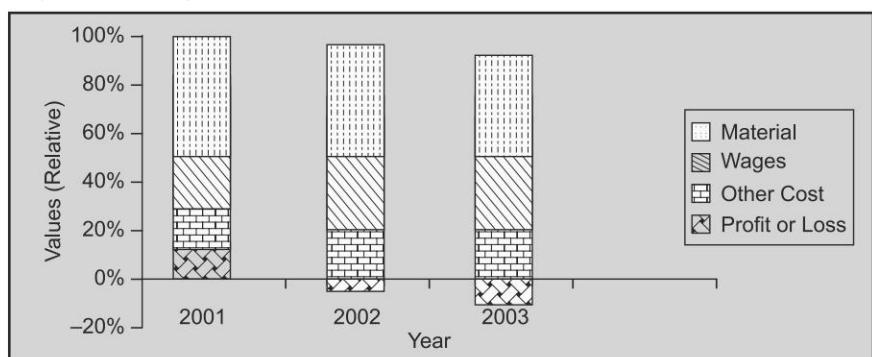


Illustration 4.21

The data gives the yearly profits (in '000 of Rupees) of the two companies, A and B.

Year	Profits in '000	
	Company A	Company B
1990	100	90
1991	120	100
1992	130	110
1993	150	120

Represent the data by means of a suitable diagram.

Solutions

Yearly profit in '000 Rupees

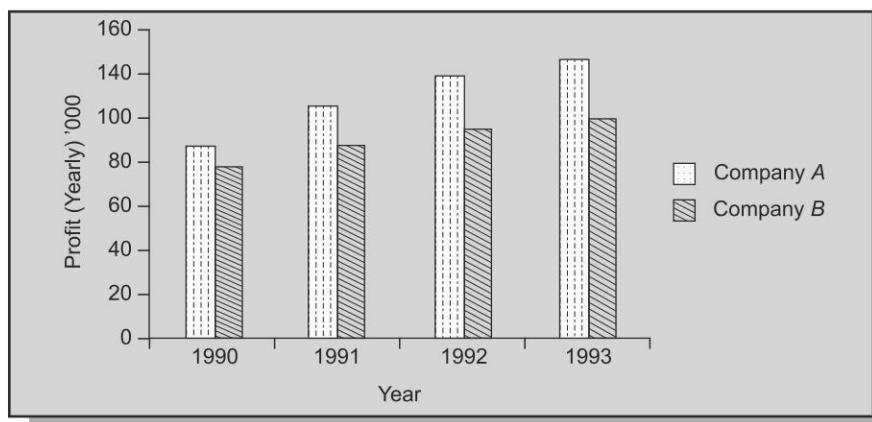


Illustration 4.22

Draw a multiple bar diagram for the following data.

Year	Sales ('00)	Gross Profit ('00)	Net Profit ('00)
1980	200	30	10
1981	210	40	20
1982	220	75	30
1983	230	60	30

Solutions

Sales, gross profit and net profit for the years from 1980 to 1983

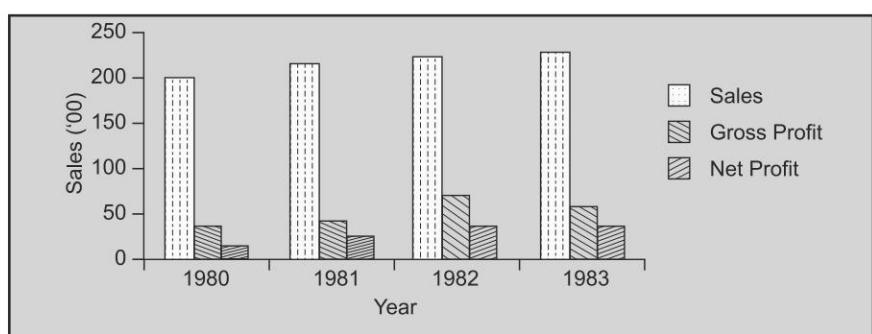


Illustration 4.23

Height	No. of trees
Below 5 feet	25
Below 10 feet	50

Below 15 feet	100
Below 20 feet	150
Below 25 feet	200
Below 30 feet	250
Below 35 feet	300

Represent the data in the form of histogram.

Solutions

Histogram of the height of trees

To draw a histogram, we have to obtain simple frequencies:

Height	No. of trees
0–5	25
5–10	25
10–15	50
15–20	56
20–25	44
25–30	55
30–35	45

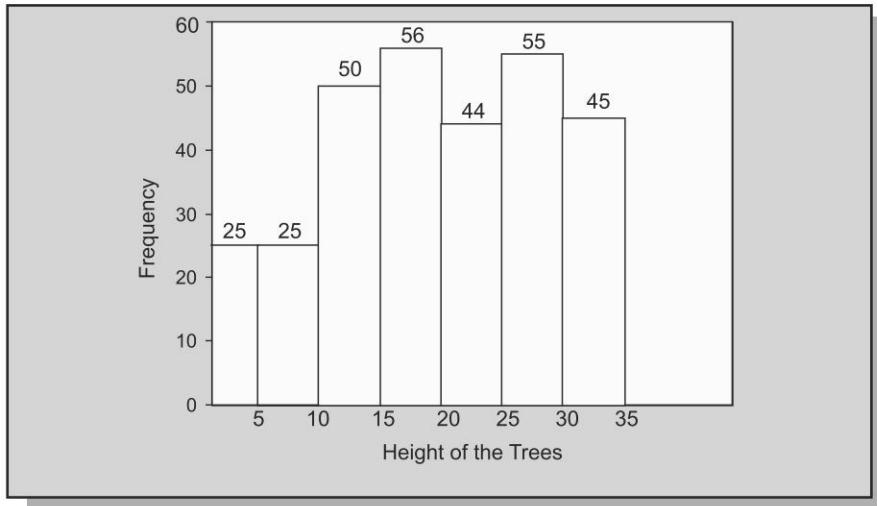


Illustration 4.24

The following figures relate to the cost of construction of a house in Delhi :

Item	Expenditure
Cement	20%
Steel	15%
Bricks	10%
Timber	18%
Labour	20%
Miscellaneous	17%

Represent the data by a suitable diagram.

Item	Expenditure	Degrees
Cement	20%	$20/100 * 360 = 72$
Steel	15%	54
Bricks	10%	36
Timber	18%	64.8
Labour	20%	72
Miscellaneous	17%	61.2
	100%	360

Pie Diagram showing the cost of construction of a house in Delhi

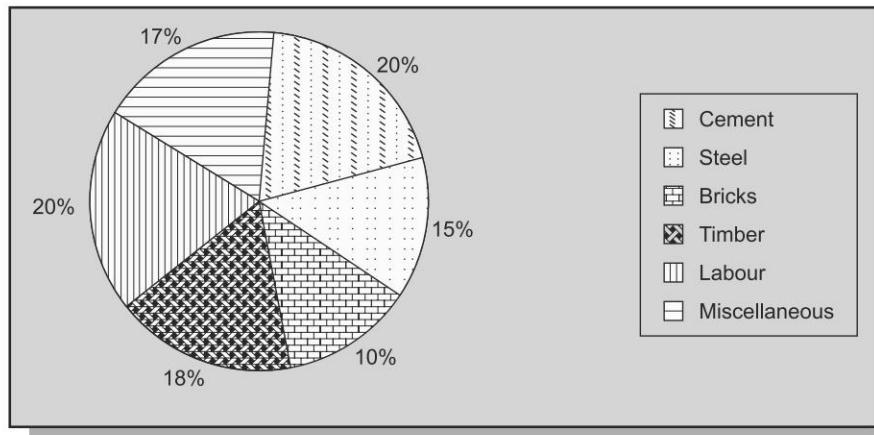


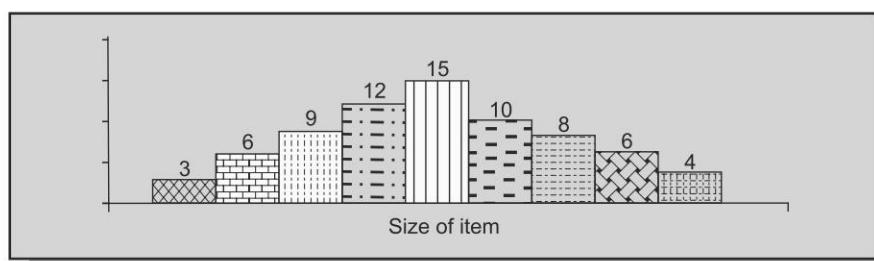
Illustration 4.25

Find out the mode of the following data graphically and check the result through calculation.

Size	Frequency
1–2	3
2–3	6
3–4	9
4–5	12
5–6	15
6–7	10
7–8	8
8–9	6
9–10	4

Solutions

Histogram

**Illustration 4.26**

Draw a histogram from the following data:

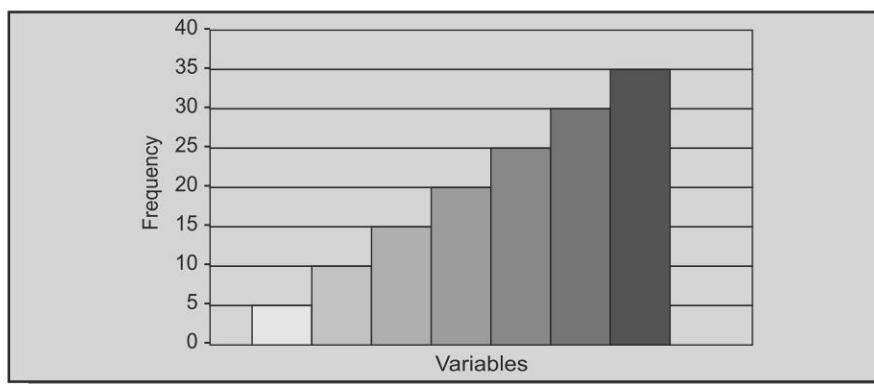
Mid Value	Frequency
10	5
20	10
30	15
40	20
50	25
60	30
70	35

Solutions

Note: We will have to find out the lower and upper value of each class and they are:

Histogram

5–15	5
15–25	10
25–35	15
35–45	20
45–55	25
55–65	30
65–75	35

**Illustration 4.27**

A firm reported that its net worth in the years 1989 to 1993 is as follows:

Year	Net Worth
1989	80
1990	100
1991	120
1992	130
1993	140

Plot the above data in the form of a semi-logarithmic graph. Can you say anything about the approximate rate of growth of its net worth?

Solutions

Year	Net Worth Y	Log y
1989	80	1.90
1990	100	2.00
1991	120	2.08
1992	130	2.11
1993	140	2.15

It reflects that the rate of growth is increasing for the whole period.

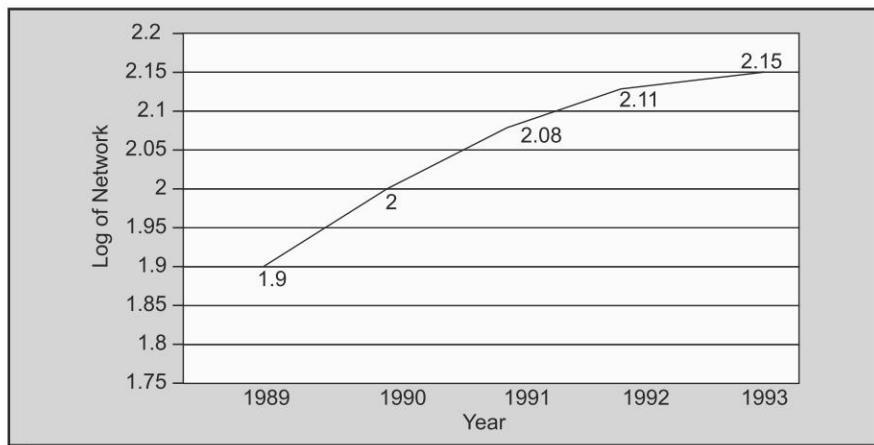
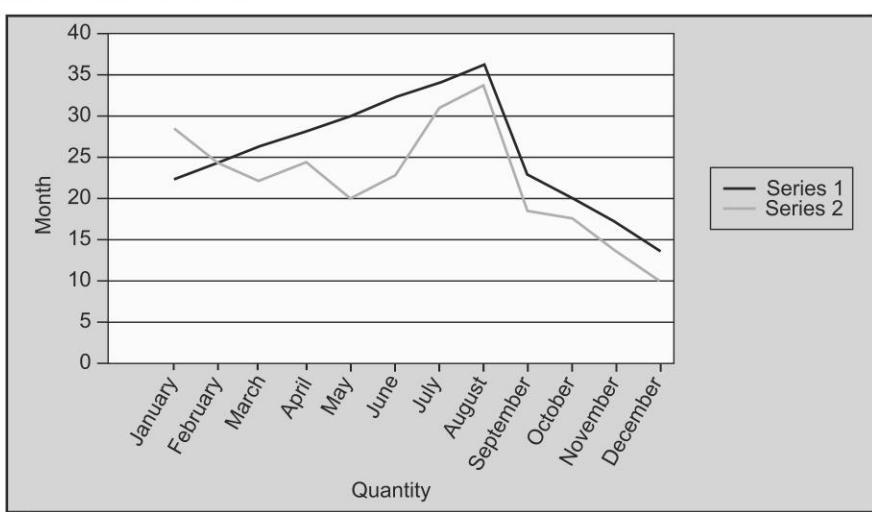


Illustration 4.28

The following table gives the value of imports and exports of a country for the year 2006 in crores of rupees. Represent by means of a graph showing also the balance of trade.

Month	Imports	Exports
January	22	28
February	24	24
March	26	22
April	28	24
May	30	20
June	32	22
July	34	31
August	36	34
September	23	19
October	20	18
November	17	14
December	14	10

Solutions**Illustration 4.29**

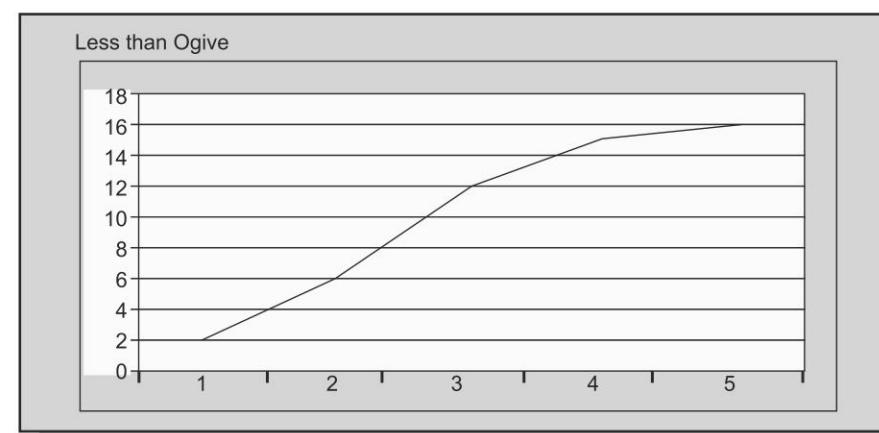
Draw a cumulative frequency curve for the following data:

Marks	Frequency
10–20	2
20–30	4
30–40	6
40–50	3
50–60	1

Solution

Cumulative frequency of the above problem.

Marks less than	20–2
Marks less than	30–6
Marks less than	40–12
Marks less than	50–15
Marks less than	60–16



SUMMARY

General Rules for Preparing Diagrams

- Suitable heading should be given.
- Scale should be mentioned.
- Diagram should be drawn with the help of drawing instruments.
- Index should be given.
- Diagram should be drawn attractively.
- Diagram should be drawn with accurate measurement.
- Diagram should neither be too big nor too small.

Type of Diagrams

- One-Dimensional Diagrams or Bar Diagram
- Two-Dimensional Diagrams
- Three-Dimensional Diagrams.
- Pictograms and Cartograms.

Types of Bar Diagram

- Simple Bar Diagram
- Sub-Divided Bar Diagram
- Multiple Bar Diagram

- Percentage Bar Diagram
- Deviation Bar Diagram

Types of Two-Dimensional Diagram

- Rectangle
- Squares
- Pie Diagram

Methods for drawing graphs for Frequency Distribution

- Histogram
 - Frequency Curve
 - Ogives or Cumulative Frequency Curve
- ⇒ Bar Diagram : One dimensional-only length of bar is important.
- ⇒ Histogram : Two dimensional-length as well as width are important.

EXERCISES

(a) Choose the best option.

1. _____ are created interest in the mind of the readers.
(a) Tables (b) Diagrams (c) Charts
2. _____ bar diagram can be drawn either or horizontal or vertical bars.
(a) Multiple (b) Simple (c) Sub-divided
3. _____ diagrams are alternative to square diagram.
(a) Circle (b) Square (c) Rectangles
4. _____ diagram ranks high in understanding.
(a) Circle (b) Square (c) Pie
5. Pie diagram is called as _____ diagram.
(a) square (b) circle (c) angular
6. Component graph also known as _____ graph.
(a) band (b) range (c) net balance
7. Net balance graph is also known as _____.
(a) range (b) band (c) silhouette
8. _____ determine median, quartiles, percentiles.
(a) Ogives (b) Frequency curve (c) Histograms
9. _____ curve should begin and end at the base line.
(a) Histogram (b) Frequency (c) Ogives

Answers

- | | | | | | |
|------|------|------|------|------|------|
| 1. b | 2. b | 3. a | 4. c | 5. c | 6. a |
| 7. c | 8. a | 9. b | | | |

(b) Fill in the blanks.

1. _____ is a visual form for presentation of statistical data.
2. Diagrams cannot be _____ further.
3. Diagrams play an important role in the _____.
4. _____ is the simplest of all the diagrams.
5. _____ are used to denote more than one phenomenon.
6. _____ bar diagram is used to depict the net deviations in different values.
7. _____ is a device of representing statistical data in pictures.
8. _____ are presented through maps.
9. _____ is not attractive.
10. _____ presentation of statistical data gives a pictorial effect.
11. Graphic is a _____ of presentation.
12. A grouped frequency distribution can be represented by a _____.
13. A _____ is drawn by smoothing the frequency polygon.
14. _____ means the graph relating to a series spread over a period of time.
15. _____ is also known as Zero graph.
16. _____ is prepared when a variable has several components.
17. _____ is a special kind of graph.
18. Z chart is very useful for _____.

Answers

- | | |
|---------------------------------|--------------------|
| 1. Diagram | 2. Analysed |
| 3. Modern advertising campaigns | 4. Line diagram |
| 5. Multiple bar diagrams | 6. Deviation |
| 7. Pictogram | 8. Cartograms |
| 9. Line diagram | 10. Graphic |
| 11. Visual form | 12. Histogram |
| 13. Frequency curve | 14. Histogram |
| 15. Range graph | 16. Belt curve |
| 17. Z chart | 18. Business firms |

(c) Theoretical Questions

1. Define diagram. State the important rules for preparing diagram.
2. Define pie diagram. State the procedure for drawing pie diagram.
3. Write short notes on
 - (a) Circle diagram (b) Pictogram
 - (c) Cartogram (d) Ogives

4. What do you understand by graphs?
5. Explain briefly the different types of graphs of time series.
6. Define graphs.
7. Explain the difference between graphs and diagrams.
8. Give two advantages of two dimensional diagram.
9. Explain Histogram.
10. Describe the construction of an Ogive.
11. State briefly the purpose served by diagrammatical presentation.
12. Describe any two methods of graphical representation of statistical data.
13. What is Z chart?
14. Explain the frequency polygon and the frequency curve.
15. Explain the briefly the concept of false base line.

(d) Practical Problems

16. The following data shows the number of accidents sustained by 314 drivers of a public utility company over a period of five years.

Number of accidents:	0	1	2	3	4	5	6	7	8	9	10	11
Drivers:	82	44	68	41	25	20	13	7	5	4	3	2

Represent the data by a line diagram.

17. The following data relating to the strength of the Indian Merchant shipping fleet gives the gross registered tonnage (GRT) as on 31st December for different years.

Year :	2001	2002	2003	2004	2005
GRT in '000 :	901	1792	2500	4464	5115

Source : Ministry of shipping and transport.

Represent the data by suitable bar diagram.

18. Represent the following data by a suitable diagram.

Items of Expenditure	Family A (Income Rs. 500)		Family B (Income Rs. 300)	
Food	150		150	
Clothing	125		60	
Education	25		50	
Miscellaneous	190		70	
Saving or Deficit	+ 10		- 30	

19. The adjoining table gives the break-up of the expenditure of a family on different items of consumption. Draw percentage bar diagram to represent the data.

Item	Expenditure(Rs.)	
Food	240	
Clothing	66	
Rent	125	
Fuel and Lighting	57	
Education	42	
Miscellaneous	190	

20. Draw a bar chart for the following data showing the percentage of total population in villages and towns.

	Percentage of total population in	
	Villages	Towns
Infants and young children	13.7	12.9
Boys and girls	25.1	23.2
Young men and women	32.3	36.5
Middle-aged men and women	20.4	20.1
Elderly Persons	8.5	7.3

21. The data below give the yearly profits (in thousand of rupees) of two companies, A and B.

Year	Profits in ('000 Rupees)	
	Company A	Company B
1994–95	120	90
1995–96	135	95
1996–97	140	108
1997–98	160	120
1998–99	175	130

Represent the data by means of a suitable diagram.

22. The following data shows the students in millions on rolls at school/university stage in India according to different class groups and sex for the year 1970–71 as on 31st March.

Stage	Boys	Girls	Total
Class I to V	35.74	21.31	57.05
Class VI to VIII	9.43	3.89	13.32
Class IX to XI	4.87	1.71	6.58
University/College	2.17	0.64	2.81

Represent the data by

- (i) Component bar diagram and
- (ii) Multiple bar diagram

23. Represent the following data relating to the military statistics at the border during the war between the two countries, A and B in 1999 by multiple bar diagram.

Category	Country A	Country B
Army Divisions	4	20
Semi-Army units	50	—
Fighter planes	75	700
Tanks	50	300
Total Troops	100,000	170,000

- 24.** Prepare a rectangular diagram from the following particulars relating to the production of a commodity in a factory.

Units produced	1000
Cost of raw materials	Rs. 5000
Direct Expenses	Rs. 2000
Indirect Expenses	Rs. 1000
Profit	Rs. 1000

- 25.** Represent the following data by a percentage sub-divided bar diagram.

Item of Expenditure	Family A	Family B
	Income Rs 5000	Income Rs 3000
Food	1500	1500
Clothes	1250	600
Education	250	500
Miscellaneous	1900	700
Savings or Deficit	100	-300

- 26.** Draw a square diagram to represent the following data:

Country	A	B	C
Yield in (kg) Per hectare	350	647	1120

- 27.** Draw a pie diagram to represent the following data of proposed expenditure by a state government for the year 1997–98.

Items	Agriculture and Rural Development	Industries and Urban Development	Health and Education	Miscellaneous
Proposed Expenditure (in million Rs.)	4200	1500	1000	500

- 28.** The following table shows the area in millions of sq.km of oceans of the world.

Ocean	Area (million sq.km)
Pacific	70.8
Atlantic	41.2
Indian	28.5
Antarctic	7.6
Arctic	4.8

Draw a pie diagram to represent the data.

- 29.** The following table gives the number of vessels as on 31st December, in Indian merchant shipping fleet for different years.

Year	:	1961	1966	1971	1975	1976
No. of vessels	:	174	231	255	330	359

Represent the data by pictogram.

- 30.** (a) Draw a bar chart to represent the following information:

Year	1952	1957	1962	1967	1972
No. of Women M.P.s	22	27	34	31	22

(b) Represent the following data with the help of a bar diagram

Year	1970–71	1971–72	1972–73	1973–74	1974–75	1975–76	1976–77
No. of Women M.P.s	4221	4655	5272	6159	6231	6578	7778

Women

M.P.s

(a) In a recent study on cause of strikes in mills, an experimenter collected the following data.

Causes	Economic	Personal	Political	Rivalry	Others
Occurrences (in Percentage)	58	16	10	6	10

Represent the data by bar chart.

(b) Below are data on the number of films made in different regional and/or other languages in India in different years.

Year	1947	1951	1961	1970	1971	1972	1973
No. of Films	281	229	303	396	433	414	448

Draw a bar chart to represent the above data.

(B.Com. MKU, MSU, BDU)

31. (a) Represent following data by a suitable diagram.

Item of Expenditure	Family A		Family B	
Food	200		250	
Clothing	100		200	
House rent	80		100	
Fuel and light	30		40	
Education	90		210	
Total	500		800	

Draw Rectangular percentage diagram.

(B.Com. MSU, CHU, BDU, BU)

(b) Represent the following data by a percentage sub-divided bar diagram.

Item of Expenditure	Family A		Family B	
	Income Rs. 500		Income Rs. 300	
Food	150		150	
Clothes	125		60	
Education	25		50	
Miscellaneous	190		70	
Savings or Deficit	10		-30	

(c) Represent the data relating to the cost of construction of two tables by a percentage diagram.

	Table I ('00 Rs)	Table II ('00 Rs)
Wood	5	10
Other materials	2	8
Labour	2	5
Other expenses	1	2
	<u>10</u>	<u>25</u>

92 Business Statistics

(d) Draw a suitable diagram to represent the following data on livelihood patterns in India, U.S.A. and U.K.

Occupation	India	U.S.A.	U.K.
Agriculture and Forestry	71%	13%	5%
Manufacture and Commerce	15%	46%	55%
Other Industries and Services	14%	41%	40%
Total	100%	100%	100%

(B.Com. CHU, BDU)

32. Draw a rectangular diagram to represent the following information.

	Factory A	Factory B
Price per unit	Rs. 15	Rs. 12
Unit produced	1000 Nos.	1200 Nos.
Raw mater/unit	Rs. 5	Rs. 5
Other expenses/unit	Rs. 4	Rs. 3
Profit/unit	Rs. 6	Rs. 4

(B.B.A. MKU, MSU, BU)

33. Draw a pie diagram to represent the distribution of a certain blood group 'O' among Gypsies, Indians and Hungarians.

Blood Group	Frequency			Total
	Gypsies	Indians	Hungarians	
'O'	343	313	344	1000

(b) A ship has four compartments labelled 1, 2, 3 and 4. The space limits of 1, 2, 3 and 4 are respectively 180,000 cubic feet, 160,000 cubic feet, 140,000 cubic feet and 120,000 cubic feet. Present the data about the different space limit in a table and draw a pie diagram to represent the above data.

(c) Represent the following data of the distribution of income of a leading company by a suitable diagram.

	Rupees in Lakhs
Raw Materials	1689
Taxes	582
Manufacturing Expenses	543
Employees	470
Other expenses	286
Depreciation	94
Dividends	75
Retained Income	51
Total	3790

(B.Com. MKU, CHU, BDU)

34. The areas of the various continents of the world in millions of square miles are presented below:

AREAS OF CONTINENTS OF THE WORLD

Continent	Africa	Asia	Europe	North America	Oceans	South America	U.S.S.R.	Total
Area (millions of Square miles)	11.7	10.4	1.9	9.4	3.3	6.9	7.9	51.5

35. Draw histogram for the following frequency distribution.

Variable	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency	12	30	35	65	45	25	18

36. Represent the adjoining distribution of marks of 100 students in the examination by a histogram.

Marks obtained	No. of students
Less than 10	4
Less than 20	6
Less than 30	24
Less than 40	46
Less than 50	67
Less than 60	86
Less than 70	96
Less than 80	99
Less than 90	100

(B.Com. MKU, MSU, CHU)

37. Represent the following data by means of a histogram.

Weekly Wages ('00 Rs.)	10–15	15–20	20–25	25–30	30–40	40–60	60–80
No. of Workers	7	19	27	15	12	12	8

(B.B.A. CHU, BDU, BU)

38. The following data show the number of accidents sustained by 313 drivers of a public utility company over a period of 5 years.

No. of Accidents	0	1	2	3	4	5	6	7	8	9	10	11
No. of Drivers	80	44	68	41	25	20	13	7	5	4	3	2

Draw a frequency polygon.

39. The following table gives the frequency distribution of the weekly wages (in '00 Rs.) of 100 workers in a factory.

Weekly Wages ('00 Rs.)	20–25	30–35	35–40	40–45	45–50	50–55	55–60	Total	
No. of Workers	24	29	34	39	44	49	54	59	64

40. Draw a histogram and frequency polygon for the following frequency distribution.

Mid-value of class interval	2.5	7.5	12.5	17.5	22.5	27.5	32.5	37.5
Frequency	7	10	20	13	17	10	14	9

(B.Com. MSU, CHU, BDU)

94 Business Statistics

41. Draw a frequency curve for the following distribution.

Age (years)	17–19	19–21	21–23	23–25	25–27	27–29	29–31
No. of Students	7	13	24	30	22	15	6

42. Draw a less than cumulative frequency curve for the following data and find from the graph the value of seventh decile.

Monthly Income	0–100	100–200	200–300	300–400	400–500	500–600	600–700	700–800	800–900
	100	200	300	400	500	600	700	800	900

No. of Workers	12	28	35	65	30	20	20	17	13	10
-----------------------	----	----	----	----	----	----	----	----	----	----

(B.B.A. CHU, BDU, BU)

43. Convert the following distribution into 'more than' frequency distribution.

Weekly Wages less than ('00 Rs.)	20	40	60	80	100
No. of Workers	41	92	156	194	201

For the data given above, draw 'less than' and 'more than' gives and hence find the value of median.

(B.Com. MKU, CHU, BDU)

44. Draw the graph of the following:

Year	1990	1991	1992	1993	1994	1995	1996	1997
-------------	------	------	------	------	------	------	------	------

Year	1990	1991	1992	1993	1994	1995	1996	1997
Yield (in million tons)	12.8	13.9	12.8	13.9	13.4	6.5	2.9	14.8

45. Plot a graph to represent the following data in a suitable manner.

Year	1990	1991	1992	1993	1994	1995	1996	1997
-------------	------	------	------	------	------	------	------	------

Imports (in million tons)	400	450	560	620	580	460	500	540
----------------------------------	-----	-----	-----	-----	-----	-----	-----	-----

Imports (million Rs.)	220	235	385	420	420	380	360	400
------------------------------	-----	-----	-----	-----	-----	-----	-----	-----

(B.Com. BDS, BU, MKU)

46. The following table gives the cost of production (in arbitrary units) of a factory in brennial averages.

Items	1988–89	1989–90	1990–91	1991–92	1992–93	1993–94	1994–95	1995–96	1996–97	1997–98
Material	37	25	35	36	35	38	22	17	26	20
Labour	10	8	11	11	11	12	7	5	8	9
Overhead	13	10	15	16	17	20	12	9	12	15
Total	60	43	61	63	63	70	41	31	46	44

47. Marks obtained by 50 students in a History paper of full marks 100 are as follows:

78	25	25	40	30	29	35	42	43	43
44	20	48	44	43	48	36	46	48	47
36	60	31	47	33	65	68	73	39	12
60	20	47	49	51	38	49	35	52	61
34	76	79	20	16	70	65	39	60	45

Arrange the data in a frequency distribution table in class intervals of length 5 units and draw a histogram to present the above data.

(B.Com. MSU, BDU, CHU)

48. (a) Draw the 'less than' and 'more than' olive curves from the data given below:

Weekly Wages less than ('00 Rs.)	0-20	20-40	40-60	60-80	80-100
No. of Workers	10	20	40	20	10

- (b) For the following distribution of wages, draw olive and hence find the value of median.

Monthly Wages	Frequency
12.5-17.5	2
17.5-22.5	22
22.5-27.5	10
27.5-32.5	14
32.5-37.5	3
37.5-42.5	4
42.5-47.5	6
47.5-52.5	1
52.5-57.5	1
Total	63

(B.Com. BU, MSU, CHU)

- (c) Below is given the frequency distribution of marks in Mathematics obtained by 100 students in a class.

Marks	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
No. of Students	7	11	24	32	9	14	2	1

49. The limits for the central 60% of the distribution from the graph.

X(less than)	5	10	15	20	25	30	35	40	45
Frequency	2	11	29	45	69	83	90	96	100

50. Represent the following data by means of a time series graph.

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Imports (Rs.'000)	167	269	263	275	270	280	282	272	265	266
Imports (Rs. '000)	307	310	280	260	275	271	280	280	260	265

Show also the net balance of trade.

(B.Com. MKU, MSU, CHU)

51. Present the following data about India by a suitable graph.

PRODUCTION IN MILLION TONS

Year	Rice	Wheat	Pulses	Other Cereals	Total
1962	30	10	10	14	64
1963	32	11	8	18	69

Year	Rice	Wheat	Pulses	Other Cereals	Total
1964	33	8.5	11.5	20	73
1965	35	12	11	20	78
1966	36	10	10	22	78
1967	38	11	9	23	81

52. Plot the following data graphically on the logarithmic scale.

Year	Total Notes Issued (in crores Rs.)	Total Notes in circulation (in crores Rs.)
2001–02	2890	2866
2002–03	3065	3020
2003–04	3242	3194
2004–05	3536	3497
2005–06	3866	3843

53. Present the following data graphically and comment on the features thus revealed.

Year	Production of steel plates (in thousand tons)	
	Unit A	Unit B
1990	30	40
1992	29	20
1994	31	10
1996	30	20
1998	30	30
2000	30	40
2002	30	60

How will the graph look like if the data are plotted on semi-logarithmic scale?

(B.Com. CHU, BDU, MKU)

5

CHAPTER

SAMPLING

5.1 INTRODUCTION

When secondary data are not available for the problem under study, a decision may be taken to collect primary data. These data may be obtained by following either population method (census method) or sample method.

5.2 POPULATION METHOD OR CENSUS METHOD

Population represents the whole area. Under the census or complete enumeration survey method, data are collected for each and every unit of the population or universe which is the complete set of items.

For example, if the average wages of workers in sugar industry in India is to be calculated, then it would be obtained from each and every worker working in the sugar industry. Then it is dividing the total wages received by the number of workers working in sugar industry. It is the average wages.

According to Simpson and Kafka, “Population or a universe means an aggregate of items possessing a common trait or traits”.

The population can be classified into two types: They are (a) Finite or Infinite (b) Real or Hypothetical.

5.2.1 Finite or Infinite Population

If the number of elements can be counted in the population, then it is called finite population.

For example, number of students in a college, number of people in a village, number of schools in a city. If the number of elements cannot be counted in the population, then it is called infinite population. For example, number of stars in the sky, number of viewers in TV Programmes, number of readers in a newspaper.

5.2.2 Real or Hypothetical Population

In real population the elements or items really exist.

For example, number of factories in a district, number of people in a city. In hypothetical population, elements may not really exist; it is also known as theoretical population. For example, tossing a coin or throwing a dice.

5.2.3 Merits of Census Method

Following are the merits of census method:

1. Data are obtained from each and every unit of the population.
2. Data results are more representative, accurate and reliable.
3. Data of complete enumeration census can be widely used as a basis for various surveys.
4. The characters of all the limits of the population can be studied.
5. Intensive study is possible.

5.2.4 Demerits

The following are the demerits of census method:

1. It requires more time and cost.
2. It is not required for all types of research.
3. Inefficient and inexperienced researcher may collect wrong information.

5.3 SAMPLING

Sampling is the process of learning about the population on the basis of a sample drawn from it. Thus, in the sampling technique instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire universe.

A sample is a subset of population units. The process of sampling involves three elements:

1. Selecting the sample
2. Collecting the information and
3. Making an inference about the population

The basic objective of sample study is to draw inference about the population. In other words, sampling is a tool, which helps to know the characteristics of the universe or population by examining only a small part of it.

5.3.1 Merits of Sample Method

The following are the merits of sample method.

1. The organisation and administration of sample surveys are easy.
2. It reduces cost, time and energy.

3. It gives accurate result.
4. It provides for detailed enquiry.

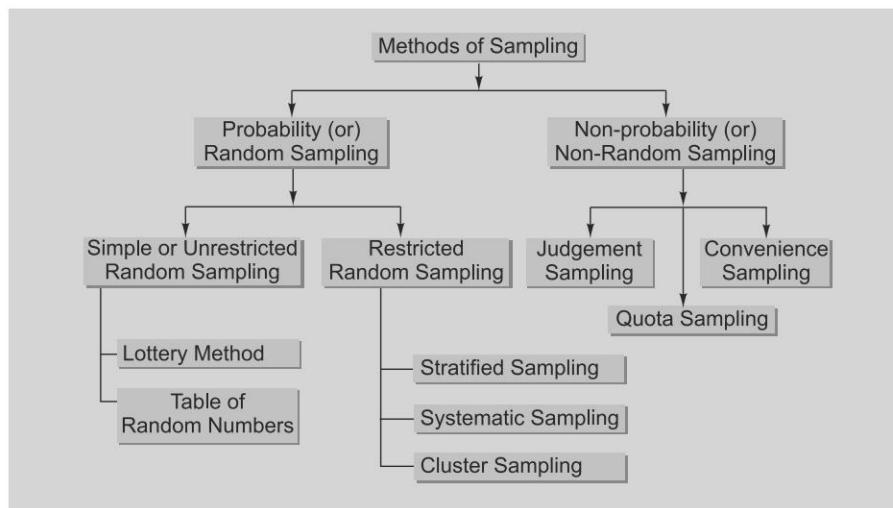
5.3.2 Demerits of Sample Method

Following are the demerits of sample method:

1. Data are not obtained from each and every unit of the population.
2. Reliable results are not possible.

5.4 METHODS OF SAMPLING

Statistical data are collected under sampling through the various methods. They are tabulated as follows:



5.4.1 Probability (or) Random Sampling

It is a method in which each item of the universe has the equal chance of being selected. This implies that the selection of sample items is independent of the person making the study.

Random Sampling can be classified into simple or unrestricted random sampling and restricted random sampling.

(a) Simple (or) Unrestricted Random Sampling It is a technique in which sample is so drawn that each and every unit of the population has an equal and independent chance of being included in the sample.

Under simple random sampling, data are collected through lottery method and table of random numbers.

100 Business Statistics

(i) Lottery Method Under this method, all the items of the universe are numbered on separate slips of paper of same size and colour. These slips are folded and put in a box or containers. Then the required number of slips are selected from the box or container.

For example, if we want to take a sample of 10 persons out of a population of 100, the procedure is to write the names of the 100 persons on separate slips of paper, fold those slips, mix them thoroughly and then make a blindfold selection of 10 slips.

(ii) Table of Random Numbers Random number table is commonly used to select the sample, when the size of the population is very large. It is a table of digits which have been generated by a random process.

Several standard tables of random numbers are available, among which the following may be specially mentioned as they have been tested extensively for randomness.

1. Tippett's (1927) random number tables consisting of 41600 random digits grouped into 10,400 sets of four-digit random numbers.
2. Fisher and Yates (1938) table of random numbers consisting of 15000 random digits arranged into 1500 sets of 10-digit random numbers.
3. Kendal and Babington Smith (1939) table of random numbers consisting of 1,00,000 digits grouped into 25000 sets of four-digit random numbers.
4. Rand Corporation (1955) table of random numbers consisting of 1,00,000 random digits grouped into 20,000 sets of five-digit random numbers.

It is important that the starting point in the table of random numbers be selected in some random fashion so that every unit has an equal chance of being selected.

Merits of Simple Random Sampling It has the following advantages:

1. There is no possibility of personal bias since the selection of items in a sample depends entirely on chance.
2. It represents the universe in a better way. As the size of the sample increases, it becomes increasingly representative of the population.
3. It provides the most reliable information at the least cost.

Limitations of Simple Random Sampling This method is however associated with the following limitations:

1. It may produce the most non-random looking results.
2. The size of the sample is usually larger under random sampling than stratified sampling.
3. Time and cost of collecting data become too large.

Restricted Random Sampling

(i) **Stratified Sampling** Stratified random sampling (or) simply stratified sampling is one of the random methods used to design a more efficient sample than obtained by the simple random procedure. In stratified random sampling, the sampling is designed so that a designated number of items is chosen from each stratum.

The following are the procedure to be followed while applying stratified random sampling technique:

1. The universe to be sampled is subdivided into groups which are mutually exclusive and include all items in the universe.
2. A simple random sample is then chosen independently from each group.

Merits Stratified sampling methods have the following advantages:

1. A more representative sample is secured.
2. Stratified sampling ensures greater accuracy.
3. Stratified samples can be more concentrated geographically.

Demerits Demerits of this method are:

1. The sample may have the effect of bias if proper stratification of the population is not done.
2. In the absence of skilled sampling supervisors, a random selection within each stratum may not be ensured.
3. Cost of observation may be quite high.

(ii) **Systematic Sampling (or) Quasi-Random Sampling** This method is popularly used in those cases where a complete list of the population from which sample to be drawn is available. The list may be prepared in alphabetical, geographical, numerical or some other order. The items are serially numbered.

It is relatively a simple technique and may be more efficient statistically. In this method, the first item is selected at random generally by following the lottery method. Subsequent items are selected by taking every K^{th} item from the list, where K refers to the sampling interval or sampling ratio symbolically.

$$K = \frac{N}{n}$$

where, K = Sampling Interval

N = Universe Size

n = Sample Size

Merits The merits of systematic sampling are:

- The systematic sampling design is simple and convenient to adopt.
- The time and work involved in sampling are relatively less.

Demerits The limitations of systematic sampling are:

1. It is less representative if we are dealing with populations having ‘hidden periodicities’.
2. Personal bias of investigators.

(iii) Multi-Stage Sampling (or) Cluster Sampling Under this method, the random selection is made of primary, intermediate and final units from a given population or stratum. It is a method of sampling which is carried out in several stages.

For example, if we want to take a sample of 5000 households from Tamil Nadu state, first, the state may be divided into number of districts and a few districts may be subdivided into a number of villages and a sample of villages may be taken randomly. At the third stage, a number of households may be selected from each of the villages selected at the second stage.

Merits Cluster sampling has its own advantages. They are:

1. It introduces flexibility in the sampling method.
2. Large area may be covered.

Demerits The demerits of cluster sampling are:

1. It is less accurate than a sample containing the same number of final stage units which have been selected by some single stage process.
2. It is a long process.

5.4.2 Non-Probability Sampling Methods

Non-Probability sampling methods can be classified into three types. They are:

- (a) Judgement Sampling
- (b) Quota Sampling
- (c) Convenience Sampling

(a) Judgement Sampling (or) Purposive (or) Deliberate Sampling In this method of sampling, the choice of sample items depends exclusively on the judgement of the investigator. In other words, the investigator exercises his judgement in the choice and includes those items in the sample, which is most typical of the universe with respect to its characteristics under investigation.

For example, if a sample of 10 students is to be selected from 60, for analysing the habits of students, the investigator would select 10 students, who in his opinion, are representative of the class.

(b) Quota Sampling In non-probability category, it is a commonly used sampling technique. It is like the stratified sampling. The population is divided into various quotas with respect to some common character. Then the required sample is to be selected from these quota.

Quota sampling is often used in public opinion studies. In quota sampling, the sampling within each cell is not done at random, the field representatives are given wide latitude in the selection of respondents to meet their quotas.

If the interviewers are carefully trained and if they follow their instructions closely, then satisfactory results can be achieved.

(c) Convenience Sampling A convenience sample is obtained by selecting 'convenient' population units. The method of convenience sampling is also called the chunk. A chunk refers to that fraction of the population being investigated which is selected neither by probability nor by judgement but by convenience.

A sample may be obtained from readily available lists such as automobile registrations, telephone directories etc., is a convenience sample and not a random sample even if the sample is drawn at random from the lists. However, convenience sampling is often used for making pilot studies. Questions may be tested and preliminary information may be obtained by the chunk before the final sampling design is decided upon.

5.5 THEORETICAL BASIS OF SAMPLING

On the basis of sample study, we can predict and generalise the behaviour of mass phenomena. This is possible because there is no statistical population whose elements would vary from each other without limit.

Theory of sampling is based on two important laws. They are:

1. Law of statistical regularity
2. Law of inertia of large numbers

5.5.1 Law of Statistical Regularity

It is derived from the mathematical theory of probability. This law points out that if a sample is taken at random from a population, it is taken at random from a population and it is likely to possess almost the same characteristics as that of the population. This law makes the desirability of choosing the sample at random.

By random selection, we mean a selection where each and every item of the population has an equal chance of being selected in the sample. A sample selected in this manner would be representative of the population. If it is satisfied, it is possible to depict fairly accurately the characteristics of the population by studying only a part of it. Hence, this law is of great practical significance because it makes possible a considerable reduction of the work before any conclusion is drawn regarding a large universe.

The results derived from sample data to be different from that of population. This is for the simple reason that the sample is only a part of whole universe. However, there would not be much difference in the results derived if the sample is representative of the universe.

5.5.2 Law of Inertia of Large Numbers

It is a law of statistical regularity. It is of great importance in the theory of sampling; it states that, other things being equal, larger the size of the sample, more accurate the results are likely to be. This is because large numbers are more stable as compared to small ones.

For example, if a coin is tossed 10 times, we should expect equal number of heads and tails. But since the experiment is tried a small number of times, it is likely that we may not get exactly 5 heads and 5 tails.

If the same experiment is carried out 1000 times, the chance of 500 heads and 500 tails would be very high. This is because, that the experiment has been carried out a sufficiently large number of times and possibility of variation in one direction compensating for others in a different direction is greater.

Similarly, if it is intended to study the variation in production of rice over a number of years and data are collected from one or two states only, the result would reflect large variations in production due to the favourable factors in operation.

5.5.3 Essentials of Sampling

It is necessary that a sample possesses the following essentials:

1. **Representativeness** A sample should be so selected that it truly represents the universe to get accurate results.
2. **Adequacy** The size of sample should be adequate to represent the characteristics of the universe.
3. **Independence** All items of the sample should be selected independently of one another and all items of the universe should have the same chance of being selected in the sample.
4. **Homogeneity** If two samples from the same universe are taken, they should give more or less the same unit.

5.5.4 Size of Sample

While adopting a sampling technique, it is important to decide about the size of the sample. It means the number of sampling units selected from the population for investigation.

If the size of sample is small, it may not represent the universe and the inference drawn about the population may be misleading. On the other hand, if the size of sample is very large, it requires a lot of time and will lead to problems of managing too. Hence, the sample size should be optimum.

Factors to be Considered While deciding on the sample size, the following factors should be considered:

1. **Size of the Universe:** The larger the size of the universe, the bigger should be the sample size.
2. **Availability of the Resources:** If the resources available are vast, a larger sample size could be taken.
3. **Degree of Accuracy:** The greater the degree of accuracy desired, the larger should be the sample size. It does not mean that bigger sample always ensure greater accuracy. If a sample is selected by experts, it may ensure better results even if the sample is small.
4. **Homogeneity or Heterogeneity of the Universe:** If the universe consists of homogeneous units, a small sample may serve the purpose but if the universe consists of heterogeneous units, a large sample may be inevitable.
5. **Nature of Study:** For an intensive and continuous study, a small sample may be studied. If it should be quite extensive in nature, then take larger sample size.
6. **Method of Sampling Adopted:** The methods of sampling should be adopted according to the size of the sample.
7. **Nature of Respondents:** The size of the sample should be selected on the basis of the nature of respondents.

5.6 SAMPLING AND NON-SAMPLING ERRORS

In statistics, the word ‘error’ is used to denote the difference between the true value and the estimated value. The term error should be distinguished from mistake or inaccuracies which may be committed in the course of making observations, counting calculations etc.

The error, mainly arise at the stage of ascertainment and processing of data in complete enumeration and sample surveys.

Errors are classified into sampling errors and non-sampling errors.

5.6.1 Sampling Errors (or) Sampling Fluctuations

The difference in value of the sample and population is called sampling errors. Only few variables are selected from the whole population, there is more chance for sampling errors. These errors can be minimised through the modern sampling theories.

In the words of Patterson, “Sampling error is the difference between the result of studying a sample and inferring a result about the population and the result of the census of the whole population.”

Sampling errors are of two types: They are: (a) Biased error (b) Unbiased error.

(a) Biased Error These errors arise from any bias in selection, estimation etc. It may occur more in deliberate sampling method than in random sampling method.

(b) Unbiased Error These errors arise due to chance differences between the members of population included in the sample and those not included. In the unbiased selection of the sample, the result may vary with the actual result of the population. These differences are called unbiased error.

Measurement of Errors Statistical errors are of two types. They are:
(a) Absolute error and (b) Relative error.

(a) Absolute Error It refers to the difference between the approximated figure and the original figure.

$$\text{Absolute error} = \text{Actual value} - \text{Estimated value}$$

$$A.E = a - e$$

In a college, the average estimated height of student is 160 cm and the actual average height is 162 cm.

Hence, the absolute error = $162 - 160 = 2$ cm.

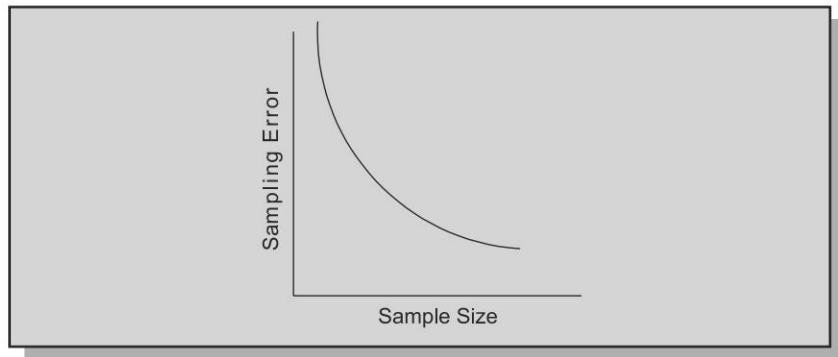
(b) Relative Error It refers to the ratio of absolute error to the estimated value.

$$R.E = \frac{a - e}{E}$$

For example, If the actual value is 1250 and the estimated value is 1200,

$$\begin{aligned} \text{Then the relative error} &= \frac{1250 - 1200}{1200} \\ &= \frac{50}{1200} = 0.042 \end{aligned}$$

Methods of Reducing Sampling Errors Once the absence of bias has been ensured, attention should be given to the random sampling errors. Such errors may be reduced to the minimum to attain the desired accuracy. Apart from reducing errors of bias, the simplest way of increasing the accuracy of a sample is to increase its size.



From this diagram, it is clear that the sampling error is very large when the size of the sample is small and the sampling error is small when the size of the sample is large.

SUMMARY

Population or universe

- In a statistical enquiry, all the items which fall within purview of enquiry.
or
- It is a set of all possible observations of the type which is to be investigated.

Finite population

When the number of observations can be counted and definite.

Infinite population

When the number of observations can't be measured on number and is infinite.

Hypothetical population or Theoretical population

It is one which does not consist of concrete objects. It exists only in the expectation. We can't count things.

Existential or real population

It is one that actually exists. It is a universe which contains persons or concrete objects.

Census Method

A complete enumeration of all items in the population. Every element of the population is included in the investigation.

Essentials of Sampling

- It must be representative.
- It must be adequate.
- It should be independent.
- It should be homogeneous.

Methods of Sampling

Random sampling method or Probability sampling

- (a) Simple or unrestricted Random sampling
- (b) Restricted Random sampling
 - Stratified sampling
 - Systematic sampling
 - Cluster sampling

Non-Random Sampling or Non-Probability Sampling

- (a) Judgement or purposive sampling
- (b) Quota sampling
- (c) Convenience sampling

Random Sampling

It is one where each item in the universe has an equal chance of known opportunity of being selected.

Simple Random Sampling

It is a technique in which each and every unit in the population has an equal and independent chance of being included in the sample.

Stratified Sampling

When the population is heterogeneous or of different segments or strata with respect to its characteristics, then it is stratified.

Systematic Sampling

It is used when a complete list of the population is available. The items should be arranged in numerical, alphabetical, geographical or any other order.

Cluster Sampling or Multi-Stage Sampling or Sampling Stages

It refers to a sampling procedure which is carried out in several procedure which is carried out in several stages. The whole population is divided in to sampling units and these units are again divided into sub-units and well continue till it reaches lease number.

Judgement Sampling or Purposive or Deliberate

In this sampling, the investigator has the power to select or reject any item in an investigation.

Quota Sampling

It is similar to stratified sampling. Universe is divided in to quota according to some characteristics in order to collect data.

Convenience or Chunk Sampling

It is a convenient slice of a population which is commonly referred to as a sample. A sample obtained from automobile registration, telephone directories etc.

Induction

Logical process of drawing general conclusions from a study of representative items.

Fundamental Principles of Statistical Theory

- Law of Statistical Regularity.
- Law of Inertia of Large numbers

Law of Statistical Regularity

Reasonably large number of items selected at random from a large group of items will, on the average be representative of characteristics of the large group of population.

Law of Inertia of large numbers

Large aggregates are more stable than small ones. The law will be true on an average.

Statistical Error

The arithmetical difference between the approximated figure and the original quantity.

Biased Errors

Errors that arise due to a bias or prejudice on the part of the enumerator or investigator in selecting, estimating or measuring instrument.

Unbiased Errors

Errors which arise in the normal course of investigation on account of chance.

Relative Error

Ratio of absolute error to the estimated value.

EXERCISES

(a) Choose the best option.

1. _____ population is the number of observations can be counted.
(a) Infinite (b) Finite (c) Real
2. _____ population is the member of observations cannot be measured.
(a) Finite (b) Infinite (c) Real
3. The universe containing persons is known as _____ population.
(a) Real (b) Finite (c) Infinite
4. Theoretical population is also known as _____ universe.
(a) Existent (b) Hypothetical

5. _____ is possible to intensive study.

- (a) Sample method (b) Census method

Answers

1. b 2. b 3. a 4. b 5. b

(b) Fill in the blanks.

1. _____ is a complete set of all possible observations.

2. Information on population be collected in _____ and _____ method.

3. Random sampling is also known as _____.

4. Non-random sampling is also called as _____.

5. _____ method is the most popular and simplest method.

6. _____ cannot be used lottery method.

7. _____ is otherwise known as multistage sampling.

8. _____ is the investigator has the power to select or reject any item in an investigation.

9. _____ sampling is called as judgement sampling.

10. Deliberate sampling is called as _____ sampling.

11. Quota sampling is similar to _____ sampling.

Answers

- | | |
|---------------------|----------------------------|
| 1. Population | 2. Census, sample |
| 3. Probability | 4. Non-probability |
| 5. lottery | 6. Table of Random numbers |
| 7. cluster sampling | 8. Judgement sampling |
| 9. purposive | 10. Judgement |
| 11. stratified | |

(c) Theoretical Questions

1. Distinguish between the census and sampling methods of collecting data and compare their merits and demerits.

(B.Com., MKU, MSU, CHU)

2. Explain the importance and significance of sample.

3. Point out the difference between a sample survey and census survey. Under what conditions are these undertaken? Explain the law which forms the basis of sampling.

(B.Com., MSU, BDU, BU)

4. Distinguish between a 'Census' and a 'Sample' enquiry and discuss briefly their comparative advantages. Explain the conditions under which each of these methods may be used with advantage.

5. Give a comparative account of the various methods of selecting a sample.
(B.Com., BDU, BU, MKU)
6. Suppose you are asked to conduct a survey of the family expenditure of the Delhi University teachers. How will you proceed?
7. What is random sampling? How can a random sample be selected? Is random sampling always better than other forms of sampling in the context of socio-economic surveys?
(B.Com., CHU, MSU, BU)
8. Write short notes of the following:
 - (a) Stratified Sampling (b) Random Sampling
 - (c) Hypothetical Population (d) Infinite Population
9. Describe the following and give their relative merits and demerits:
 1. Judgement or Purposive Sampling
 2. Cluster Sampling
 3. Multistage Sampling
 4. Quota Sampling
10. Explain the merits of sampling technique over the census technique.
(B.Com., MKU, MSU, BU)
11. Enumerate the various steps in conducting a Sample Survey.
(B.Com., BU, CHU, BDU)
12. Describe the method of conducting Sample Surveys.
13. Distinguish between Sampling and Census method of Survey.
14. Write a short notes on:
 1. Law of Statistical Regularity
 2. Law of Inertia of Large Numbers
(B.Com., MKU, MSU, BDU)
15. Explain the terms: Sampling unit, Sampling frame and Sampling errors.
(B.Com., CHU, BU, MKU)

6

CHAPTER

MEASURES OF CENTRAL TENDENCY

6.1 INTRODUCTION

The word average is one of the most commonly used in day-to-day conversation. For example, we often talk of average strength of a class, average score in cricket, average height, average income, average life of an Indian. Average is a single value that represents a group of values. Averages are, generally the central part of the distribution and therefore, they are also called measures of central tendency. Averages are the values which lie between the two extreme observations (i.e., the smallest and the largest observations) of the distribution.

Averages are very much useful—

- (i) for describing the distribution in concise manner;
- (ii) for comparative study of different Distribution, and
- (iii) for computing various other statistical measures such as Dispersion, Skewness and kurtosis.

6.2 DEFINITIONS

The following are the important definitions:

Average is an attempt to find one single figure to describe whole of figures.
— Clark

Average is a value which is typical or representative of a set of data.
— Muny R. Speigal

An average is a single number describing some features of a set of data.
— Wallis and Roberts

An average is a typical value in the sense that it is sometime employed to represent all the individual values in a series or of variable.
— Ya-Lun-Chou

An average value is a single value within the range of data that is used to represent all of the values in the series. Since an average is somewhere within the range of data, it is also called a measure of central value.

— Croxton and Cowton

From the above definitions, average means a representative value of a group of data. It represents the character of the entire data.

6.2.1 Objectives of Average

The following are the important objectives of the measure of central tendency.

- (i) To get a single value for the entire group
The single value that describes the characteristics of the entire group. This one value may represent thousands or millions of values. From single value, one can get an idea about the entire data.
- (ii) To facilitate comparison
Measures of central tendency, by reducing the volume of data to one single figure, enable comparison to made. Comparison can be made either at a point of time or over a period of time. For example, average height of B.Com. students can be compared with average height of B.Sc. students.
- (iii) To take policy decisions
Policy decisions can be taken if the data are presented in one single value.

6.2.2 Characteristics of a Good Average or Requisites of a Good Average

The following are important characteristics and properties of a good average.

- 1. Easy to understand:** The single value should not be a complicated one. It should be readily comprehensible and should be computed with sufficient ease.
- 2. It should be simple to calculate:** The procedure adopted for calculating the averages should be simple. It should not involve heavy arithmetical calculation.
- 3. It should be rigidly defined:** The definition should be clear and unambiguous so that it leads to one and only interpretation by the different persons.
- 4. It should be based on all the observations:** In the computation of ideal average, the entire set of data should be used and there should not be any omission of information.
- 5. It should not be affected much by extreme observations:** Sometimes the observations may contain either extremely low value or high value or both of them. This will affect the reliability of the single value as a true representative item. So, proper descriptions should be given regarding the extreme values while presenting the averages.
- 6. Capable of further algebraic treatment:** Averages should be eligible for further algebraic treatment which is for further analysis so that its utility is enhanced.
- 7. Sampling stability:** The sample selected for the study should be capable of giving same data when approached by different group of people.

6.3 TYPES OF AVERAGES

The following are the important types of averages:

1. Arithmetic mean
2. Geometric mean
3. Harmonic mean
4. Median
5. Mode

6.3.1 Arithmetic Mean

The most popular and widely used measure of central tendency is Arithmetic mean. It is called average. It can be obtained by dividing the sum of all the observations by the total number of observations.

There are two types of arithmetic mean. They are:

- (i) Simple Arithmetic mean (ii) Weighted Arithmetic mean

A. Simple Arithmetic Mean—Individual Observations

I. Direct Method

$$\text{Arithmetic mean} = \frac{\text{Sum of all the values}}{\text{Number of observations}}$$

$$\bar{x} = \frac{\sum x}{N}$$

where \bar{x} represents Arithmetic mean

Σx represents sum of all the values

N represents number of observations.

Illustration 6.1

Calculate Arithmetic mean of heights of 10 students in a B.Com. class.

Roll No.	1	2	3	4	5	6	7	8	9	10
Height cms	160	157	162	155	158	159	161	154	152	156

Solutions

Calculation of Arithmetic Height

Roll No.	Heights (in cms)
1	160
2	157
3	162
4	155
5	158
6	159

Contd.

Roll No.	Heights (in cms)
7	161
8	154
9	152
10	156
N = 10	$\Sigma x = 1574$

$$\text{Arithmetic mean } \bar{x} = \frac{\sum x}{N} = \frac{1574}{10} \\ = 157.4 \text{ cms.}$$

2. Short-cut Method When the actual mean is in fraction, this method is used. It is based on assumed mean. Any value can be taken as assumed mean but it should be closely related with the data. Deviations are taken from the assumed mean. The formula is

$$\bar{X} = A \pm \frac{\sum d}{N}$$

where,

\bar{X} = Arithmetic mean

A = Assumed mean

D = Deviations of items from the Assumed mean ($x-A$)

N = Number of observations.

Illustration 6.2

Marks obtaining 10 students in statistics are given below.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks	46	64	55	47	74	78	88	90	59	63

Calculate Arithmetic mean by short-cut method.

Solutions

Calculation of Arithmetic mean

Roll No.	Marks X	Deviations $\Sigma d = x - A$
1	46	-28
2	64	-10
3	55	-19
4	47	-27
5	74	0
6	78	4
7	88	14
8	90	16
9	59	-15
10	63	-11
N = 10		$\Sigma d = -76$

$$\begin{aligned}\text{Arithmetic mean } \bar{x} &= A \pm \frac{\sum d}{N} = 74 + \frac{-76}{10} \\ &= 74 - 7.6 \\ &= 66.4\end{aligned}$$

B. Simple Arithmetic Mean—Discrete Series

1. Direct Method Arithmetic mean can be calculated through direct method by applying the following formula,

$$\bar{x} = \frac{\sum fx}{N}$$

where,
 f stands for frequency
 x stands for variables
 N stands for total number of observations.

Illustration 6.3

Calculate the mean value of income per month of families in a town from the following data.

Income (Rs.'000)	25	26	27	28	29	30
No. of families	3	6	9	14	10	8

Solutions

Calculation of Arithmetic mean

Income Per day Rs. (x)	No. of families f	fx
25	3	75
26	6	156
27	9	243
28	14	392
29	10	290
30	8	240
$N = 50$		$\Sigma fx = 1396$

$$\begin{aligned}\text{Arithmetic mean } \bar{x} &= \frac{\sum fx}{N} = \frac{1396}{50} \\ &= 27.92\end{aligned}$$

Ans: The mean value of income of the families in the town is 27.92.

2. Short-cut Method This method is an easy way by calculating the arithmetic mean.

Steps

1. Take any value as assumed mean (A).
2. Find out deviations of each variable from the assumed mean (d).
3. Multiply the deviation with the respective frequencies (fd).
4. Add up the products (Σfd).
5. Apply the formula—

$$\bar{x} = A \pm \frac{\sum fd}{N}$$

x = mean

A = Assumed mean

Σfd = sum of total deviations

N = Total frequency

Illustration 6.4

Marks obtained by students of a class in statistics are given below.

Marks	35	40	45	50	55	60	65	70	75	80
No. of Students	13	18	22	19	14	17	25	15	20	17

Calculate Arithmetic mean by short-cut method.

Solutions

Calculation of Arithmetic mean through short-cut method

Marks (x)	No. of students f	(A = 65)	$d = x - 65$	fd
35	13		-30	-390
40	18		-25	-450
45	22		-20	-440
50	19		-15	-285
55	14		-10	-140
60	17		-5	-85
65	25		0	0
70	15		5	75
75	20		10	200
80	17		15	255
$N = 180$			$\Sigma fd = -1260$	

$$\begin{aligned}\bar{x} &= A \pm \frac{\sum fd}{N} \\ &= 65 - \frac{1260}{180} = 65 - 7\end{aligned}$$

$$\bar{x} = 58$$

Ans. The Mean marks obtained by 180 students in statistics is 58.

C. Simple Arithmetic Mean—Continuous Series

I. Direct Method In continuous series, variables are represented by class intervals. Each class interval has its own frequency.

The following procedure is to be adopted for calculating arithmetic mean in a continuous series.

1. Find out the mid-value of each group or class. The mid-value is obtained by adding the lower limit and upper limit of the class and dividing the total by two. For example, in a class interval, say 20–30, the mid-value is 25 ($20 + 30/2 = 50/2 = 25$). Its symbol is m .
2. Multiply the mid-value of each class by the frequency of the class. In other words, m will be-multiplied by f .
3. Add up all the products—(Σfm).
4. Σfm is divided by N .

Apply the formula:

$$\bar{x} = \frac{\sum fm}{N}$$

where

f = frequency

m = mid-point; and

N = total no. of frequencies ($N = \Sigma f$)

Illustration 6.5

Calculate Arithmetic mean from the following data.

Production in tonnes	No. of factories
20–30	15
30–40	14
40–50	17
50–60	22
60–70	20
70–80	18
80–90	14

Solutions

Calculation of Arithmetic mean

Production in tonnes X	No. of factories	Mid-value	Fm
20–30	15	25	375
30–40	14	35	490
40–50	17	45	765
50–60	22	55	1210
60–70	20	65	1300
70–80	18	75	1350

80–90	14	85	1190
N = 120			$\Sigma fm = 6680$

$$\bar{x} = \frac{\Sigma fm}{N} = \frac{6680}{120}$$

$$\bar{x} = 55.7$$

Ans. The mean production of the factories is 55.7 tonnes.

2. Short-cut Method This method is based on assumed arithmetic mean. Assumed arithmetic mean may be taken either from the mid-points of the class intervals or not.

Steps

1. Find the mid-value of each class or group – (m).
2. Assume any one of the mid-value as an average (A).
3. Find out the deviations of the mid-value of each from the assumed mean (d).
4. Multiply the deviations of each class by its frequency (fd).
5. Add up the products of step 4 (Σfd).
6. Apply the formula—

$$\bar{x} = A \pm \frac{\Sigma fd}{N}$$

where

A = Assumed mean

Σfd = Sum of total deviations

N = Number of items

Illustration 6.6

Calculate Arithmetic mean through short-cut method from the following data.

Income Per day Rs.	No. of families	Income Per day Rs.	No. of families
10–20	16	60–70	14
20–30	22	70–80	21
30–40	25	80–90	19
40–50	20	90–100	24
50–60	18	100–110	21

Solutions

Calculation of Arithmetic mean through short-cut method

Income per day Rs. x	No. of families f	Mid pts m	$D = m - 55$ ($A = -55$)	fd
10–20	16	15	-40	-640
20–30	22	25	-30	-660

Income Per day Rs. x	No. of families f	Mid pts m	D = m - 55 (A = - 55)	fd
30–40	25	35	- 20	- 500
40–50	20	45	- 10	- 200
50–60	18	55	0	0
60–70	14	65	10	140
70–80	21	75	20	420
80–90	19	85	30	570
90–100	24	95	40	960
100–110	21	105	50	1050
$\Sigma f = 200$				$\Sigma fd = 1140$

$$\begin{aligned}\bar{x} &= A \pm \frac{\sum fd}{N} \\ &= 55 + \frac{1140}{200} \\ &= 55 + 5.7 \\ \bar{x} &= 60.7\end{aligned}$$

Ans. The Mean income of the families is Rs. 60.7.

3. Step Deviation Method Step Deviation method is used to simplify the calculation of arithmetic mean for continuous series.

Steps

- Find out the mid-value of each class or group (m).
- Assume any one of the mid-value as an average (A).
- Find out the deviations of the mid-value of each from the assumed mean (d).
- Deviations are divided by a common factor (c).
- Multiply the d' of each class by its frequency (fd').
- Add up the products of above step (5) – ($\Sigma fd'$).
- Then apply the formula—

$$\bar{x} = A \pm \frac{\sum fd'}{N} \times c$$

where,

\bar{x} = Mean

A = Assumed mean

$\Sigma fd'$ = Sum of the deviations

N = Number of items

C = Common factor

Illustration 6.7

The following are the data related with the production of a product during May in 10 factories:

Production in factories	No. of factories
0–10	17
10–20	25
20–30	20
30–40	19
40–50	20
50–60	22
60–70	18
70–80	23
80–90	19
90–100	17

Calculate Arithmetic mean through step deviation method.

Solutions

Production in tonnes	No. of factories	m	$d' = \frac{(m - a)}{c}$	fd'
x	f			
0 – 10	17	5	-5	-85
10 – 20	25	15	-4	-100
20 – 30	20	25	-3	-60
30 – 40	19	35	-2	-38
40 – 50	20	45	-1	-20
50 – 60	22	55	0	0
60 – 70	18	65	1	18
70 – 80	23	75	2	46
80 – 90	19	85	3	57
90 – 100	17	95	4	68
$N = 200$		$\Sigma fd' = -114$		

$$\bar{x} = A + \frac{\sum fd'}{N} \times c$$

$$\bar{x} = 55 - \frac{114}{200} \times 10 = 55 - 5.7$$

$$\bar{x} = 49.3$$

Hence, the mean production of the product in May for 10 factories is 49.3 tonnes.

Step Deviation method can also be applied in the direct method of calculating arithmetic mean of continuous series. The following formula can be applied for this purpose.

$$\bar{x} = \frac{\sum fm'}{N} \times C$$

where

f = frequency

m' = mid points divided by common factor

$(m' = m/c)$ = common factor

Illustration 6.8

Calculate Arithmetic mean from the following data.

Income per day (Rs.)	No. of employees
0–100	14
100–200	13
200–300	16
300–400	21
400–500	17
500–600	12
600–700	12

Solutions

Calculation of Arithmetic mean

Income	No. of employees <i>f</i>	<i>m</i>	<i>m/50</i> <i>m'</i>	<i>fm'</i>
0–100	14	50	1	14
100–200	13	150	3	39
200–300	16	250	5	80
300–400	21	350	7	147
400–500	17	450	9	153
500–600	12	550	11	132
600–700	12	650	13	156
<i>N = 105</i>				$\Sigma fm' = 721$

$$\begin{aligned}\bar{x} &= \frac{\sum fm'}{N} \times C \\ &= \frac{721}{105} \times 50 \\ &= 343.33\end{aligned}$$

Ans. The average income is Rs. 343.33.

Illustration 6.9

The following are the data related with the monthly income of 200 families in a town.

Monthly Income (Rs.)	No. of families
Above 0	200
Above 100	195
Above 200	180
Above 300	175
Above 400	160

Monthly Income (Rs.)	No. of families
Above 500	155
Above 600	140
Above 700	135
Above 800	125
Above 900	120
Above 1000	100

Calculate Arithmetic mean.

Solutions

Calculation of Arithmetic mean

Monthly Income (Rs.)	No. of families	Mid points <i>m</i>	$d' = \frac{(m - a)}{c}$ (A = 450)	fd'
0–100	5	50	-4	-20
100–200	15	150	-3	-45
200–300	5	250	-2	-10
300–400	15	350	-1	-15
400–500	5	450	0	0
500–600	15	550	1	15
600–700	5	650	2	10
700–800	10	750	3	30
800–900	5	850	4	20
900–1000	20	950	5	100
$N = 100$			$\Sigma fd' = -85$	

$$\bar{x} = A \pm \frac{\sum fd'}{N} \times C$$

$$= 450 + \frac{85}{100} \times 100$$

$$= 450 + 85$$

$$\bar{x} = 535$$

Ans. The Mean monthly income of 200 families is Rs. 535.

D. Weighted Arithmetic Mean

Weighted arithmetic mean should be applied to find out the average. In this method, the weights for observations should be given on the basis of their relative importance. The formula under this method is

$$\bar{xw} = \frac{\sum wx}{\sum w}$$

where

\bar{xw} = weighted arithmetic mean

x = value of observations

w = weights for the observations

Illustration 6.10

Calculate the average mark scored by a student in 5 different subjects. The details of the marks are given below.

Tamil 95, English 85, Maths 55, Science 65, Social 80.

The weight given for these subjects are Maths 5, Science 4, Social 3, Tamil 2 and in English 1.

Solutions

Calculation of Weighted Arithmetic mean

Subject	Weight w	Marks X	wX
Maths	5	55	275
Science	4	65	260
Social	3	80	240
Tamil	2	95	190
English	1	85	85
$\Sigma w = 15$		$\Sigma wX = 1050$	

$$\bar{x}_w = \frac{\sum wX}{\sum w} = \frac{1050}{15}$$

$$\bar{x}_w = 70 \text{ marks}$$

Correcting and Incorrect Arithmetic Mean Sometimes the arithmetic mean may be wrongly calculated. This is due to the mistakes in copying, oversight and taking wrong values. It is easy to calculate the correct arithmetic mean from the incorrect one.

Steps

1. The sum of the observations (Σx) deduct the wrong observations.
2. Add the correct observations.
3. Divide the correct sum of observations (Σx) by the number of observations.
4. This will give the correct arithmetic mean.

Illustration 6.11

The arithmetic mean of the sales in 40 factories was calculated as Rs. 2320. Later, it was found out that the sales for two factories were wrongly calculated as Rs. 2540 and Rs. 2260 instead of Rs. 2450 and Rs. 2870.

Solutions

$$\bar{x} = \frac{\sum x}{N}$$

Hence, $\sum x = N\bar{x} = 40 \times 2320 = 92,800$

Less incorrect figure ($2540 + 2260$)

i.e. 4800 from $\sum x$, $92,800 - 4,800 = 88,000$

Add : correct figure ($2450 + 2870$) = 5,320

Corrected $\sum x = 93,320$

$$\text{Correct } \bar{x} = \frac{93,320}{40} = 2,333$$

Illustration 6.12

The arithmetic mean of marks obtained in maths by the students of two classes is 93 and 77 respectively. The number of students in these two classes is 50 and 60 respectively. Calculate the mean marks obtained by 100 students of these two classes.

Solutions

The combined arithmetic mean,

$$\bar{x}_{12} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$$

$$N_1 = 50, N_2 = 60, x_1 = 93, x_2 = 77$$

$$\begin{aligned}\bar{x}_{12} &= \frac{(50 \times 93) + (60 \times 77)}{50 + 60} \\ &= \frac{4650 + 4620}{110} = \frac{9270}{110}\end{aligned}$$

$$\bar{x}_{12} = 84.27$$

Illustration 6.13

The mean wages of 250 workers in a factory is Rs. 95. If the arithmetic mean of 90 workers in one section is Rs. 102 calculate the arithmetic mean of wages for the other section.

Solutions

$$\bar{x}_{12} = 95, x_1 = 102, N_1 = 90, N_2 = 250 - 90 = 160$$

$$\bar{x}_{12} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$$

$$\bar{x}_{12} = \frac{(90 \times 102) + (160 \times \bar{x}_2)}{90 + 160}$$

$$\begin{aligned}
 95 &= \frac{9180 + 160\bar{x}_2}{250} \\
 95 \times 250 &= 9180 + 160\bar{x}_2 \\
 160\bar{x}_2 &= 23750 - 9180 \\
 \bar{x}_2 &= \frac{14570}{160} \\
 \bar{x}_2 &= 91.06
 \end{aligned}$$

Merits of Arithmetic Mean Arithmetic mean is the simplest measurement of Central Tendency of a series. It is widely used because:

- (i) It is easy to understand.
- (ii) It is easy to calculate.
- (iii) It is used in further mathematical analysis.
- (iv) It can be rigidly defined.
- (v) It represents the entire group of data in one single value.
- (vi) It provides a good basis for comparison.
- (vii) It is based on the value of every item in the series.
- (viii) It can be used for further analysis and algebraic treatment.
- (ix) The nature of two or more groups of distribution can be easily compared through arithmetic mean.
- (x) The mean is a more stable measure of central tendency.

Demerits of Arithmetic Mean

- (i) It is difficult to calculate mean for open-end frequency distribution.
- (ii) The qualities like intelligence, efficiency etc. could not be studied in this method.
- (iii) It will not represent the data in a reasonable manner, if the distribution is an abnormal one.
- (iv) The mean is unduly affected by the extreme items.
- (v) It is unrealistic.
- (vi) It may lead to a false conclusion.
- (vii) It cannot be accurately determined if even one of the values is not known.
- (viii) It cannot be located by observation or the graphic method.
- (ix) It gives greater importance to bigger items of a series and lesser importance to smaller items.

E. Geometric Mean

Meaning Geometric mean may be defined as N^{th} root of the product of N observations of a series.

For two observations, square root should be taken; for three observations, third root should be taken. For four observations, fourth root should be taken and so on. It can be represented as,

$$\text{G.M.} = N\sqrt{x_1 + x_2 + x_3 + \cdots + x_n}$$

Calculation of G.M. for individual observations

$$\text{G.M.} = \text{Antilog} \left(\frac{\sum \log x}{N} \right)$$

Geometric mean-Individual Series

Steps

1. Find out the logarithm of each value or the size of item from the log table $-\log x$.
2. Add all the values of $\log x - \sum \log x$.
3. The sum of $\log (\sum \log x)$ is divided by the number of items. $\sum \log x / N$
4. Find out the antilog of the quotient (from step 3). This is the geometric mean of the data.

Illustration 6.14

Calculate geometric mean of the following:

82 72 37 750 976

Solutions

X	log of x
82	1.9138
72	1.8573
37	1.5682
750	2.8751
976	2.9894
$\Sigma \log x = 11.2038$	

$$\text{G.M.} = \text{Antilog} \left(\frac{\sum \log x}{N} \right)$$

$$= \text{Antilog} \frac{11.2038}{5}$$

$$= \text{Antilog } 2.2408$$

$$\text{G.M.} = 174.1$$

Geometric Mean—Discrete Series**Steps**

1. Find out the logarithm of each value — $\log x$.
2. Multiply the log of each size by the frequency — $f \log x$.
3. Add all the products thus we get $\sum f \log x$
4. Divide the total of products by the total frequency (N)

$$\left(\frac{\sum f \log x}{N} \right)$$

5. The Antilog of the step 4 is the result.

$$= \text{Antilog} \left(\frac{\sum f \log x}{N} \right)$$

Illustration 6.15

Calculate geometric mean for the following data.

Production	145	135	149	146	150
No. of factories	6	4	2	3	1

Solutions

Calculation of geometric mean

Production (in tonnes) x	No. of factories	$\log x$	$f \log x$
145	6	2.1614	12.9684
135	4	2.1303	8.5212
149	2	2.1732	4.3464
146	3	2.1644	6.4932
150	1	2.1761	2.1761
$N = 16$		$\sum f \log x = 34.5053$	

$$\begin{aligned}
 \text{G.M.} &= \text{Antilog} \left(\frac{\sum f \log x}{N} \right) \\
 &= \frac{34.5053}{16} = \text{Antilog of } 2.1566 \\
 &= 1434 \\
 &= 143.4 \text{ tonnes.}
 \end{aligned}$$

Geometric Mean—Continuous Series**Steps**

1. Find the mid-value of each class— m .
2. Find the logarithm of the mid-value $\log m$.
3. Multiply the logs of m by their respective frequency $f \log m$.
4. Add up all the products— $\sum f \log m$
5. Divide $f \log m$ by N — $\sum f \log m / N$
6. Find out the antilog of the result of step 5 and this will give the answer.

The formula is : G.M. = Antilog $\frac{\sum f \log m}{N}$

Illustration 6.16

Find out the Geometric mean.

Yield of Wheat	No. of farms
10.5–13.5	9
13.5–16.5	19
16.5–19.5	23
19.5–22.5	7
22.5–25.5	4

Solutions

Calculation of Geometric Mean

Mid-value (maunds)	$\log m$	f	$f \log m$
12	1.0792	9	9.7128
15	1.1761	19	22.3459
18	1.2553	23	28.8719
21	1.3222	7	9.2554
24	1.3802	4	5.5208
$N = 62 \quad \sum f \log m = 75.7068$			

$$\text{G.M.} = \text{Antilog} \frac{\sum f \log m}{N}$$

$$= \frac{75.7068}{62} = 1.2211$$

$$= 1660$$

$$= 166 \text{ maunds}$$

Merits of Geometric Mean

1. It is rigidly defined.
2. It is based on all observations.

130 Business Statistics

3. It can be used for further statistical analysis.
4. Geometric mean can be used for averaging ratios, rates and percentages.
5. It is suitable for calculating index numbers.
6. It is less affected by the extreme values.
7. It is useful in studying economic and social data.
8. It is an average most suitable when large weights have to be given to small items and small weights to the large items.
9. It is capable of further algebraic treatments.

Demerits of Geometric Mean

1. It is difficult to calculate and understand.
2. If any of the value in a series is zero, geometric mean will also be zero. Hence, it cannot be calculated for such kind of series.
3. It cannot be calculated when the distribution has open-end class.
4. Non-mathematical persons cannot do calculations.
5. It has restricted application.

F. Harmonic Mean

Harmonic mean like geometric mean is a measure of central tendency in solving special types of problems. Harmonic mean is the reciprocal of the arithmetic average of the reciprocal of values of varies items in the variable.

Harmonic Mean—Individual Series**Steps**

1. Find out the reciprocal of each size that is, $1/x$ (for easy calculation, refer log table)
2. Add all the reciprocals of all values ($\sum 1/x$)
3. Apply the formula—

$$\text{H.M.} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

or

$$\text{H.M.} = \frac{N}{\sum 1/x}$$

$x_1 + x_2 + x_3 \dots x_n$ refer to the various values in the observations.

Illustration 6.17

Compute Harmonic mean for the following data.

Family	1	2	3	4	5
Income	70	75	42	36	40

Calculate the Harmonic mean.

Solutions

Calculation of Harmonic mean

Family	Income	Reciprocals (1/x)
1	70	.01426
2	75	.01333
3	42	.02318
4	36	.02778
5	40	.02500
N = 5		$\Sigma 1/x = .10355$

$$\text{H.M.} = \frac{N}{1/x_1 + 1/x_2 + 1/x_3 + \dots + 1/x_n}$$

$$\frac{N}{\sum 1/x} = \frac{5}{0.10355} = \text{Rs. } 48.29$$

Harmonic Mean—Discrete Series

Steps

- Find out the reciprocal of each item ($1/x$).
- Multiply the reciprocal ($1/x$) of each size by its frequency—($f 1/x$).
- Add up all the products— $\sum f(1/x)$.
- Apply the formula—

$$\text{H.M.} = \frac{N}{\sum f(1/x)}$$

Illustration 6.18

Calculate Harmonic mean for the following data.

Size of items	1	2	3	4	5
Frequency	9	6	5	8	2

Calculate the Harmonic mean.

Solutions

Calculation of Harmonic mean

Size of items	Frequency	Reciprocals (1/x)	Product of Reciprocal($f 1/x$)
1	9	.1250	1.1250
2	6	.1429	0.8574
3	5	.1111	0.5555
4	8	.0909	0.7272
5	2	.1000	0.2000
$\Sigma f = 30$		$\Sigma f \left(\frac{1}{x}\right) = 3.4651$	

$$\begin{aligned} \text{H.M.} &= \frac{N}{\sum f(1/x)} \\ &= \frac{30}{3.4651} = 8.658 \end{aligned}$$

Harmonic Mean—Continuous Series**Steps**

1. Find out the mid-value of each class m .
2. Find out the reciprocal of each mid-value— $1/m$.
3. Multiply the reciprocal of each mid-value by its frequency $f 1/m$.
4. Add up all the products $\Sigma f 1/m$.
5. Apply the formula—

$$\text{H.M.} = \frac{N}{f_1 1/m_1 + f_2 1/m_2 + \dots + f_n 1/m_n}$$

or

$$= \frac{N}{\sum f 1/m}$$

Illustration 6.19

Calculate Harmonic mean for the following data.

Marks	10–20	20–30	30–40	40–50	50–60
Frequency	8	20	35	45	55

Calculate the Harmonic mean.

Solutions

Calculation of Harmonic mean

Marks	Mid value (m)	Frequency f	Reciprocals ($1/m$)	Product of Reciprocal ($f 1/m$)
10–20	15	8	.06667	0.533
20–30	25	20	.04	0.8
30–40	35	35	.02857	.42855
40–50	45	45	.02222	.28886
50–60	55	55	.01818	.14544
$N = 163$			$\Sigma(f 1/m) = 2.19585$	

$$\begin{aligned} \text{H.M.} &= \frac{N}{\sum(f 1/m)} \\ &= \frac{163}{2.19585} = 74.23 \end{aligned}$$

Merits of Harmonic Mean

1. It is rigidly defined.
2. It is based on all the observations of the series.
3. It is suitable in case of series having wide dispersion.
4. It is suitable for further mathematical treatment.
5. It gives less weight to large items and more weight to small items.

Demerits of Harmonic Mean

1. It is difficult to calculate and understandable.
2. It is not popular.
3. All the values must be available for computation.
4. It is usually a value which does not exist in series.

G. Median

Meaning Median is the middle value of the group of data arranged in an order, either on an ascending order or descending order. It divides the entire group of data into two equal parts.

- (i) The first part should be the observations which are smaller than median.
- (ii) The second part is the observations which are greater than median.

If the total number of observations is in odd numbers then the median can be ascertained without any difficulty.

Definitions

The median, as its name indicates, is the value of the middle item in a series, when items are arranged according to magnitude. — **Yan Lun Chou**

Median of a series is the value of the items actual or estimated when a series is arranged in order of magnitude which divides the distribution into two parts.

— **Se Crist**

Median—Individual Observations Median refers to the middle value in a distribution. In a series of individual observations, if the total number of items is an odd figure, the value of the middle item is the median value.

The formula for calculating median is

$$\text{Median, } M = \left[\frac{N+1}{2} \right]^{\text{th}} \text{ item}$$

Illustration 6.20

Sl. No.	1	2	3	4	5	6	7
Weight in kg	48	59	67	56	50	60	78

Find out the median weight of students.

Solutions

Sl. No.	Weight in kg	Weight in kg Ascending Order
1	48	48
2	59	50
3	67	56
4	56	59
5	50	60
6	60	67
7	78	78

$$\text{Median, } M = \left[\frac{N+1}{2} \right]^{\text{th item}}$$

$$M = \left[\frac{7+1}{2} \right] = 4^{\text{th}} \text{ item}$$

Hence, the median weight is 59 kilograms.

Median—Discrete Series

$$M = \left[\frac{N+1}{2} \right]^{\text{th item}}$$

Illustration 6.21

Calculate median from the following data.

Wages (Rs.)	35	45	55	65	75
No. of Workers	19	12	15	10	14

Solutions

Calculation of Median

Wages (Rs.)	No. of Workers <i>f</i>	Cumulative Value c.f.
35	19	19
45	12	31
55	15	46
65	10	56
75	14	70
<i>N = 70</i>		

$$\text{Median, } M = \left[\frac{N+1}{2} \right]^{\text{th item}}$$

$$= \left[\frac{70+1}{2} \right]^{\text{th item}} \\ = 35.5^{\text{th}} \text{ item.}$$

Hence, the median value of wages is Rs. 55

Median—Continuous Series

$$M = L + \left[\frac{N/2 - c.f.}{f} \right] \times i$$

L = Lower value of the median class.

$c.f.$ = Cumulative frequency of the class preceding the median class.

f = The frequency of the median class.

i = The class interval of the median class.

Steps

- (i) Find out the median by using $N/2$.
- (ii) Find out the class in which median lies.
- (iii) Apply the formula.

Illustration 6.22

The monthly income of 10 families in a town is stated below:

Monthly Income Rs.	No. of families
10–20	5
20–30	7
30–40	15
40–50	20
50–60	22
60–70	12
70–80	9
80–90	8
90–100	7

Calculate the median value.

Solutions

Calculation of median

Monthly Income Rs.	No. of families f	Cumulative Value $c.f.$
10–20	5	5
20–30	7	12
30–40	15	27
40–50	20	47

Monthly Income Rs.	No. of families f	Cumulative Value c.f.
50–60	22	69
60–70	12	81
70–80	9	90
80–90	8	98
90–100	7	105
N = 105		

$$\text{Median } M = L + \left[\frac{N/2 - c.f.}{f} \right] \times i$$

Hence, median class = 50–60

$L = 50; f = 22; c.f. = 47; i = 50-60 = 10; N/2 = 52.5$

$$\begin{aligned}\text{Median } M &= 50 + \frac{52.5 - 47}{22} \times 10 \\ &= 50 + \frac{5.5}{22} \times 10 \\ &= 50 + 2.5 = 52.5\end{aligned}$$

Hence, the median value of monthly income is Rs. 52.5.

Illustration 6.23

Compute the median from the following data:

Mid-value	Frequency
5	20
15	22
25	25
35	47
45	62
55	45
65	22

Solutions

Calcualtion of Median

Mid-value m	Class Intervals x	Frequency f	Cumulative Value c.f
5	0–10	20	20
15	10–20	22	42
25	20–30	25	67
35	30–40	47	114
45	40–50	62	176
55	50–60	45	221
	60–70	22	243
N = 243			

$$\begin{aligned}\text{Median } M &= L + \left[\frac{N/2 - c.f}{f} \right] \times i \\ &= 40 + \frac{121.5 - 114}{62} \times 10 \\ &= 40 + 1.21 = 41.21\end{aligned}$$

Hence, the median is 41.21.

Illustration 6.24

The details regarding the yearly production of 100 factories are stated in the following data. Find out the median.

Production in tonnes x	No. of factories f
100–400	25
401–500	15
501–600	21
601–700	20
701–800	23
801–900	16
901–1000	20

Solutions

Calculation of median

Production in tonnes x	No. of factories f	Cumulative Value c.f.
100–400.5	25	25
400.5–500.5	15	40
500.5–600.5	21	61
600.5–700.5	20	81
700.5–800.5	23	104
800.5–900.5	16	120
900.5–1000.5	20	140
N = 140		

$$\begin{aligned}\text{Median } M &= L + \left[\frac{N/2 - c.f}{f} \right] \times i \\ &= 600.5 + \frac{70 - 61}{20} \times 100 \\ &= 600.5 + 45 = 645.5\end{aligned}$$

Hence, mid-value of production is 645.5 tonnes.

Merits of Median

1. It is easy to understand and easy to compute.
2. It can be calculated through graphical method.

3. It is quite rigidly defined.
4. It is used for further analysis of data.
5. It can be calculated even though the extremes of the variables are not known.
6. It can be located even by mere inspection.
7. Its value generally lies in the distribution.
8. It can be calculated even from qualitative phenomena, that is, honesty, character etc.
9. It is amenable to further algebraic process.
10. It eliminates the effect of extreme items.

Demerits of Median

1. Where the number of items is large, the pre-requisite process, that is arranging the items is difficult process.
2. It ignores the extreme items.
3. In case of continuous series, the median is estimated, but not calculated.
4. It is more affected by fluctuations of sampling than in mean.
5. Median is not amenable to further algebraic manipulation.
6. Calculation of median for continuous series is complicated.
7. The use of median for further algebraic analysis is very limited.
8. Arrangement of data on the basis of ascending or descending order is must for calculating median.

Other Measures of Median Median is termed as position measures of data.

H. Quartiles

Quartiles may be defined as those values which divide the total frequencies into four equal parts. They are termed as Q_1 , Q_2 , Q_3 and Q_4 . Formula for calculation of quartiles are:

$$Q_1 = N/4^{\text{th}} \text{ items}$$

$$Q_2 = 2N/4 \text{ or } N/2^{\text{th}} \text{ item (median)}$$

$$Q_3 = 3N/4^{\text{th}} \text{ item}$$

Q_2 is called median.

Quartiles—Individual and Discrete Series The method for calculating the quartiles is the same as that for median. The following steps may be noted:

1. Find out the cumulative frequency.
2. Then apply the formula.

First Quartile

(Lower Quartile) or Q_1 = Size of $(N+1)/4^{\text{th}}$ item

Third Quartile

(Upper Quartile) or Q_3 = Size of $3(N+1)/4^{\text{th}}$ item

*Individual Series***Illustration 6.25**

In the series given below find the first quartile and third quartile.

46	55	39	28	40	67	53
----	----	----	----	----	----	----

Solutions

Data: 28 39 40 46 53 55 67
 $N = 7$

$$Q_1 = \frac{N+1}{4}^{\text{th}} \text{ item} = \frac{7+1}{4}^{\text{th}} \text{ item} = 2^{\text{th}} \text{ item} = 39$$

$$Q_3 = \frac{3(N+1)}{4}^{\text{th}} \text{ item} = \frac{3(7+1)}{4}^{\text{th}} \text{ item} = 6^{\text{th}} \text{ item} = 55$$

Illustration 6.26

Wages Rs.	40	50	60	70	80	90
No. of workers	14	18	20	19	16	13

Find Q_1 and Q_3 .

Solutions

Wages (Rs.)	No. of Workers f	Cumulative Value $c.f$
40	14	14
50	18	32
60	20	52
70	19	71
80	16	87
90	13	100
$N = 100$		

$$Q_1 = \frac{N+1}{4}^{\text{th}} \text{ item} = \frac{100+1}{4}^{\text{th}} \text{ item} = 25.25^{\text{th}} \text{ item} = 50$$

$$Q_3 = \frac{3(N+1)}{4}^{\text{th}} \text{ item} = \frac{3(100+1)}{4}^{\text{th}} \text{ item} = 75.75^{\text{th}} \text{ item} = 70$$

*Continuous Series***Steps**

1. Find out the cumulative frequency.
2. Find out the first quartile item ($N/4$).
3. Find out the class containing the first quartile item.

140 Business Statistics

4. The first quartile is estimated by applying the following formula:

$$Q_1 = L_1 + \frac{N/4 - c.f.}{f} \times i$$

5. Find out the third quartile item ($3N/4$).
 6. Find out the class containing the third quartile item.
 7. The upper quartile is estimated by applying the following formula:

$$Q_3 = L_1 + \frac{3N/4 - c.f.}{f} \times i$$

Illustration 6.27

Find out the quartiles from the following data.

Marks	10–20	20–30	30–40	40–50	50–60
No. of Students	14	12	10	15	19

Solutions

Calculation of Q_1 and Q_3

Marks	No. of Students	Cumulative Value c.f.
10–20	14	14
20–30	12	26
30–40	10	36
40–50	15	51
50–60	19	70
N = 70		

$$Q_1 = L_1 + \frac{N^{\text{th}}}{4} \text{ item } \frac{70}{4} = 17.5^{\text{th}} \text{ item}$$

Hence $c.f. = 14$, $L = 20$, $f = 12$, $i = 10$

$$\begin{aligned} Q_1 &= L_1 + \frac{N/4 - c.f.}{f} \times i \\ &= 20 + \frac{17.5 - 14}{12} \times 10 \\ &= 20 + 2.92 = 22.92 \end{aligned}$$

$$Q_3 = \frac{3N^{\text{th}}}{4} \text{ item} = \frac{3 \times 70}{4} = 52.5^{\text{th}} \text{ item}$$

Hence $c.f. = 51$, $L_1 = 50$, $f = 19$, $i = 10$

$$\begin{aligned} Q_3 &= L_1 + \frac{3N/4 - c.f.}{f} \times i \\ &= 50 + \frac{3 \times 17.5 - 51}{19} \times 10 \\ &= 50 + 0.79 = 50.79 \end{aligned}$$

Illustration 6.28

The following are the information relating to daily wages of workers in a factory:

Wages	0–10	10–20	20–30	30–40
No. of workers	17	12	19	18

Calculate Q_3 , D_3 , P_2 , P_{22} and P_{74} .

Solutions

Calculation of Quartiles, Deciles and Percentiles

Wages	No. of Workers	Cumulative No. of Workers c.f
0–10	17	17
10–20	12	29
20–30	19	48
30–40	18	66

$$Q_3 = \frac{3N^{\text{th}}}{4} \text{ item} = \frac{3 \times 66}{4} = \frac{198}{4} = 49.5^{\text{th}} \text{ item}$$

Hence $c.f = 48$, $L = 30$, $f = 18$, $i = 10$

$$\begin{aligned} Q_3 &= L_1 + \frac{3N/4 - c.f}{f} \times i \\ &= 30 + \frac{3 \times 66/4 - 48}{18} \times 10 \\ &= 30 + 0.83 = 30.83 \end{aligned}$$

$$D_3 = \frac{3N^{\text{th}}}{10} \text{ item} = \frac{3 \times 66}{10} = \frac{198}{10} = 19.8^{\text{th}} \text{ item}$$

Hence $c.f = 17$, $L = 10$, $f = 12$, $i = 10$

$$\begin{aligned} D_3 &= L + \frac{3N/10 - c.f}{f} \times i \\ &= 10 + \frac{3 \times 66/10 - 17}{12} \times 10 \\ &= 10 + 2.33 = 12.33 \end{aligned}$$

$$P_2 = \frac{2N^{\text{th}}}{100} \text{ item} = \frac{2 \times 66}{100} = 1.32^{\text{th}} \text{ item}$$

Hence $c.f = 0$, $L = 0$, $f = 17$, $i = 10$

$$\begin{aligned} P_2 &= L_1 + \frac{2N/100 - c.f}{f} \times i \\ &= 0 + \frac{2 \times 66/100 - 0}{17} \times 10 \\ &= 0 + \frac{1.32 - 0}{17} \times 10 \end{aligned}$$

$$P_2 = 0.78$$

$$P_{22} = \frac{22N^{\text{th}}}{100} \text{ item} = \frac{22 \times 66}{100} = 14.52^{\text{th}} \text{ item}$$

Hence $c.f = 0$, $L = 0$, $f = 17$, $i = 10$

$$\begin{aligned} P_2 &= L_1 + \frac{22N/100 - c.f}{f} \times i \\ &= 0 + \frac{22 \times 66/100 - 0}{17} \times 10 \end{aligned}$$

$$\begin{aligned} P_{22} &= 0 + 8.541 \\ &= 8.541 \end{aligned}$$

$$P_{74} = \frac{74N^{\text{th}}}{100} \text{ item} = \frac{74 \times 66}{100} = 48.84^{\text{th}} \text{ item}$$

Hence $c.f = 48$, $L = 30$, $f = 18$, $i = 10$

$$\begin{aligned} P_{74} &= L + \frac{74N/100 - c.f}{f} \times i \\ &= 30 + \frac{74 \times 66/100 - 48}{18} \times 10 \\ P_{74} &= 30 + 0.467 \\ &= 30.467 \end{aligned}$$

I. Mode Mode means the value that occurs most frequently in a statistical distribution.

Mode—Individual Series In individual observation, mode can be easily located without applying any formula. The data which appeared the maximum number of time can be treated as mode for that group of data. In individual observation, these may be chance of obtaining one mode, two modes, three modes or more than three modes.

Illustration 6.29

Calculate mode from the following data:

75	86	75	83	84	75	83
----	----	----	----	----	----	----

Solutions

In data 75, 86, 75, 83, 84, 75, 83. 75 repeated 3 times and 83 repeated 2 times. Hence, mode = 75

Illustration 6.30

Calculate mode from the following data.

60	50	35	45	30	35	45	65	70
----	----	----	----	----	----	----	----	----

Solutions

In this data, 60, 50, 35, 45, 30, 35, 45, 65 and 70. 35 and 45 repeated 2 times each. Hence both these data can be taken as modes.

Answer Mode (i) = 35 (Bimodal)
Mode (i) = 45

Illustration 6.31

Calculate mode from the following data.

70 72 73 56 37 62 55 85

Solutions

In this group of data, 70, 72, 73, 56, 37, 62, 55 and 85. There is no mode since each item appeared once.

Mode—Discrete Series We cannot depend on the method of inspection to find out the mode. In such situations, it is suggested to analysis table to find out the mode. First we prepare grouping table and then an analysis table.

Steps

1. Prepare a grouping table with 6 columns.
2. Write the size of the item in the margin.
3. In column 1, write the frequencies against respective items.
4. In column 2, the frequencies are grouped in twos.
(1 and 2, 3 and 4, 5 and 6 and so on)
5. In column 3, the frequencies are grouped in twos, leaving the first frequencies.
(2 and 3, 4 and 5, 6 and 7 and so on)
6. In column 4, the frequencies are grouped in threes.
(1, 2 and 3, 4, 5 and 6, 7, 8 and 9 and so on)
7. In column 5, the frequencies are grouped in threes, leaving the first frequency.
(2, 3 and 4, 5, 6 and 7, 8, 9 and 10 and so on)
8. In column 6, the frequencies are grouped in threes, leaving the first two frequencies.
(3, 4 and 5, 6, 7 and 8 and so on)
In all the processes, mark down the maximum frequencies by bold letters or by a circle.
9. After grouping the frequencies table, an analysis table is prepared to show the exact size, which has the highest frequency.

Illustration 6.32

Calculate the mode from the following:

Size	20	21	22	23	24	25	26	27	28
Frequency	20	22	25	29	30	18	14	13	12

Solutions

Grouping Table

Size	Frequency →					
	1	2	3	4	5	6
20	20					
21	22	42				
22	25		47	67		
			(20 + 22 + 25)			
23	29	54			76	
		(25 + 29)			22 + 25 + 29	
24	30		59			84
			29+30			(25 + 29 + 30)
25	18	48		77		
				(29 + 30 + 18)		
26	14		32			
27	13	27				
28	12		25	39		

Analysis Table

Col. No.	Size of item containing maximum frequency					
1	21	22	23	24	25	
2		1	1			
3			1	1		
4			1	1	1	
5		1	1			
6	1	1	1	1		
Total	1	3	5	3	1	

Hence mode = 23; in the mere inspection, it appeared that mode in 24 but it is proved that mode = 23.

Mode—Continuous Series In a continuous series, to find out the mode, we need one step more than those used for discrete series. As explained in the discrete series, model class is determined by preparing grouping table and analysis tables. Then we apply the following formula:

$$Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where,

Z = Mode

L_1 = Lower limit of the model class

f_1 = Frequency of the model class

f_0 = Frequency of the class preceding the model class

f_2 = Frequency of the class succeeding the model class

i = Class interval

Illustration 6.33

Calculate the mode from the following distribution.

Marks	No. of Students
10–20	14
20–30	17
30–40	25
40–50	21
50–60	16
60–70	19
70–80	14
80–90	13

Solutions

Model class = 30 – 40, since it is repeated maximum number of times.

Hence $f_1 = 25$; $f_2 = 21$; $f_0 = 17$; $L = 30$; $i = 10$

$$z = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$\Delta_1 = f_1 - f_0 = 25 - 17 = 8$$

$$\Delta_2 = f_1 - f_2 = 25 - 21 = 4$$

$$z = 30 + \frac{8}{8+4} \times 10$$

$$= 30 + 6.67$$

$$= 36.67$$

Illustration 6.34

Calculate mode from the following data:

Size	Frequency
40–50	14
50–60	15
60–70	25
70–80	19
80–90	21
90–100	24

Size	Frequency
100–110	18
110–120	23
120–130	17
130–140	18

Solution

Grouping Table

Size	Frequency	1	2	3	4	5
40–50	14					
50–60	15	29				
60–70	25		40	54		
70–80	19	44			59	
80–90	21		40			65 (25 + 19 + 21)
90–100	24	45 (21 + 24)		64 (19 + 21 + 24)		
100–110	18		42 (24 + 18)		63 (21 + 24 + 18)	65 (24 + 18 + 23)
110–120	23	41		58		
120–130	17		40			
130–140	18	35			58	

Analysis Table

Col. No.	Size of item containing maximum frequency					
	60–70	70–80	80–90	90–100	100–110	110–120
1			1	1	1	
2		1			1	
3		1	1	1		
4			1	1	1	
5	1		1	1	1	1
Total	1	2	4	5	3	1

Through it appears that 60–70 is the model class, the real model class is 90–100.

$$\text{Hence, } f_1 = 24; f_2 = 18; f_0 = 21; L = 90; i = 10$$

$$\Delta_1 = f_1 - f_0 = 24 - 21 = 3$$

$$\Delta_2 = f_1 - f_2 = 24 - 18 = 6$$

$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$z = 90 + \frac{3}{3+6} \times 10$$

$$= 90 + 3.33$$

$$= 93.33$$

6.4 MISCELLANEOUS ILLUSTRATIONS

6.4.1 Discrete series—Direct Method

Illustration 6.35

Calculate mean from the following data:

Value	2	4	6	8	10	12	14	16
Frequency	42	60	56	20	52	17	20	33

Solutions

Calculation of mean

x	f	fx
2	42	84
4	60	240
6	56	336
8	20	160
10	52	520
12	17	204
14	20	280
16	33	528
$N = 300$		$\Sigma fx = 2352$

$$\bar{X} = \frac{\sum fx}{N}$$

\bar{X} = Arithmetic mean

Σfx = The sum of product

N = Total number of items

$\Sigma fx = 2352$

$N = 300$

$$\bar{X} = \frac{\sum fx}{N} = \frac{2352}{300}$$

$$\bar{X} = 7.84$$

6.4.2 Short-cut Method

$$\bar{X} = A \pm \frac{\sum fd}{N}$$

A = Assumed mean

N = Total frequency

Σfd = Sum of total deviations

Illustration 6.36

Calculate mean from the following data:

x	2	4	6	8	10	12	14	16
f	42	60	56	20	52	17	20	33

Solutions

Calculation of mean

x	f	d = x - A (A= 8)	fD
2	42	-6	-252
4	60	-4	-240
6	56	-2	-112
8	20	0	0
10	52	2	104
12	17	4	68
14	20	6	120
16	33	8	264
N = 300		$\Sigma fd = -48$	

$$\begin{aligned}\bar{X} &= A \pm \frac{\sum fd}{N} \\ &= 8 - \frac{48}{300} \\ &= 8 - 0.16 \\ \bar{X} &= 7.84\end{aligned}$$

6.4.3 Continuous Series Direct Method

Illustration 6.37

From the following data, find out mean sales:

Sales per shop	No. of shop.
10 – 20	5
20 – 30	9
30 – 40	10
40 – 50	13
50 – 60	15
60 – 70	14
70 – 80	9

Solutions

Calculation of mean

x	f	Mid point (m)	fm
10 – 20	5	15	75
20 – 30	9	25	225
30 – 40	10	35	350
40 – 50	13	45	585
50 – 60	15	55	825
60 – 70	14	65	910
70 – 80	9	75	675
N = 75		$\Sigma fm = 3645$	

$$\bar{X} = \frac{\sum fm}{N}$$

$$\sum fm = 3645, N = 75$$

$$= \frac{3645}{75}$$

$$\bar{X} = 48.6$$

6.4.4 Short-cut Method

x	f	Mid point (m)	d = (m - A)	m - 45	fd
10 – 20	5	15	-30	-30	-150
20 – 30	9	25	-20	-20	-180
30 – 40	10	35	-10	-10	-100
40 – 50	13	45	0	0	0
50 – 60	15	55	10	10	150
60 – 70	14	65	20	20	280
70 – 80	9	75	30	30	270
N = 75		$\Sigma fd = 270$			

$$\bar{X} = A \pm \frac{\Sigma fd}{N}$$

$$\Sigma fd = 270, N = 75, A = 45$$

$$= 45 + \frac{270}{75} = 45 + 3.6$$

$$\bar{X} = 48.6$$

6.4.5 Step Deviation Method

x	F	Mid point (x)	m - A	A = 45	$d' = d/c$	fd'
				d		
10 – 20	5	15	-30	-30	-3	-15
20 – 30	9	25	-20	-20	-2	-18
30 – 40	10	35	-10	-10	-1	-10
40 – 50	13	45	0	0	0	0
50 – 60	15	55	10	10	1	15
60 – 70	14	65	20	20	2	28
70 – 80	9	75	30	30	3	27
75				27		

150 Business Statistics

$$\sum fd = 27, N = 75, A = 45, C = 10$$

$$\begin{aligned}\bar{X} &= A \pm \frac{\sum fd}{N} \times C \\ &= 45 + \frac{27}{75} \times 10 \\ &= 45 + 3.6 \\ \bar{X} &= 48.6\end{aligned}$$

Illustration 6.38

Calculate mean from the following data:

Value	Frequency
Less than 100	4
Less than 200	10
Less than 300	15
Less than 400	25
Less than 500	30
Less than 600	35
Less than 700	45
Less than 800	65

Solutions

Value	f	Mid point (x)	A = 450, d' m-A	d/c	d'	fd'
0 – 100	4	50	-400	-4	-16	
100 – 200	(10–4) 6	150	-300	-3	-18	
200 – 300	(15–10) 5	250	-200	-2	-10	
300 – 400	(25–15) 10	350	-100	-1	-10	
400 – 500	(30–25) 5	450	0	0	0	
500 – 600	(35–30) 5	550	100	1	5	
600 – 700	(45–35) 10	650	200	2	20	
700 – 800	(65–45) 20	750	300	3	60	
	65			-4	31	

$$\sum fd' = 31, N = 65, A = 450, C = 100$$

$$\begin{aligned}X &= A \pm \frac{\sum fd'}{N} \times C \\ &= 450 + \frac{31}{65} \times 100 \\ &= 450 + \frac{3100}{65} \\ &= 450 + 47.69\end{aligned}$$

$$\bar{X} = 497.69$$

Illustration 6.39

From the following information pertaining to 100 workers, calculate average wage paid to workers.

Wages (Rs.)	No. of workers
More than 50	100
More than 60	90
More than 70	65
More than 80	60
More than 90	57
More than 100	40
More than 110	32
More than 120	25

Solutions

x	f	Mid point (x)	d = m - A m - 95	d' = d/c	fd'
50–60	(100–90) 10	55	-40	-4	-40
60–70	(90–65) 25	65	-30	-3	-75
70–80	(65–60) 5	75	-20	-2	-10
80–90	(60–57) 3	85	-10	-1	-3
90–100	(57–40) 17	95	0	0	0
100–110	(40–32) 8	105	10	1	8
110–120	(32–25) 7	115	20	2	14
120–130	25	125	30	3	75
100			-40	4	$\Sigma fd' = -31$

$$\Sigma fd' = -31, N = 100, A = 95$$

$$\begin{aligned}\bar{X} &= A \pm \frac{\sum fd'}{N} \\ &= 95 \pm \frac{-31}{100} \\ &= 95 - 0.31\end{aligned}$$

$$\bar{X} = 94.69$$

6.4.6 Inclusive Class Intervals**Illustration 6.40**

Calculate mean from the following data:

Class interval	0–9	10–19	20–29	30–39	40–49	50–59
Frequency	1	3	9	11	14	2

Solutions

Class interval	f	Mid point (x)	d = m-A	m-95	d' = d/c	fd'
0– 9	1	4.5	– 20	– 2	– 2	
10 – 19	3	14.5	– 10	– 1	– 3	
20 – 29	9	24.5	0	0	0	
30 – 39	11	34.5	10	1	11	
40 – 49	14	44.5	20	2	28	
50 – 59	2	54.5	30	3	6	
N = 40				3	$\Sigma fd' = 40$	

$$\Sigma fd' = 40, N = 40, A = 24.5$$

$$\begin{aligned}\bar{X} &= A \pm \frac{\Sigma fd'}{N} \\ &= 24.5 + \frac{40}{40} \\ &= 24.5 + 1 \\ \bar{X} &= 25.5\end{aligned}$$

6.4.7 Correcting Incorrect Mean (Mis-read Items)**Illustration 6.41**

The average marks secured by 18 students were 26. But it was discovered that an item 32 was missed as 23. Find the correct mean of marks.

Solutions

$$N = 18 \quad x = 26$$

$$\bar{X} = \frac{\sum x}{N} \quad \Sigma x = X \times N = 26 \times 18 = 468$$

$$\text{Wrong } \Sigma x = 468$$

$$\begin{aligned}\text{Correct } \Sigma x &= \text{incorrect } \Sigma x - \text{wrong items} + \text{current items} \\ &= 468 - 23 + 32\end{aligned}$$

$$\text{Correct } \Sigma x = 477$$

$$\bar{X} = \frac{\Sigma x}{N} = \frac{477}{18} = 26.5$$

$$\bar{X} = 26.5$$

Illustration 6.42

The mean of 200 items was 92. Later on, it was discovered that an item 32 was mis-read as 23 and another item 86 was misread as 68; it was also found that the number of items was 180 and not 200. Find the correct mean.

Solutions

Wrong aggregate of 200 items = $200 \times 92 =$	18,400
Less: Wrong value of items included 23 and 68 =	<u>91</u>
	18,309
Add: Correct value of items to be included 32 and 86	<u>118</u>
Correct Σx	<u>18,427</u>
Correct No. of items = 180	
$\bar{X} = \frac{\Sigma x}{N} = \frac{18,427}{180}$	
	$\bar{X} = 102.37$

6.4.8 Combined Arithmetic Mean

$$X_{12} = \frac{N_1 X_1 + N_2 X_2}{N_1 + N_2}$$

$$X_{123} = \frac{N_1 X_1 + N_2 X_2 + N_3 X_3}{N_1 + N_2 + N_3}$$

Illustration 6.43

In a factory, there are 50 skilled, 125 semi-skilled and 75 unskilled workers. It has been observed that on an average a unit length of a particular fabric is woven by skilled workers in 1.5 hours, by a semi-skilled workers in 2 hours and by an unskilled workers in 2.5 hours. After a training of 1 year, the semi skilled workers are expected to become skilled and unskilled workers to become semi-skilled. How much less time will be required after 1 year of training for weaving the unit length of fabric by an average worker?

Solutions

Average time per worker before training

$$= \frac{(200 \times 1.5) + (125 \times 2) + (75 \times 2.5)}{200 + 125 + 75}$$

$$= \frac{300 + 250 + 187.5}{400}$$

$$= \frac{737.5}{400} = 1.84 \text{ hrs.}$$

After training the composition of workers is as follows:

$$\text{Skilled worker} = 50 + 125 = 175$$

$$\text{Semi-skilled workers} = 75$$

$$\text{Unskilled workers} = \text{Nil}$$

Average time per workers after one year training is:

$$\begin{aligned} &= \frac{(175 \times 1.5) + (75 \times 2)}{175 + 75} \\ &= \frac{262.5 + 150}{250} \\ &= 1.65 \text{ hrs.} \end{aligned}$$

After one year training, 0.19 hour less than would be required.

6.4.9 Missing Values and Missing Frequency

Illustration 6.44

Find out the missing values of the variant for the following distribution whose mean is 28.03.

x	8	16	23	29	?	54
f	4	8	24	45	15	4

Solutions

Computation of missing figure

x	f	fx
8	4	32
16	8	128
23	24	552
29	45	1305
?x	15	15x
54	4	216
N = 100		2233+15x

$$\bar{X} = \frac{\sum fx}{N}$$

$$\frac{28.03}{1} = \frac{2233+15x}{100}$$

$$2803 = 2233 + 15x$$

$$15x = 2803 - 2233$$

$$\bar{X} = \frac{570}{15}$$

$$\bar{X} = 38$$

6.4.10 Missing Frequency

Illustration 6.45

For certain frequency table which is only partly reproduced here, the mean was found to be 6.56.

x	f
2	23
4	?
6	?
8	5
10	27
12	10
	100

Calculate the missing frequencies.

Solutions

Computation of missing figures

X	f	fx
2	23	46
4	f_1	$4f_1$
6	f_2	$6f_2$
8	5	40
10	27	270
12	10	120
	$65 + f_1 + f_2 = 100$	$476 + 4f_1 + 6f_2$

$$\text{We know } 100 = 86 + f_1 + f_2$$

$$\begin{aligned} 100 - 65 &= f_1 + f_2 \\ 35 &= f_1 + f_2 \end{aligned} \tag{1}$$

$$\bar{X} = \frac{\sum fx}{\sum f} = 6.56 = \frac{476 + 4f_1 + 6f_2}{100}$$

$$\begin{aligned} 656 &= 476 + 4f_1 + 6f_2 \\ 656 - 476 &= 4f_1 + 6f_2 \\ 180 &= 4f_1 + 6f_2 \end{aligned} \tag{2}$$

Sub (1) $\times 4$ from (2),

$$\begin{array}{r} 4f_1 + 6f_2 = 180 \\ 4f_1 + 4f_2 = 140 \\ \hline (-) \quad (-) \quad (-) \\ 2f_2 = 40 \\ f_2 = 40/2 = 20 \\ f_2 = 20 \end{array}$$

Substituting in (1), we get

$$\begin{aligned} f_1 + f_2 &= 35 \\ f_1 + 20 &= 35 \\ f_1 &= 35 - 20 \\ f_1 &= 15 \\ f_1 &= 15, f_2 = 20 \end{aligned}$$

156 Business Statistics

6.4.11 Median

Individual Series

Illustration 6.46

Following are the marks scored by 9 students; find out the median marks.

Roll No.	Marks
1	54
2	23
3	81
4	75
5	56
6	82
6	64
7	36
8	28

Solutions

Ascending order

Roll No.	Marks
2	23
9	28
8	36
1	54
5	56
7	64
4	75
3	81
6	82

$$\text{Median} = \text{Size of } \frac{(N+1)^{\text{th}}}{2} \text{ item}$$

$$= \frac{9+1}{2} = \frac{10^{\text{th}}}{2} \text{ item}$$

Value of the 5th item = 56

Median = 56.

Illustration 6.47

Find out the median from the following.

52 53 56 37 33 60 67 61

Solutions

Ascending order

S.No.	Value
1	33
2	37
3	52
4	53
5	56
6	60
7	61
8	67

$$\text{Median} = \text{Size of } \frac{(N+1)^{\text{th}}}{2} \text{ item}$$

$$= \frac{8+1}{2} = \frac{9}{2} = 4.5 \text{ hrs.}$$

$$\text{value of } 4.5^{\text{th}} \text{ item} = \frac{53+56}{2}$$

$$= \frac{109}{2} = 54.5$$

$$\text{Median} = 54.5.$$

Illustration 6.48

Computation of median from the following data:

x	10	10.5	11	11.5	12	12.5	13
f	5	8	14	8	14	20	17

Solutions

Computation of Median

x	f	cf
10	5	5
10.5	8	13
11	14	27
11.5	8	35
12	14	49
12.5	20	69
13	17	86
		86

$$\text{Median} = \text{Size } \frac{(N+1)^{\text{th}}}{2} \text{ item}$$

$$\begin{aligned}
 &= \frac{86+1^{\text{th}}}{2} \text{ item} \\
 &= \frac{87}{2} = 43.5^{\text{th}} \text{ item} \\
 \text{Value of } 43.5^{\text{th}} \text{ item} &= 12 \\
 \text{Median} &= 12.
 \end{aligned}$$

Median—Continuous Series

Illustration 6.49

From the following data, find out median.

Marks	Frequency
10 – 25	3
25 – 40	10
40 – 55	20
55 – 70	13
70 – 85	2
85 – 100	2

Solutions

Calculation of median

X	f	cf
10 – 25	3	3
25 – 40	10	13
40 – 55	20	33
55 – 70	13	46
70 – 85	2	48
85 – 100	2	50

Median = Size of $N/2$ of the item

$$= \frac{50}{2} = 25^{\text{th}} \text{ item}$$

Median class = 40 – 55

$$M = L + \frac{N/2 - cf}{f} \times i$$

$$M = 40 + \frac{25-13}{20} \times 15 = 40 + 9 = 49$$

Median = 49

6.4.12 Quartiles

Lower quartile (Q_1) = Size of $\frac{(N+1)^{\text{th}}}{4}$ item

$$\text{Median } (Q_2) = \text{Size of } \frac{2 \times (N+1)^{\text{th}}}{4} \text{ item}$$

$$\text{Upper quartile } (Q_3) = \text{Size of } \frac{3 \times (N+1)^{\text{th}}}{4} \text{ item}$$

Illustration 6.50

In the series given below find the first Quartile and third Quartile.

2	4	6	8	10	12	14	16	18	20	22
---	---	---	---	----	----	----	----	----	----	----

Solutions

$$\begin{aligned} Q_1 &= \text{Size of } \frac{(N+1)^{\text{th}}}{4} \text{ item} \\ &= \text{Size of } \frac{(11+1)^{\text{th}}}{4} \text{ item} \\ &= 12/4^{\text{th}} \text{ item} \end{aligned}$$

Value of 3rd item = $Q_1 = 6$

$$\begin{aligned} Q_3 &= \text{Size of } \frac{3(N+1)^{\text{th}}}{4} \text{ item} \\ &= \text{Size of } \frac{3(11+1)^{\text{th}}}{4} \text{ item} \\ &= 36 / 4 = 9^{\text{th}} \text{ item} \end{aligned}$$

Value of 9th item is 18

$$Q_3 = 18$$

Illustration 6.51

Find Q_1 and Q_3 of the following series:

Size of shoes	Frequency
4	5
6	9
8	11
10	22
12	13
14	15
16	18

Solutions

Calculation of Q_1 and Q_2

x	f	cf
4	5	5
6	9	14

x	f	cf
8	11	25
10	22	47
12	13	60
14	15	75
16	18	93

$$\begin{aligned} Q_1 &= \text{Size of } \frac{(N+1)^{\text{th}}}{4} \text{ item} \\ &= \text{Size of } \frac{(93+1)^{\text{th}}}{4} \text{ item} \\ &= 94/4 = 23.5^{\text{rd}} \text{ item} \end{aligned}$$

Value of 23.5th item = 8

$$Q_1 = 8$$

$$\begin{aligned} Q_3 &= \text{Size of } \frac{3(N+1)^{\text{th}}}{4} \text{ item} \\ &= \text{Size of } \frac{3(93+1)^{\text{th}}}{4} \text{ item} \\ &= 70.5^{\text{th}} \text{ item} \end{aligned}$$

Value of 70.5th item = 14

$$Q_3 = 14$$

Continuous Series

Illustration 6.52

Calculate the value of the Quartile one and Quartile three from the following.

Value	Frequency
4 – 6	3
6 – 8	4
8 – 10	3
10 – 12	10
12 – 14	12
14 – 16	15
16 – 18	14

Solutions

x	f	cf
4 – 6	3	3
6 – 8	4	7
8 – 10	3	10
10 – 12	10	20
12 – 14	12	32
14 – 16	15	47
16 – 18	14	61

$$\begin{aligned}
 &= \text{Size of } \left[\frac{N}{4} \right]^{\text{th}} \text{ item} \\
 &= \text{Size of } \left[\frac{61}{4} \right]^{\text{th}} \text{ item} \\
 &= \frac{61}{4} = 15.25^{\text{th}} \text{ item}
 \end{aligned}$$

Q_1 class = 10–12

$$\begin{aligned}
 Q_1 &= L + \frac{N/4 - cf}{f} \times c \\
 &= 10 + \frac{15.25 - 10}{10} \times 2 \\
 &= 10 + 1.05
 \end{aligned}$$

$Q_1 = 11.05$

$$\begin{aligned}
 Q_3 &= \text{Size of } \frac{3(N+1)}{4}^{\text{th}} \text{ item} \\
 &= \text{Size of } \frac{3(61)}{4}^{\text{th}} \text{ item} \\
 &= 3 \times 61/4 = 45.75^{\text{th}} \text{ item}
 \end{aligned}$$

Q_3 class = 14–16

$$\begin{aligned}
 Q_3 &= L + \frac{3N/4 - cf}{f} \times i \\
 &= 14 + \frac{45.75 - 32}{15} \times 2 \\
 &= 14 + \frac{13.75}{15} \times 2 \\
 &= 14 + 1.83
 \end{aligned}$$

$Q_3 = 15.83$

Illustration 6.53

Find Lower quartile, Median, Upper quartile, Decile seven, sixtieth percentile for the following Frequency distribution.

Wages	5–10	10–15	15–20	20–25	25–30	30–35	35–40
No. of workers	2	6	22	42	86	64	18

Solutions

x	f	cf
5–10	2	2
10–15	6	8
15–20	22	30

x	f	cf
20–25	42	72
25–30	86	158
30–35	64	222
35–40	18	240
	240	240

$$Q_1 = \text{size of the } \frac{N^{\text{th}}}{4} \text{ item}$$

$$= \frac{240}{4} = 60^{\text{th}} \text{ item}$$

Value of 60th item = 20–25

$$Q_1 = L + \frac{N/4 - cf}{f} \times c$$

$$= 20 + \frac{60 - 30}{42} \times 5$$

$$= 20 + 3.57$$

$$Q_1 = 23.57$$

$$\text{Median} = \text{Size of } \frac{N^{\text{th}}}{2} \text{ item}$$

$$= \frac{240}{2} = 120^{\text{th}} \text{ item}$$

Value of 120th item = 25–30

$$M = L + \frac{N/2 - cf}{f} \times c$$

$$= 25 + \frac{120 - 72}{86} \times 5$$

$$= 25 + 2.79$$

$$M = 27.79$$

$$Q_3 = \text{Size of } \left(\frac{3 \times N}{4} \right)^{\text{th}} \text{ item}$$

$$= \frac{3 \times 240}{4} = 180^{\text{th}} \text{ item}$$

value of 180th item = 30–35

$$Q_3 = L + \frac{3N/4 - cf}{f} \times c$$

$$= 30 + \frac{180 - 158}{64} \times 5$$

$$= 30 + 1.72$$

$$Q_3 = 31.72$$

$$\begin{aligned}
 D_7 &= \text{Size of } \left(\frac{7N}{10} \right)^{\text{th}} \text{ item.} \\
 &= \text{Size of } = \frac{7 \times 240}{10} 168^{\text{th}} \text{ item} \\
 \text{Value of item } 168^{\text{th}} &= 30-35 \\
 D_7 &= L + \frac{7N/10 - cf}{f} \times c \\
 &= 30 + \frac{168 - 158}{64} \times 5 = 30 + \frac{10}{64} \times 5 \\
 D_7 &= 30.78 \\
 P_{60} &= \text{Size of } \frac{60N^{\text{th}}}{100} \text{ item} \\
 &= \text{Size of } \frac{60 \times 240}{100} = 144^{\text{th}} \text{ item} \\
 \text{Value of } 144^{\text{th}} &= 25-30 \\
 P_{60} &= L + \frac{60N/100 - cf}{f} \times c \\
 &= 25 + \frac{144 - 72}{86} \times 5 \\
 &= 25 + \frac{72}{86} \times 5 = 25 + 4.19 \\
 P_{60} &= 29.19
 \end{aligned}$$

6.4.13 Mode

Individual Series

Illustration 6.54

From the following data compute mode:

85 75 6 82 85 73 60 85 64 53

Mode = 85

Discrete Series

Illustration 6.55

Calculate the mode from the following.

Size	Frequency
10	10
11	12
12	15

Size	Frequency
13	19
14	20
15	8
16	4

Solutions

Grouping Table

Size	1	2	3	4	5	6
10	10					
11	12	22				
12	15		27			
13	19		(34)			
14	(20)			37		
15	8		(39)			(46)
16	4	28	12			(54)

Analysis Table

Size	1	2	3	4	5	6
10						
11				X		1
12		X		X	X	3
13		X	X	X	X	(5)
14	X		X	X	X	4
15				X		1
16						

The mode is 13 as this size of items repeats five times but through inspection, we say the mode is 14 because the size 14 occurs 20 times. But this wrong decision is revealed by analysis table.

Continuous Series

$$Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Z = Mode

L_1 = Lower limit of the model class

f_1 = Frequency of the model class

f_0 = Frequency of the class preceding the model class

f_2 = Frequency of the class succeeding the model class

i = Class interval

Illustration 6.56

Compute the mode from the following series.

Size	Frequency
0 – 10	10
10 – 20	12
20 – 30	16
30 – 40	14
40 – 50	10
50 – 60	8
60 – 70	17
70 – 80	5
80 – 90	4

Solutions

Grouping Table

Size of items	f_1	1	2	3	4	5
0 – 10	10	22				
10 – 20	$12 f_0$		(28)	(38)		
20 – 30	$16 f_1$	(30)			(42)	(40)
30 – 40	$14 f_2$		24			
40 – 50	10	18		32		
50 – 60	8		25			
60 – 70	(17)	22		26	35	30
70 – 80	5		9			
80 – 90	4					

Analysis Table

Size	1	2	3	4	5	6
0 – 10				X		1
10 – 20			X	X	X	3
20 – 30		X	X	X	X	(5)
30 – 40		X			X	3
40 – 50					X	1
50 – 60						
60 – 70	X					1
70 – 80						
80 – 90						

Illustration 6.57

From the above analysis table, we find that the modal class is 20–30 as its frequencies occur for the maximum times. Mode is estimated by the formula

$$\begin{aligned}
 Z &= L_1 \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \\
 &= 20 + \frac{16 - 12}{(2 \times 16) - 12 - 14} \times 10
 \end{aligned}$$

$$\begin{aligned}
 &= 20 + \frac{4}{32 - 12 - 14} \times 10 \\
 &= 20 + \frac{4}{6} \times 10 \\
 &= 20 + 6.67 \\
 Z &= 26.67
 \end{aligned}$$

Illustration 6.58

If in moderately asymmetrical frequency distribution, the values of median and arithmetic mean are 36 and 39 respectively. Estimate the value of the mode.

Solutions

$$\begin{aligned}
 \text{Mean} - \text{Mode} &= 3(\text{mean} - \text{median}) \\
 39 - \text{Mode} &= 3(39 - 36) \\
 39 - \text{Mode} &= 3(3) \\
 39 - \text{Mode} &= 9 \\
 -\text{Mode} &= 9 - 39 \\
 -\text{Mode} &= -30 \\
 \text{Mode} &= 30
 \end{aligned}$$

6.4.14 Geometric Mean

Geometric mean is defined as N th root of the product of N items.
 Geometric mean = Antilog of $\log x/N$

Illustration 6.59

Calculate geometric mean of the following:

25	36	27	41	47
----	----	----	----	----

Solutions

$$\begin{array}{ll}
 X & \log x \\
 \hline
 25 & 1.3979 \\
 36 & 1.5563 \\
 27 & 1.4317 \\
 41 & 1.6128 \\
 47 & 1.6721 \\
 \hline
 \Sigma \log x & \underline{\underline{7.6708}} \\
 \text{G.M.} & = \text{Antilog } \frac{\Sigma \log x}{N} \\
 & = \text{Antilog } \frac{7.6708}{5} = \text{Antilog } 1.53416 \\
 & = 34.20
 \end{array}$$

Discrete Series

$$\text{G.M.} = \text{Antilog of } \frac{\sum f \log x}{N}$$

Illustration 6.60

The following table gives the height of 30 persons in sample survey. Calculate geometric mean.

Height in cms	130	135	140	145	148	152	160	162	165
No. of persons	3	2	3	7	9	2	1	2	1

Solutions

Calculation of Geometric Mean

Height (x)	Frequency (f)	Log x	f log x
130	3	2.1139	6.3417
135	2	2.1303	4.2606
140	3	2.1461	6.4383
145	7	2.1614	15.1298
148	9	2.1703	19.5327
152	2	2.1818	4.3636
160	1	2.2041	2.2041
162	2	2.2095	4.4190
165	1	2.2175	2.2175
	30		60.4723

$$\begin{aligned}\text{G.M.} &= \text{Antilog of } \frac{\sum f \log x}{N} \\ &= \text{Antilog of } \frac{60.4723}{30} \\ &= \text{Antilog of } 2.0157\end{aligned}$$

G.M. height = 103.6 cms

Continuous Series**Illustration 6.61**

Find out the geometric mean from the following data:

x	f
0–4	5
4–8	9
8–12	19
12–16	23
16–20	7
20–24	5
24–28	2

Solutions

Calculation of Geometric Mean

x	f	Mid x	Log m	f log m
0–4	5	2	0.3010	0.6020
4–8	9	6	0.7782	4.6692
8–12	19	10	1.0000	10.0000
12–16	23	14	2.1461	16.0455
16–20	7	18	1.2553	22.5954
20–24	5	22	1.3424	29.5328
24–28	2	26	1.4150	36.7900
	70			120.2349

$$\begin{aligned} \text{G.M.} &= \text{Antilog of } \frac{\sum f \log x}{N} \\ &= \text{Antilog of } \frac{120.2349}{70} \\ &= \text{Antilog of } 1.7176 \\ \text{G.M.} &= 52.19 \end{aligned}$$

6.4.15 Harmonic Mean

Harmonic mean is the reciprocal of the arithmetic average of reciprocal of values of various items in the variables.

$$\text{H.M.} = \frac{N}{\sum 1/X}$$

Individual Series

Illustration 6.62

The yearly incomes of 8 families in rupees in a certain village are given below:

Family	1	2	3	4	5	6	7	8
Income	75	85	65	100	600	350	400	360

Solutions

Income (x)	Reciprocal (1/x)
75	0.01333
85	0.01176
65	0.01538
100	0.01000
600	0.00167
350	0.00286
400	0.00250
360	0.00278
$\Sigma 1/x = 0.06028$	

$$\begin{aligned}\text{H.M.} &= \frac{N}{\sum 1/X} \\ &= \frac{8}{0.06028} \\ \text{H.M.} &= 132.71\end{aligned}$$

Discrete Series**Illustration 6.63**

Calculate H.M. from the following data:

Size of items	8	10	12	14	16	18
Frequency	8	12	18	10	4	16

Solutions

Calculation of Harmonic Mean

x	f	Reciprocal (1/x)	f(1/x)
8	8	0.1250	1.0000
10	12	0.1000	1.2000
12	18	0.0833	1.4994
14	10	0.0714	0.7140
16	4	0.0625	0.2500
18	16	0.0556	0.8896
	68		$\Sigma f(1/x) = 5.5530$

$$\begin{aligned}\text{H.M.} &= \frac{N}{\sum f 1/X} \\ &= \frac{68}{5.5530} \\ \text{H.M.} &= 12.25\end{aligned}$$

Continuous Series**Illustration 6.64**

Calculate H.M. of the following data:

x	f
10 – 20	30
20 – 30	26
30 – 40	16
40 – 50	12
50 – 60	30
60 – 70	14
70 – 80	12

Solutions

Calculation of Harmonic Mean

X	f	Mid x	Reciprocal $1/m$	$f(1/m)$
10–20	30	15	0.06667	2.0001
20–30	26	25	0.04000	1.0400
30–40	16	35	0.02857	0.4571
40–50	12	45	0.02222	0.2666
50–60	30	55	0.01818	0.5454
60–70	14	65	0.01538	0.2153
70–80	12	75	0.01333	0.1599
	140			4.6853

$$\begin{aligned} \text{H.M.} &= \frac{N}{\sum f(1/m)} \\ &= \frac{140}{4.6853} \end{aligned}$$

H.M. = 29.88

Illustration 6.65

The values of mode and median for a moderately skewed distribution are 64.2 and 68.6 respectively. Find the value of the mean.

$$\begin{aligned} \text{Mean} &= \text{Mode} + (3/2) (\text{Median} - \text{Mode}) \\ &= 64.2 + (3/2) (68.6 - 64.2) \\ &= 64.2 + (3/2) (4.4) \\ &= 64.2 + 6.6 \end{aligned}$$

Mean = 70.8

Illustration 6.66

In a moderately asymmetrical distribution, the values of mode and mean are 32.1 and 35.4 respectively. Find the median value.

$$\begin{aligned} \text{Median} &= 1/3 (2 \text{ Mean} + \text{Mode}) \\ &= 1/3 (2 \times 35.4 + 32.1) \\ &= 1/3 \times 102.9 \end{aligned}$$

Median = 34.3

Illustration 6.67

If the mean and median of a moderately asymmetrical series are 26.8 and 27.9 respectively. What would be its most Probable mode?

$$\begin{aligned} \text{Mode} &= 3 \text{ Median} - 2 \text{ Mean} \\ &= 3 \times 27.9 - 2 \times 26.8 \\ &= 83.7 - 53.6 \end{aligned}$$

Mode = 30.1

SUMMARY

Average: It represents the whole value and it lies between the minimum and maximum value of the data. An average is a single number describing some features of the set of data.

Objectives of Average:

- To find out single value which represents the entire data
- To facilitate policy decision-making
- To facilitate easy comparison of data.

Characteristics of a Good Average:

- It should be easy to understand.
- It should be simple to calculate.
- It should be rigidly defined.
- It should represent all the data.
- It shouldn't be affected by the extreme value.
- It should be eligible for further algebraic treatment.
- It should have sampling stability.

Types of Average: • Mean • Median • Mode**Classification of Mean:**

- Arithmetic mean
- Geometric mean
- Harmonic mean

Arithmetic Mean:

It can be obtained by dividing the sum of all the observations by the total number of the observations.

Arithmetic mean for open-end classes:

If the lower limit of the first class interval and upper limit of the last class interval are not known it is called open-end classes.

Geometric Mean:

Nth root of the product of n observation of a series.

Harmonic Mean:

The total number of observations divided by the sum of reciprocals of the numbers.

Median:

Middle value of the group of data arranged in an order either on an ascending order or descending order.

Methods of determining median:

- One ogive method
- Two ogives methods.

Ogive Means cumulative frequency curve

Mode The value that occurs most frequently in a statistical distribution.

FORMULAE

Arithmetic Mean

1. Individual Series

(A) Direct Method

$$\bar{X} = \frac{\sum X}{N}$$

or

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

(B) Short-cut Method

$$\bar{X} = A \pm \frac{\sum d}{N}$$

(C) Step Deviation Method $\bar{X} = A \pm \frac{\sum d'}{N} \times C$

(A) \bar{X} = Arithmetic Mean

$\sum X$ = The sum of variables

N = Number of observations

(B) A = Assumed Mean

$\sum d$ = Sum of the deviations

N = Number of items

A = Assumed Mean

(C) $\sum d'$ = Sum of the deviations

C = Common factor

2. Discrete Series

(A) Direct Method

$$\bar{X} = \frac{\sum fx}{N}$$

(B) Short-cut Method

$$\bar{X} = A \pm \frac{\sum fd}{N}$$

(C) Step Deviation Method

$$\bar{X} = A \pm \frac{\sum fd'}{N} \times C$$

(A) $\sum fx$ = Sum of the products of the frequencies and the variable X

N = Total number of items i.e., $\sum f$

(B) A = Assumed Mean
 $\sum fd$ = Sum of total deviations

N = Total Frequency
 A = Assumed Mean

(C) $\sum fd'$ = Sum of deviations
 N = Total Frequency
 C = Common Factor

3. Continuous Series

(A) Direct Method $\bar{X} = \frac{\sum fm}{N}$

(B) Short-cut Method $\bar{X} = A \pm \frac{\sum fd}{N}$

(C) Step Deviation Method $\bar{X} = A \pm \frac{\sum fd'}{N} \times C$

(A) $\sum fm$ = Sum of the products of the frequencies and the mid points
 N = The total frequencies

(B) A = Assumed Mean
 $\sum fd$ = Sum of total deviations
 N = Total frequency

(C) A = Assumed Mean
 $\sum fd'$ = The sum of deviations
 N = Number of items
 C = Common factor

Median (Med)

1. Individual Series

Med : Size of $\left(\frac{N+1}{2}\right)^{\text{th}}$ item

2. Discrete Series

Med : Size of $\left(\frac{N+1}{2}\right)^{\text{th}}$ item

3. Continuous Series

Size of $\left(\frac{N}{2}\right)^{\text{th}}$ item

$\frac{N}{2}$ = Median item

$$\text{Med} = L_1 + \frac{N/2 - c.f}{f} \times i$$

L_1 = Lower limit of the median class

f = Frequency of the median class

$c.f$ = Cumulative frequency of the class proceeding of median class
(Total frequency of all lower classes)

i = Class interval of median class

L_2 = Upper limit of the median class

$$\frac{N}{2} = \text{Median item}$$

i = Class interval

$$\text{Med} = L_2 - \frac{N/2 - c.f}{f} \times i$$

Quartiles

1. Quartile one (Q_1)

Q_1 = Lower Quartile

Individual series and discrete series

$$\text{Size of } \left(\frac{N+1}{4} \right)^{\text{th}} \text{ item}$$

N = Number of items

2. Continuous Series

$$\text{Size of } \left(\frac{N}{4} \right)^{\text{th}} \text{ item}$$

L_1 = Lower limit of the lower quartile class

= Lower Quartile item

$$Q_1 = L_1 + \frac{N/4 - c.f}{f} \times i$$

$c.f$ = Cumulative frequency of the class proceeding of lower quartile class

F = Frequency of the lower quartile class

i = Class interval of lower quartile class

Quartile Three (Q_3)

Q_3 = Upper Quartile

1. Individual Series and Discrete Series

$$\text{Size of } 3 \left(\frac{N+1}{4} \right)^{\text{th}} \text{ item}$$

N = Number of items

2. Continuous Series

Size of $\left(\frac{3N}{4}\right)^{\text{th}}$ item

$$Q_3 = L_1 + \frac{\frac{3N}{4} - c.f}{f} \times i$$

L_1 = Lower limit of the upper Quartile class

$$\frac{3N}{4} = \text{Upper Quartile item}$$

$c.f$ = Cumulative frequency of the class proceeding of upper quartile class

F = Frequency of the upper quartile class

i = Class interval of upper quartile class

Deciles**1. Individual Series and Discrete Series**

$$D_5 = \text{Size of } 5 \left(\frac{N+1}{10} \right)^{\text{th}} \text{ item}$$

D_5 = Decile 5

2. Continuous Series

Size of $m \left(\frac{5N}{10} \right)^{\text{th}}$ item

$$D_5 = L_1 + \frac{\frac{5N}{10} - c.f}{f} \times i$$

L_1 = Lower limit of the class

$$\frac{5N}{4} = D_5 \text{ item}$$

$c.f$ = Cumulative frequency of the class proceeding of D_5 class

F = Frequency of D_5 class

i = Class interval of D_5 class

Percentiles**1. Individual and Discrete Series**

$$P_{70} = \text{Size of } \frac{70(N+1)}{100}^{\text{th}} \text{ item}$$

P_{70} = Percentile 70

N = Number of items

2. Continuous Series

P_{70} = Size of $\frac{70N^{\text{th}}}{100}$ item

$$P_{70} = L_1 + \frac{\frac{70N}{100} - c.f}{f} \times i$$

L_1 = Lower limit of P_{70} class

$\frac{70N}{100}$ = Percentile P_{70} item

$c.f$ = Cumulative frequency of the class proceeding of P_{70} class

F = Frequency of P_{70} class

i = Class interval of P_{70} class

Mode (Z)

1. Individual Series Most common value

2. Discrete Series By inspection or grouping method

3. Continuous Series By grouping

$$(Z) \text{ or } M_0 = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Z = Mode

L_1 = Lower limit of the modal class

f_1 = Frequency of the modal class

f_0 = Frequency of the class

f_2 = Frequency of the class succeeding the model class

Proceeding the modal class

i = Class interval

Alternative Formula

$$Z = L_1 + \frac{\Delta_1}{\Delta_1 - \Delta_2} \times i$$

$$\Delta_1 = f_1 - f_0$$

$$\Delta_2 = f_1 - f_2$$

4. Empirical Mode

$$Z = 3 \text{ Median} - 2 \text{ Mean}$$

Geometric Mean**1. Individual Series**

$$G.M = \text{Antilog of } \frac{\log X_1 + \log X_2 + \dots + \log X_n}{N}$$

or

$$G.M = \text{Antilog} \left(\frac{\sum \log X}{N} \right)$$

$\Sigma \log X$ = The sum of the values of $\log X$
 N = Number of items

2. Discrete Series

$$\text{G.M.} = \text{Antilog} \left(\frac{\sum f \log X}{N} \right)$$

$\Sigma f \log X$ = Total of the products
 N = The total frequency

3. Continuous Series

$$\text{G.M.} = \text{Antilog} \left(\frac{\sum f \log m}{N} \right)$$

Harmonic Mean (H.M.)

1. Individual Series

$$\text{H.M.} = \frac{N}{\frac{1}{X_1} + \frac{1}{X_2} + \dots + \frac{1}{X_n}}$$

$$\text{H.M.} = \frac{N}{\sum \frac{1}{x}}$$

X_1, X_2, \dots, X_n = Variables

2. Discrete Series

$$\text{H.M.} = \frac{N}{\sum f \left(\frac{1}{X} \right)}$$

3. Continuous Series

$$\text{H.M.} = \frac{N}{\sum f \left(\frac{1}{m} \right)}$$

4. Weighted Arithmetic Mean

$$\bar{X}_W = \frac{\sum WX}{\sum W}$$

5. Weighted Geometric Mean

$$\text{G.M.W.} = \text{Antilog} \left(\frac{\sum (\log \times W)}{\sum W} \right)$$

6. Weighted Harmonic Mean

$$\text{H.M.W.} = \frac{\sum W}{\left(\frac{1}{a} \times W_1 \right) x \left(\frac{1}{b} \times W_2 \right)}$$

7. Combined Arithmetic Mean

$$\bar{X}_{12} = \frac{N_1 \times \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$\bar{X}_{123} = \frac{N_1 \times \bar{X}_1 + N_2 \bar{X}_2 + N_3 \bar{X}_3}{N_1 + N_2 + N_3}$$

$\bar{X}_{12}; \bar{X}_{123}$ = The combined means

$\bar{X}_1; \bar{X}_2; \bar{X}_3$ = Arithmetic mean of first group, second group and third group

N_1, N_2, N_3 = Number of items in first, second and third groups

EXERCISES**(a) Choose the best option:**

1. Mean, Median and Mode are known as
 - (a) Average of position
 - (b) Mathematical Average
 - (c) Measures of Central Tendency.
2. The most popular method of measuring the representative value is:
 - (a) Arithmetic Mean
 - (b) Harmonic Mean
 - (c) Geometric Mean
3. If the lower limit of the first class interval and upper limit of the last class interval are not known, it is called
 - (a) Open-end classes
 - (b) Closed-end classes
 - (c) Mid-end classes
4. When the total number of observations are divided by the sum of reciprocals of the numbers it is known as
 - (a) Harmonic Mean
 - (b) Geometric Mean
 - (c) Arithmetic Mean
5. When the middle value of the group of data are arranged in an order either on an ascending or descending order it is known as

(a) Mean	(b) Median	(c) Mode
----------	------------	----------

6. When the values of total frequencies are divided into four parts, then it is
 - (a) Quartiles
 - (b) Quintiles
 - (c) Deciles
7. When the value of total frequencies are divided into 10 parts, then it is
 - (a) Deciles
 - (b) Percentiles
 - (c) Quintiles
8. When the value occurs most frequently in a statistical distribution, it is termed as
 - (a) Mean
 - (b) Median
 - (c) Mode
9. Which of the following can be used for averaging ratios, percentages, rate of change and index numbers?
 - (a) Geometric Mean
 - (b) Harmonic Mean
 - (c) Mode

Answers

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

(b) Fill in the blanks:

10. In a symmetrical distribution, the values of mean, median and mode will _____.
 - (a) coincide
 - (b) differ
 - (c) change
11. When the value of total frequencies are divided into eight parts, then it is _____.
 - (a) quartiles
 - (b) octiles
 - (c) percentiles
12. Ogive means
 - (a) Cumulative frequency curve
 - (b) Non-cumulative frequency curve
 - (c) Measuring of data
13. If all the items of the data are of same value, then the arithmetic mean would be equal to _____.
 - (a) Mean
 - (b) Mode
 - (c) Geometric Mean and Harmonic Mean
14. The sum of the deviations of all the observations from arithmetic mean is equal to _____.
 - (a) zero
 - (b) one
 - (c) two
15. A short way of expressing an arithmetical result is _____.
 - (a) median
 - (b) kurtosis
 - (c) mean

Answer

10. a 11. b 12. a 13. b 14. a 15. c

(c) Theoretical Questions:

1. What are the measures of central value of tendency? Describe their characteristics and state what considerations determine the use of a particular measure. **(B.Com., MKU, BDU, BU)**
2. What do you understand by the term average in statistics? Explain its significance in statistical work. **(B.Com., MSU, BDU, CHU, BU)**
3. Define mean and median. Mention its merits and demerits.
4. What is the relationship between mean, median and mode?
5. Distinguish between
 - (i) Simple mean and weighted mean
 - (ii) Geometric mean and harmonic mean
 - (iii) Crude and standardised death rates.
6. Under what circumstances are Harmonic mean most suitable?
7. Can the value of mean, median and mode be the same in symmetrical distribution? If yes, state the situation.
8. Briefly explain the role of grouping and analysing table in calculation of mode.
9. Enumerate the circumstances under which weighted average should be used in preference to simple average.
(B.Com., MKU, MSU, BDU, CHU)
10. State two important objects of measures of central value.
11. Define the following terms with an example
 - (i) Class interval
 - (ii) Class frequency
12. Prove that for any two quantities $A.M. \geq G.M \geq H.M.$
13. Define average. What are the uses of averages?
(B.Com., MSU, BU, CHU)
14. "The Arithmetic mean is the best among all the averages". Give reasons. **(B.Com., MKU, MSU, BDU)**
15. How do you determine median and mode graphically?
16. Calculate mean for the following data by Direct method.

70 65 55 75 80 85 65 70 95

Answer. 73.33

17. Calculate mean from the following data (Short-cut method)

130 170 140 200 160 140 130 180 150 190

Answer. 159

18. Calculate arithmetic mean from the following data.

55 65 75 85 90 40 50 60 25 35

Answer. 58

19. Calculate mean for the following series by Direct method.

x	:	5	10	15	20	25	30	35	40
f	:	6	17	28	34	18	11	9	7

Answer. 20.58 (B.Com., MKU, MSU)

20. Calculate mean for the following data by Short-cut method.

x	:	20	30	40	50	60	70	80
f	:	5	2	3	10	3	2	5

Answer. 50

21. Find arithmetic mean from the following data.

Wages	0–50	50–100	100–150	150–200	200–250	250–300
No. of Workers	20	25	35	28	24	19

In a factory

Answer. $\bar{x} = 147.52$ (B.Com., MKU, MSU, BDU, CHU)

22. Calculate mean from the following data.

Workers	5–15	15–25	25–35	35–45	45–55	55–65
No. of Students	8	12	6	14	7	3

Answer. 31.8

23. Find arithmetic mean from the following data.

Class Interval	3–5	6–8	9–11	12–14	15–17	18–20
Frequency	3	7	16	34	17	3

Answer. 12.4 (B.Com., MSU, BDU, BSU)

24. Find Arithmetic mean from the following series.

Marks	Below 10	10–30	30–60	60–100	100–150	Above 150
No. of Students	4	6	8	10	4	8

Answer. 81

25. Find mean from the following data.

Class Interval	Below 50	50–70	70–90	90–110	110–130	130–150
Frequency	18	34	66	110	51	21

Answer. 115.43 (B.Com., CHU, MSU, BDU)

26. Find the Geometric mean from the following data.

7 18 65 91 103

Answer. G.M. = 20.62 (B.Com., MKU, MSU)

27. Compute Median from the following data.

45 25 50 60 70 80 37 92

Answer. 55

28. Compute Median, Q_1 and Q_3 for the following data.

15	20	25	28	16	17	9	11
----	----	----	----	----	----	---	----

Answer. $m = 16.5$, $Q_1 = 12$, $Q_3 = 23.75$

29. Compute Median, Q_1 and Q_3 for the following data.

x	:	8	12	20	25	30	40
f	:	9	16	28	46	20	10

Answer. 25, 20, 25

30. Compute Median, Q_1 and Q_3 for the following data.

x	:	0–4	4–8	8–12	12–16	16–20	20–24	24–28	28–32
f	:	4	17	36	90	123	110	66	14

Answer. 18.7, 14.58, 22.73

31. Compute Mode from the following data.

Marks	16	18	22	16	15	16	14	10	11	16
--------------	----	----	----	----	----	----	----	----	----	----

Answer. 16

32. From the following data, find Mode

x	:	5	10	15	20	25	30	35	40	45
f	:	1	3	4	9	11	12	6	2	2

Answer. 25

33. Find Mode from the following data.

x	:	20–25	25–30	30–35	35–40	40–45	45–50	50–55	55–60
f	:	4	16	48	52	57	18	9	4

Answer. 37. 22 (B.Com., MKU, MSU, BDU)

34. Find arithmetic mean and median from the following.

Marks below	10	20	30	40	50	60	70	80
No. of students	15	35	60	84	96	127	198	250

Answer. $\bar{x} = 50.4$, Med.= 59.35 (B.Com., MKU, BDU, BU, CHU)

35. Average marks in statistics of 10 students of a class was 68. A new student took admission with the marks whereas the existing students left the college. If the marks of these students were 40 and 39. Find the average marks of the remaining students.

Answer. 22.22 (B.Com., MKU, MSU, BDU)

36. Find out the median and mode from the following table.

No. of days absent	No. of students
Less than 5	29
Less than 10	224
Less than 15	465
Less than 20	582
Less than 25	634
Less than 30	644

Less than 35	650
Less than 40	653
Less than 45	655

Answer. Med. 12.75, Mode = 11.35

(B.Com., MKU, MSU, CHU, BDU)

37. Find the missing frequency from the following distribution of daily sales of shops, given that the median value of shops is Rs. 2400.

Sale in Rs. '000	0–10	10–20	20–30	30–40	40–50
No. of shops	5	25	?	18	7

Answer. 25

38. In the frequency distribution of 100 families given below, the number of families corresponding to expenditure groups 20–40 and 60–80 are missing from the table. However, the median is known to be 50. Find the missing frequencies.

Expenditure	0–20	20–40	40–60	60–80	80–100
No. of families	14	?	27	?	15

Answer. 23, 21

(B.Com., MKU, MSU, CHU)

39. For a certain frequency table which has only been partly reproduced here, the mean was found to be 1.46.

No. of accidents	0	1	2	3	4	5	Total
Frequency	46	?	?	25	10	5	200

Answer. 38, 76

(B.Com., CHU, MSU, BDU)

40. An incomplete distribution is given below:

Variable	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Frequency	10	20	?	40	?	25	15

1. You are given that the median value is 35. The total frequency is 170, find out the missing frequencies.
2. Calculate the A.M. of the computed data.

7

CHAPTER

MEASURES OF DISPERSION

7.1 MEANING

The measures of central tendency give one single value which represents the entire data. It tells something about the general level of magnitude of the distribution but it fails to give its complete description. It doesn't tell how the observations in a series are distributed relative to this measure, that is, how the items scatter around the measures of central tendency.

Hence, there is a need to know the complete character of the distribution. Measures of dispersion give a clear picture of the degree of extension of the data from the central value.

7.2 DEFINITION OF DISPERSION

Brooks and Dick defines dispersion as, *Dispersion or spread is the degree of the scatter or variation of the variable about a central value.*

A.E. Bowley defines dispersion as, *Dispersion is the measure of the variations of the item.*

W.I. King defines dispersion as, *The term dispersion is used to indicate the facts that within a given group, the items differ from one another in size or in other words, there is lack of uniformity in their sizes.*

7.3 OBJECTIVES OR IMPORTANCE OF THE MEASURES OF DISPERSION

The main objectives of dispersion may be summarised as follows:

7.3.1 To Gauge the Reliability of Average

An average will be satisfactory measure of typical size only when it is derived from the data that are homogeneous. A measure of dispersion in conjunction

with a measure of central tendency gives a description of the structure of the distribution and the place of individual items in it.

7.3.2 To Control the Variation of Data from the Central Value

The second basic purpose is to determine the nature and causes of variation in order to control the variation itself. It is important not as merely supplementary to the average, but because the scatter in a distribution may itself be significant.

It helps to measure the extent of variation from the standard quality of various works carried in industries. Hence, a corrective measure can be taken after identifying the causes for the deviation.

7.3.3 To Compare two or more Series Regarding their Variability/Uniformity

On the basis of measures of dispersion, the extent of variability between two or more series can be compared. It is useful to find out the degree of uniformity in the two or more sets of data. A greater amount of dispersion means lack of uniformity in the degree.

7.3.4 To obtain other Statistical Measures for Further Analysis of Data

The measure of variation helps to make detailed analysis of data like correlation analysis, regression analysis, theory of estimation and testing of hypotheses.

7.4 METHODS OF MEASURING DISPERSION

The important methods of measuring dispersion are as follows:

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation and
5. Lorenz Curve

7.4.1 Range

It is the difference between the largest value and the smallest value of the variables (in the frequency distribution).

Hence, Range = Largest value – Smallest value.

$$R = L - S$$

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

Illustration 7.1

Find the range of weights of 10 students from the following 65, 19, 86, 15, 17, 18, 8, 4, 9, 7.

Solutions

$$\begin{aligned}\text{Range} &= L - S \\ &= 86 - 4 \\ \text{Range} &= 82\end{aligned}$$

$$\begin{aligned}\text{Coefficient of Range} &= \frac{L - S}{L + S} \\ &= \frac{86 - 4}{86 + 4} \\ &= \frac{82}{90}\end{aligned}$$

$$\text{Coefficient of Range} = 0.91$$

Merits

1. It is simple to compute and understand.
2. It gives a rough but quick answer.

Demerits

1. It is not reliable, because it is affected by the extreme items.
2. Usually frequency distribution may be concentrated in the middle of the series; but range depends on extreme items; it is an unsatisfactory measure.
3. It cannot be applied to open-ended classes.
4. It is not suitable for mathematical treatment.
5. According to King, “*Range is too indefinite to be used as a practical measure of dispersion.*”

7.4.2 Quartile Deviation

It is otherwise called as inter-quartile range. It is the difference between the third quartile and the first quartile divided by 2.

$$\text{Hence, Quartile Deviation Q.D.} = \frac{Q_3 - Q_1}{2}$$

where $Q_3 = 3^{\text{rd}}$ quartile

$Q_1 = 1^{\text{st}}$ quartile

$$\text{Co-efficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

The values of Q_3 and Q_1 may also be obtained through the following equation:

$$Q_3 = \text{Median} + Q.D$$

$$Q_1 = \text{Median} - Q.D$$

Quartile Deviation in Individual Series

Illustration 7.2

Calculate quartile deviation from the following data: 45, 35, 50, 65, 60, 40, 70 and also calculate co-efficient of quartile deviation.

Solutions

First arrange the data in ascending order; first variables are arranged 35, 40, 45, 50, 60, 65, 70 according to this order.

$$\begin{aligned} Q_1 &= \frac{N+1}{4} \text{ th item} \\ &= \frac{7+1}{4} = \frac{8}{4} = 2 \text{nd item i.e., } 40 \\ Q_3 &= \frac{3(N+1)}{4} \text{ th item} \\ &= \frac{3(7+1)}{4} = \frac{3 \times 8}{4} \\ &= \frac{24}{4} = 6 \text{th item i.e., } 65 \end{aligned}$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{65 - 40}{2} = \frac{25}{2} = 12.5$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{65 - 40}{65 + 40} = \frac{25}{105}$$

$$\text{Co-efficient of quartile deviation} = 0.238$$

Quartile Deviation under Discrete Series

Illustration 7.3

Find out the value of quartile deviation and co-efficient of quartile deviation.

Marks	40	50	60	70	80	85	90
No. of Students	2	5	10	8	7	9	8

Solutions

Calculation of Quartiles

Marks <i>x</i>	No. of students <i>f</i>	Cumulative No. of students <i>c.f</i>
40	2	2
50	5	7
60	10	17
70	8	25
80	7	32
85	9	41
90	8	49

$$Q_1 = \frac{N+1}{4} \text{ th item}$$

$$\begin{aligned}\frac{49+1}{4} &= \text{th item} \\ &= \frac{50}{4} = 12.5 \text{th item}\end{aligned}$$

12.5th item comes under cumulative frequency of 17.

i.e., $= 60 \text{ marks}$

$$\begin{aligned}Q_3 &= \frac{3(N+1)}{4} \text{ th item} \\ &= \frac{3(49+1)}{4} \text{ th item} \\ &= \frac{3 \times 50}{4} = \frac{150}{4} = 37.5 \text{th item}\end{aligned}$$

37.5th item comes under the cumulative frequency of 41.

i.e., $= 85 \text{ marks}$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{85 - 60}{2} = \frac{25}{2} = 12.5 \text{ marks}$$

$$\begin{aligned}\text{Co-efficient of Q.D.} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\ &= \frac{85 - 60}{85 + 60} = \frac{25}{145}\end{aligned}$$

$$\text{Co-efficient of Q.D.} = 0.17$$

Quartile Deviation under Continuous Series

Illustration 7.4

Compute quartile deviation and co-efficient of quartile deviation from the following data.

x	20–40	40–60	60–80	80–100	100–120
f	15	22	25	19	25

Solutions

Calculation of Q.D.

x	f	c.f
20–40	15	15
40–60	22	36
60–80	25	62
80–100	19	81
100–120	25	106

$$Q_1 = \frac{N}{4} \text{ th item} = \frac{106}{4} \text{ th item}$$

$$Q_1 = 26.5 \text{th item}$$

which (26.5th item) lies in 40–60. Hence, $L = 40$, $c.f = 15$, $f = 22$ and $i = 20$ (L = Lower limit; $c.f$ = Previous cumulative frequency; F = Current frequency; i = Class interval)

$$\begin{aligned} Q_1 &= L + \left[\frac{N/4 - cf}{f} \right] \times i \\ &= 40 + \frac{26.5 - 15}{22} \times 20 \\ &= 40 + \frac{11.5}{22} \times 20 \\ &= 40 + 10.45 = 50.45 \end{aligned}$$

$$Q_3 = \frac{3N}{4} \text{ th item} = \frac{3 \times 106}{4}$$

$$\frac{318}{4} = 79.5 \text{th item}$$

which lies in 80 – 100. Hence, $L = 80$, $c.f = 62$, $f = 19$, $i = 20$

$$Q_3 = L + \left(\frac{3N/4 - cf}{f} \right) \times i$$

$$\begin{aligned}
 &= 80 + \left(\frac{79.5 - 62}{19} \right) \times 20 \\
 &= 80 + \frac{79.5}{19} \times 20 = 80 + 18.42 = 98.42
 \end{aligned}$$

$$\begin{aligned}
 \text{Co-efficient of Q.D.} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\
 &= \frac{98.42 - 50.45}{98.42 + 50.45} = \frac{47.97}{148.87} = 0.32
 \end{aligned}$$

7.4.3 Mean Deviation or Average Deviation

Range and Quartile deviation are calculated with the help of two data of a series. For the measurement of dispersion, only two extreme values are considered in the calculation of range and in quartile deviation only quartiles are taken.

Mean Deviation is a measure of deviation based on all the items in a distribution. While calculating the sum of the deviation, minus (-) should be ignored and it should be treated as positive (+) value.

Calculation of Mean Deviation

Mean Deviation for Individual Observations

$$\text{M.D.} = \frac{\sum |D|}{N}$$

$\sum |D|$ indicates the sum of deviations from the mean or median or mode. $|D|$ can be read as modules D . N indicates the total frequencies.

$$\text{Co-efficient of Mean Deviation} = \frac{\text{M.D.}}{\text{Mean or Median or Mode}}$$

For calculating co-efficient of M.D., divide M.D. by mean if deviation are taken from arithmetic mean or if deviations are taken from median by median or if deviations are taken from mode by mode.

Mean Deviation for Discrete Series

$$\text{M.D.} = \frac{\sum f |D|}{N}$$

where f represents frequency.

Here the deviations from mean or median or mode of the variables should be multiplied by respective frequencies and divided by sum of frequency.

Mean Deviation for Continuous Series

$$\text{M.D.} = \frac{\sum f |D|}{N}$$

The deviations of the mid points of class intervals from mean or median or mode should be multiplied by the respective frequencies. Then the sum of these deviations

should be divided by the total number of frequencies to get the mean deviation.

Illustration 7.5

Find out the value of mean deviation and its co-efficient from the data given below :

40, 60, 35, 10, 70, 80, 15, 20, 30, 20

Solutions

Calculation of mean deviation

X	 D = (x - \bar{x}) (x - 38)
40	2
60	22
35	3
10	28
70	32
80	42
15	23
20	18
30	8
20	18
$\Sigma X = 380$	$\Sigma D = 196$

$$\bar{x} = \frac{\sum x}{N} = \frac{380}{10} = 38$$

$$\text{M.D.} = \frac{\sum |D|}{N} = \frac{196}{10} = 19.6$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{Mean}} = \frac{19.6}{38} = 0.516$$

Illustration 7.6

Calculate mean deviation from mean, median and mode from the following data. Also calculate the co-efficient of mean deviation.

55, 15, 20, 55, 35, 55, 45

Solutions

Calculation of M.D.

x	$ D = x - \bar{x}$ ($x - 40$)	$ D = x - \text{Median}$ ($x - 45$)	$ D = x - \text{mode}$ ($x - 55$)
15	25	30	40
20	20	25	35
35	5	10	20
45	5	0	10
55	15	10	0
55	15	10	0
55	15	10	0
$\Sigma x = 280$	$\Sigma D = 100$	$\Sigma D = 95$	$\Sigma D = 105$

$$\sum x = 280$$

$$\bar{x} = \frac{\sum x}{N} = \frac{280}{7} = 40$$

$$\text{Median} = \frac{N+1}{2} \text{ th item}$$

$$= \frac{7+1}{2} \text{ th item} = 4\text{th item i.e., } 45$$

$$\text{M.D. based on Mean} = \frac{\sum |D|}{N} = \frac{100}{7} = 14.3$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{Mean}} = \frac{14.3}{40} = 0.36$$

$$\text{M.D. based on Median} = \frac{\sum |D|}{N} = \frac{95}{7} = 13.57$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{Median}} = \frac{13.57}{45} = 0.302$$

$$\text{M.D. based on Mode} = \frac{\sum |D|}{N} = \frac{105}{7} = 15$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{Mode}} = \frac{15}{55} = 0.27$$

Illustration 7.7

The details regarding the sales of goods in factories situated in Q city are given below.

Sales (in tonnes)	100	200	300	400	500	600
No. of factories	2	8	11	10	6	5

Calculate the mean deviation and its co-efficient.

Solutions

Calculation of M.D.

Sales	No. of factories	Deviations from assumed mean $d(x - 300)$	fd
x	f		
100	2	- 200	- 400
200	8	- 100	- 800
300	11	0	0
400	10	+ 100	+ 1000
500	6	+ 200	+ 1200
600	5	+ 300	+ 1500
$N = 42$			$\Sigma fd = 2500$

$$\text{Arithmetic Mean } \bar{x} = A \pm \frac{\sum fd}{N}$$

$$= 300 + \frac{2500}{42} = 359.52$$

$ D = (x - \bar{x})$	$f D $
+ 259.52	519.04
159.52	1276.16
59.52	654.72
40.48	404.80
140.48	842.88
240.48	1202.40
$\Sigma f D = 4900$	

$$\text{M.D.} = \frac{\sum f|D|}{N} = \frac{4900}{42} = 116.67$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{Mean}} = \frac{116.67}{359.52} = 0.32$$

Illustration 7.8

Calculate mean deviation and co-efficient of mean deviation from the following data.

<i>x</i>	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
<i>f</i>	14	18	20	25	30	15	19	21

Solutions

<i>x</i>	<i>f</i>	<i>m</i>	$\frac{m - 45}{10}$ <i>d'</i>	<i>fd'</i>	$m - \bar{x}$ (<i>m</i> – 41.36)	<i>f D </i>
0–10	14	5	-4	-56	36.36	509.04
10–20	18	15	-3	-54	26.36	474.48
20–30	20	25	-2	-40	16.36	327.20
30–40	25	35	-1	-25	6.36	159.00
40–50	30	45	0	0	3.64	0
50–60	15	55	+1	+15	13.64	204.60
60–70	19	65	+2	+38	23.64	449.16
70–80	21	75	+3	+63	33.64	706.44
<i>N = 162</i>			$\Sigma fd' = -59$		$\Sigma D = 2829.92$	

$$\bar{x} = A \pm \frac{\sum fd'}{N} \times i = 45 - \frac{59 \times 10}{162} = 45 - 3.64 = 41.36$$

$$\text{M.D.} = \frac{\sum f|D|}{N} = \frac{2829.92}{162} = 17.469$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{Mean}} = \frac{17.469}{41.36} = 0.4224$$

Advantages of Mean Deviation

1. It is easy to calculate.
2. It is less affected by extreme observations.
3. It is not much affected by fluctuations of sampling.
4. It is more comprehensive measure than range and quartile deviation, as it is based on all the observations of the series.

Disadvantages of Mean Deviation

1. It is not well-defined measure of dispersion since deviations can be taken from any measure of central tendency.
2. It is not suitable for further statistical analysis.
3. It is a non-algebraic method since the positive and negative signs for the deviations are treated as positive only.

7.4.4 Standard Deviation

Standard Deviation is the square root of the sum of squared deviations from average. It is a prominent and widely used measure of dispersion.

In this method, deviations should be taken only from the arithmetic mean. It is denoted by the small Greek letter sigma (σ).

Calculation of Standard Deviation for Individual Observation

$$\text{S.D. } \sigma = \sqrt{\frac{\sum x^2}{N}} \text{ or } \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

$x \rightarrow$ Deviations from arithmetic mean, i.e., $(x - \bar{x})$

$N \rightarrow$ Total number of variables

If the observation is a complicated one, then it takes time to find out the arithmetic mean. To overcome this difficulty, short-cut method can be adopted to calculate the standard deviation.

$$\text{S.D. } \sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N} \right)^2}$$

$d \rightarrow$ Refers to deviation from Assumed mean i.e., $(x - A)$

Calculation of Standard Deviation for Discrete Series

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

where, σ = Standard deviation

f = frequencies

x = deviations from arithmetic mean i.e., $x - \bar{x}$

N = Total number of frequencies i.e., $\sum f$.

If the calculation of A.M is difficult, then the short-cut method can be applied to calculate the standard deviation.

Formula for S.D. under short-cut method is

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2}$$

$d \rightarrow$ Deviations from assumed mean i.e., $d = x - A$

$f \rightarrow$ Frequency

$N \rightarrow$ Total number of frequencies i.e., Σf .

If the variables got common factor, then the calculation for standard deviation may further be simplified through step deviation method.

Formula for S.D. under Step-deviation method

$$\sigma = \sqrt{\frac{\sum fd'}{N} - \left(\frac{\sum fd'}{N} \right)^2} \times C$$

196 Business Statistics

$d \rightarrow$ Deviations from assumed mean divided by a common factor i.e.,

$$d' = \frac{x - A}{C}$$

$C \rightarrow$ Common factor for the observations

$f \rightarrow$ Frequencies

$N \rightarrow$ Total number of frequencies i.e., $N = \sum f$

Calculation of Standard Deviation for Continuous Series In this series, the mid points of the class intervals should be taken into account instead of the variable.

(i) Direct Method

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

$x \rightarrow$ Deviation from arithmetic mean i.e., $(x = x - \bar{x})$

$f \rightarrow$ Frequency

$N \rightarrow$ Total number of frequencies i.e., $N = \sum f$

(ii) Short-cut Method

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2}$$

$d \rightarrow$ Deviations of mid-points from the assumed mean i.e., $d = m - A$

(iii) Step Deviation Method

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N} \right)^2} \times C$$

$d' \rightarrow$ Deviation of mid-points from assumed mean divided by a common factor

$$\text{i.e., } d' = \frac{M - A}{C}$$

$C \rightarrow$ Common Factor

It is the most common method adopted for calculating standard deviation in continuous series.

Illustration 7.9

Calculate standard deviation from the following data.

29, 47, 38, 74, 65, 92, 56, 83, 101

Solutions

Calculation of standard deviation.

x	$x = (x - 65)$	x^2
29	-36	1296
47	-18	324
38	-27	729
74	9	81
65	0	0
92	27	729
56	-9	81
83	18	324
101	36	1296
$\Sigma x = 585$		$\Sigma x^2 = 4860$

$$\bar{x} = \frac{\sum x}{N} = \frac{585}{9} = 65$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{4860}{9}} = \sqrt{540}$$

$$\sigma = 23.24$$

Illustration 7.10

Calculate standard deviation from the following data:

29, 26, 42, 65, 92, 83, 87

Solutions

Short-cut method

Calculation of standard deviation

x	Assumed mean $A = 60$ $d = (x - 60)$	d^2
29	-31	961
26	-34	1156
42	-18	324

x	Assumed mean $A = 60$ $d = (x - 60)$	d^2
65	+ 5	25
92	+ 32	1024
83	+ 23	529
87	+ 27	729
	$\Sigma d = 4$	$\Sigma d^2 = 4748$

$$\begin{aligned}\sigma &= \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2} \\ &= \sqrt{\frac{4748}{7} - \left(\frac{4}{7}\right)^2} \\ &= \sqrt{678.29 - 0.32} = \sqrt{677.97} = 26.038\end{aligned}$$

Illustration 7.11

Find out standard deviation for the following data.

Sales (in tonnes)	100	150	175	200	250	300	350
No. of factories	4	10	14	24	18	10	6

Solutions

Calculation of standard deviation by step deviation method.

Sales (in tonnes) x	f	$d' = (x - 200)$ 50	d'^2	fd'	fd'^2
100	4	- 2	4	- 8	16
150	10	- 1	1	- 10	10
175	14	- 0.5	0.25	- 7	3.5
200	24	0	0	0	0
250	18	+ 1	1	+ 18	18
300	10	+ 2	4	+ 20	40
350	6	+ 3	9	+ 18	54
$N = 86$			$\Sigma fd' = 31$	$\Sigma fd'^2 = 141.50$	

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C \\ &= \sqrt{\frac{141.50}{86} - \left(\frac{31}{43}\right)^2} \times 50 \\ &= \sqrt{1.65 - 0.52} \times 50 \\ &= \sqrt{1.13} \times 50 = 1.06 \times 50 = 53\end{aligned}$$

Illustration 7.12

Calculate standard deviation from the following data:

Marks	10	20	30	40	50	60	70
No. of students	6	5	12	3	5	4	5

Solutions

Marks	No. of students	fx	$x = (x - \bar{x})$	x^2	fx^2
x			$x - 37$		
10	6	60	-27	729	4374
20	5	100	-17	289	1445
30	12	360	-7	49	588
40	3	120	3	9	27
50	5	250	13	169	845
60	4	240	23	529	2116
70	5	350	33	1089	5445
$N = 40$		$\Sigma fx = 1480$			$\Sigma fx^2 = 14840$

$$\bar{x} = \frac{\sum fx}{N} = \frac{1480}{40} = 37$$

$$\sigma = \sqrt{\frac{\sum fx^2}{N}} = \sqrt{\frac{14840}{40}} = \sqrt{371} = 19.26$$

Illustration 7.13

Calculate standard deviation for the information given below.

Height (in cm)	147	150	152	154	156	158
No. of students	2	3	9	4	7	3

Solutions

Calculation of S.D. through short-cut method.

Height (in cm) <i>x</i>	No. of students <i>f</i>	<i>A</i> = 152 <i>d</i> = (<i>x</i> – <i>A</i>) = <i>x</i> – 152	<i>d</i> ²	<i>fd</i>	<i>fd</i> ²
147	2	– 5	25	– 10	50
150	3	– 2	4	– 6	12
152	9	0	0	0	0
154	4	+ 2	4	8	16
156	7	+ 4	16	28	112
158	3	+ 6	36	18	108
<i>N</i> = 28				$\Sigma fd = 38$	$\Sigma fd^2 = 298$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{298}{28} - \left(\frac{38}{28}\right)^2} \\ = \sqrt{10.64 - 1.85} = \sqrt{8.79} = 2.96$$

Illustration 7.14

Compute the standard deviation from the following data.

Expenditure (Rs)	50–100	100–150	150–200	200–250	250–300
No. of families	20	10	30	5	10

Solutions

Calculation of standard deviation by direct method.

Expenditure (Rs) <i>x</i>	No. of families <i>f</i>	<i>m</i>	<i>fm</i>	<i>x</i> = <i>m</i> – \bar{x} <i>m</i> – 158.3	<i>x</i> ²	<i>fx</i> ²
50–100	20	75	1500	– 83.3	6938.89	138777.80
100–150	10	125	1250	– 33.3	1108.89	11088.90
150–200	30	175	5250	16.7	278.89	8366.70
200–250	5	225	1125	66.7	4448.89	22244.45
250–300	10	275	2750	116.7	13618.89	136188.90
<i>N</i> = 75			Σfm = 11875		Σfx^2 = 316666.75	

$$\bar{x} = \frac{\sum fm}{N} = \frac{11875}{75} = 158.33$$

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

$$= \sqrt{\frac{316666.75}{75}} = \sqrt{4222.22} = 64.98$$

Illustration 7.15

Calculate standard deviation from the following data:

x	25–50	50–75	75–100	100–125	125–150	150–175
f	15	14	21	13	17	19

Solutions

Calculation of standard deviation by short-cut method.

x	f	m	d = m - A m - 112.5	fd	d²	fd²
25 – 50	15	37.5	- 75	- 1125	5625	84375
50 – 75	14	62.5	- 50	- 700	2500	35000
75 – 100	21	87.5	- 25	- 525	625	13125
100 – 125	13	112.5	0	0	0	0
125 – 150	17	137.5	+ 25	+ 425	625	10625
150 – 175	19	162.5	+ 50	+ 950	2500	47500
N = 99				Σfd = - 975		Σfd² = 190625

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{190625}{99} - \left(\frac{-975}{99}\right)^2}$$

$$= \sqrt{1925.51 - 96.99} = \sqrt{1828.52} = 42.76$$

Illustration 7.16

Calculate standard deviation from the following distribution:

x	50–100	100–150	150–200	200–250	250–300	300–350
f	10	13	8	18	7	4

Solutions

Calculation of S.D. by step-deviation method.

<i>x</i>	<i>f</i>	<i>m</i>	<i>d' = m - A/c</i> <i>(m - 225/c)</i>	<i>fd'</i>	<i>d'^2</i>	<i>fd'^2</i>
50–100	10	75	-3	-30	9	90
100–150	13	125	-2	-26	4	52
150–200	8	175	-1	-8	1	8
200–250	18	225	0	0	0	0
250–300	7	275	1	7	1	7
300–350	4	325	2	8	4	16
<i>N = 60</i>			<i>Σfd' = -49</i>		<i>Σfd'^2 = 173</i>	

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C \\ &= \sqrt{\frac{173}{60}} = \sqrt{\left(\frac{-49}{60}\right)^2} \times 50 \\ &= \sqrt{2.88 - 0.82} \times 50 = \sqrt{2.06} \times 50 \\ &= 1.4353 \times 50 = 71.76\end{aligned}$$

Mathematical Properties of Standard Deviation The important mathematical properties of standard deviation are:

- (i) Combined Standard Deviation
- (ii) Coefficient of variation and
- (iii) Variance

(i) Combined Standard Deviation It can be calculated by following the same method of calculating the combined mean of two or more than two groups. It may be denoted by σ_{12} .

Formula for combined standard deviation of two groups is

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

or

$$\sigma_{12} = \sqrt{\frac{N_1 (\sigma_1^2 + d_1^2) + N_2 (\sigma_2^2 + d_2^2)}{N_1 + N_2}}$$

where,

- σ_{12} denotes combined standard deviation
- σ_1 denotes standard deviation of the first group
- σ_2 denotes standard deviation of the second group
- d_1 denotes the difference between the arithmetic mean of first group with the combined arithmetic mean of the two groups, that is, $(\bar{x}_1 - \bar{x}_{12})$
- d_2 denotes the difference between the arithmetic mean of the second group with the combined arithmetic mean of the two groups, that is, $(\bar{x}_2 - \bar{x}_{12})$
- N_1 denotes the total number of variables of the first group
- N_2 denotes the total number of variables of the second group

$$\bar{x}_{12} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$$

where,

\bar{x}_{12} → Combined arithmetic mean

\bar{x}_1 → Arithmetic mean of the first group

\bar{x}_2 → Arithmetic mean of the second group

The combined standard deviation of three or more groups can also be calculated by modifying the formula for σ_{12} .

Formula for combined standard deviation of three groups is

$$\sigma_{123} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_3 \sigma_3^2 + N_1 d_1^2 + N_2 d_2^2 + N_3 d_3^2}{N_1 + N_2 + N_3}}$$

or

$$\sqrt{\frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2) + N_3(\sigma_3^2 + d_3^2)}{N_1 + N_2 + N_3}}$$

Illustration 7.17

From the following particular calculate combined standard deviation.

$$N_1 = 80; \bar{x}_1 = 45; \sigma_1 = 9$$

$$N_2 = 70; \bar{x}_2 = 55; \sigma_2 = 11$$

Solutions

Hence,

$$\bar{x}_{12} = \frac{N_1 \bar{x}_1 + N_2 \bar{x}_2}{N_1 + N_2}$$

$$\begin{aligned}
 &= \frac{[80 \times 45] + [70 \times 55]}{80 + 70} \\
 &= \frac{3600 + 3850}{150} = \frac{7450}{150} = 49.67
 \end{aligned}$$

$$\bar{x}_{12} = 49.67$$

$$d_1 = \bar{x}_1 - \bar{x}_{12} = 45 - 49.67 = -4.67$$

$$d_2 = \bar{x}_2 - \bar{x}_{12} = 55 - 49.67 = 5.33$$

∴ Combined S.D. or σ_{12}

$$\begin{aligned}
 &= \sqrt{\frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N_1 + N_2}} \\
 &= \sqrt{\frac{80[(9)^2 + (-4.67)^2] + 70[(11)^2 + (5.33)^2]}{80 + 70}} \\
 &= \sqrt{\frac{80[81 + 21.81] + 70[121 + 28.41]}{150}} \\
 &= \sqrt{\frac{80[102.81] + 70[149.41]}{150}} \\
 &= \sqrt{\frac{8224.8 + 10458.7}{150}} = \sqrt{\frac{18683.5}{150}} \\
 &= \sqrt{124.56} = 11.16
 \end{aligned}$$

$$\sigma_{12} = 11.16$$

Illustration 7.18

The details regarding the marks obtained by students of 3 departments in a subjects are stated below.

Department	No. of students	Mean Marks	Standard Deviation
Maths	40	65	8.2
Commerce	30	38	6.8
Economics	50	45	7.7

Calculate the combined standard deviation.

Solutions

$$N_1 = 40; N_2 = 30; N_3 = 50; \bar{x}_1 = 65; \bar{x}_2 = 38; \bar{x}_3 = 45; \sigma_1 = 8.2; \sigma_2 = 6.8; \sigma_3 = 7.7$$

$$\begin{aligned}\bar{x}_{123} &= \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3} \\ &= \frac{(40 \times 65) + (30 \times 38) + (50 \times 45)}{40 + 30 + 50} \\ &= \frac{2600 + 1140 + 2250}{120} = \frac{5990}{120} = 49.92\end{aligned}$$

$$d_1 = \bar{x}_1 - \bar{x}_{123} = 65 - 49.92 = 15.08$$

$$d_2 = \bar{x}_2 - \bar{x}_{123} = 38 - 49.92 = -11.92$$

$$d_3 = \bar{x}_3 - \bar{x}_{123} = 45 - 49.92 = -4.92$$

$$\begin{aligned}\sigma_{123} &= \sqrt{\frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2) + N_3(\sigma_3^2 + d_3^2)}{N_1 + N_2 + N_3}} \\ &= \sqrt{\frac{40[(8.2)^2 + (15.08)^2] + 30[(6.8)^2 + (-11.92)^2] + 50[(7.7)^2 + (-4.92)^2]}{40 + 30 + 50}} \\ &= \sqrt{\frac{40(67.24 + 227.41) + 30(46.24 + 142.09) + 50(59.29 + 24.21)}{120}} \\ &= \sqrt{\frac{(40 \times 294.65) + (30 \times 188.33) + (50 \times 83.5)}{120}} = \sqrt{\frac{21610.90}{120}} = \sqrt{180.09} = 13.42\end{aligned}$$

(ii) Co-efficient of Variation The relative measure of standard deviation is called the co-efficient of variation. Standard deviation is only an absolute measure of dispersion. The consistency, reliability, uniformity or stability of data can be known through the standard deviation.

Hence, the co-efficient of variation technique is adopted. If the co-efficient of variation is greater then the data are more variable and less consistent, less uniform, less homogeneous, and less stable and vice versa.

The co-efficient of variation can be measured through the following formula:

$$C.V. = \frac{\sigma}{\bar{x}}$$

where σ = Standard deviation

\bar{x} = Arithmetic mean and

C.V. = Co-efficient of variation

If it is termed in percentage, then

$$C.V. = \frac{\sigma}{\bar{x}} \times 100$$

This formula was suggested by Karl Pearson.

(iii) Variance It is the square of standard deviation. Hence, variance is σ^2 or $\sigma = \sqrt{\text{Variance}}$.

Relation between Standard Deviation and other Measures of Dispersion
There exists a relationship between the standard deviation and quartile deviation and between standard deviation and mean deviation.

Of these deviations, quartile deviation is the smallest and standard deviation is the largest one. The mean deviation lies between the standard deviation and the quartile deviation.

$$Q.D. = \frac{2}{3} \sigma \text{ and } M.D. = \frac{4}{5} \sigma$$

$$\text{Hence, } \sigma = \frac{3}{2} Q.D. \text{ or } \sigma = \frac{5}{4} M.D.$$

$$\text{Therefore, } \frac{3}{2} Q.D. = \frac{5}{4} M.D., M.D. = \frac{6}{5} Q.D.$$

Illustration 7.19

The data relating to the yearly sales of a product in two factories are given below:

Yearly sales in tonnes

Factory	1	2	3	4	5	6	7	8	9	10	11	12
P	300	500	450	540	490	530	600	460	410	560	590	450
Q	700	1200	200	150	1500	1000	900	800	100	250	950	850

Find out which factory is more efficient and which factory is more consistent.

Solutions

Calculation of co-efficient of variation for factory P.

Sales <i>X</i>	$x - \bar{X} = x$ $x = (X - 490)$	x^2
300	-190	36100
500	10	100
450	-40	1600

Contd.

Sales <i>X</i>	$x - \bar{x} = x$ $x = (X - 490)$	x^2
540	50	2500
490	0	0
530	40	1600
600	110	12100
460	-30	900
410	-80	6400
560	70	4900
590	100	10000
450	-40	1600
$\sum x = 5880$		$\sum x^2 = 77800$

$$\bar{X} = \frac{\sum x}{N} = \frac{5880}{12} = 490$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum x^2}{N}} \\ &= \sqrt{\frac{77800}{12}} = \sqrt{6483.33} = 80.52\end{aligned}$$

$$\text{Co-efficient of variation} = \frac{\sigma}{\bar{x}} = \frac{80.52}{490} = 0.16$$

Calculation of co-efficient of variation for factory *Q*.

<i>X</i>	$x = X - \bar{x}$ $(X - 700)$	x^2
700	0	0
1200	500	250000
200	-500	250000
150	-550	302500
1300	600	360000
1000	300	90000
900	200	40000
800	100	10000
100	-600	360000
250	-450	202500
950	250	62500
850	150	22500
$\sum X = 8400$		$\sum x^2 = 1950000$

$$\bar{x} = \frac{\sum x}{N} = \frac{8400}{12} = 700$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{1950000}{12}} = \sqrt{162500} = 403.11$$

$$\begin{aligned}\text{Co-efficient of variation} &= \frac{\sigma}{\bar{x}} \\ &= \frac{403.11}{700} = 0.58\end{aligned}$$

Answer Factory Q is more efficient than factory P in terms of sales of the goods, since the average yearly sales for factory Q is more than that of P . As far as consistency of sales is concerned, factory P is more consistent than factory Q , since co-efficient of variation for factory P is less than that of the factory Q .

Illustration 7.20

Monthly expenditure of the sample families in two towns are given below.

Monthly Expenditure (Rs.)	No. of Families	
	Town S	Town T
600	24	15
850	17	16
1100	13	17
1350	21	18
1600	19	17
1850	14	19
2100	12	18

Find out which town spends more and which one is spending more consistently.

Solutions

Calculation of co-efficient of variation for town S .

Monthly Expenditure X	No. of Families f	$d' = \frac{X - A}{C}$ $\left(\frac{X - 1350}{250} \right)$	d'^2	fd'	fd'^2
600	24	-3	9	-72	216
850	17	-2	4	-34	68

Monthly Expenditure <i>X</i>	No. of Families <i>f</i>	$d' = \frac{X - A}{C}$ $\left(\frac{X - 1350}{250} \right)$	d'^2	fd'	fd'^2
1100	13	-1	1	-13	13
1350	21	0	0	0	0
1600	19	1	1	19	19
1850	14	+2	4	28	56
2100	12	+3	9	36	108
$N = 120$				$\sum fd' = -36$	$\sum fd'^2 = 480$

$$\bar{X} = A \pm \frac{\sum fd'}{N} \times C$$

$$= 1350 \pm \frac{-36}{120} \times 250 = 1350 - 75 = 1275$$

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N} \right)^2} \times C$$

$$= \sqrt{\frac{480}{120} - \left(\frac{-36}{120} \right)^2} \times 250$$

$$= 1.98 \times 250 = 495$$

Calculation of co-efficient of variation for town *T*.

Monthly Expenditure Rs. Expenditure	No. of families <i>f</i>	$d' = \frac{X - A}{C}$ $\left(\frac{X - 1350}{250} \right)$	d'^2	fd'	fd'^2
600	15	-3	9	-45	135
850	16	-2	4	-32	64
1100	17	-1	1	-17	17
1350	18	0	0	0	0
1600	17	+1	1	+17	17
1850	19	+2	4	+38	76
2100	18	+3	9	+54	162
$N = 120$				$\sum fd' = 15$	$\sum fd'^2 = 471$

210 Business Statistics

$$\begin{aligned}\bar{x} &= A + \frac{\sum fd'}{N} \times C = 1250 + \frac{15}{120} \times 250 \\ &= 1250 + 31.25 = \text{Rs.}1281.25 \\ \sigma &= \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C = \sqrt{\frac{471}{120} - \left(\frac{15}{120}\right)^2} \times 250 \\ &= \sqrt{3.93 - 0.02} \times 250 = \sqrt{3.91} \times 250 = 1.98 \times 250 = 495 \\ \text{C.V.} &= \frac{\sigma}{\bar{x}} = \frac{495}{1281.25} = 0.3863\end{aligned}$$

Answer Town *S* spends more than *T*. Also, town *S* is more consistent than *T*.

Illustration 7.21

Following are the information related with the marks obtained by 20 students of a class in two subjects. Find out which one got the homogeneity and which one got the highest marks in the class.

Marks	0–10	10–20	20–30	30–40	40–50
Subject I	4	2	3	8	3
Subject II	5	4	6	3	2

Solutions

Calculation of co-efficient of variation.

Marks points	Mid <i>m</i>	<i>f</i> ₁ Subject I	<i>f</i> ₂ Subject II	$\left(\frac{X - 25}{5}\right)$	<i>d'</i>	<i>f</i> ₁ <i>d'</i>		<i>f</i> ₂ <i>d'</i>		<i>f</i> ₁ <i>d'</i> ²		<i>f</i> ₂ <i>d'</i> ²	
						<i>f</i> ₁ <i>d'</i>	<i>f</i> ₂ <i>d'</i>	<i>f</i> ₁ <i>d'</i> ²	<i>f</i> ₂ <i>d'</i> ²				
0–10	5	4	5	-4	-4	16	-16	-20	64	80			
10–20	15	2	4	-2	-2	4	-4	-8	8	16			
20–30	25	3	6	0	0	0	0	0	0	0			
30–40	35	8	3	+2	+2	4	+16	+6	32	12			
40–50	45	3	2	+4	+4	16	+12	+8	48	32			
		20	20			$\sum f_1 d'$	$\sum f_2 d'$	$\sum f_1 d'^2$	$\sum f_2 d'^2$				
						= 8	= -14	= 152	= 140				

$$A \pm \frac{\sum f_1 d'}{N} \times C' = 25 + \frac{8}{20} \times 5 = 25 + 2 = 27$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum f_1 d'^2}{N} - \left(\frac{\sum f_1 d'}{N}\right)^2} \times C \\ &= \sqrt{\frac{152}{20} - \left(\frac{8}{20}\right)^2} \times 5 = \sqrt{7.6 - 0.16} \times 5 = \sqrt{7.44} \times 5 = 2.723 \times 5 = 13.65 \\ \text{C.V.} &= \frac{\sigma}{\bar{x}} = \frac{13.65}{27} = 0.51 \\ \bar{x}_2 &= A \pm \frac{\sum f_2 d'}{N} \times C = 25 - \frac{14}{20} \times 5 = 25 - 3.5 = 21.5 \\ \sigma &= \sqrt{\frac{\sum f_2 d'^2}{N} - \left(\frac{\sum f_2 d'}{N}\right)^2} \times C \\ &= \sqrt{\frac{140}{20} - \left(\frac{-14}{20}\right)^2} \times 5 = \sqrt{7 - 0.49} \times 5 = \sqrt{6.51} \times 5 \\ &= 2.55 \times 5 = 12.75 \\ \text{C.V.} &= \frac{\sigma}{\bar{x}} = \frac{12.75}{21.5} = 0.59\end{aligned}$$

Therefore, subject I got the homogeneity and also got the highest marks in the class.

7.4.5 Lorenz Curve

Lorenz curve is a graphic method of measuring dispersion. This method was devised by Max O. Lorenz, a famous economic statistician. It can be drawn by converting the value into percentages.

Procedure for Drawing Lorenz Curve

1. Calculate the cumulative values of variables and frequencies.
2. Calculate the cumulative percentage of all the variables and frequencies.
3. X-axis should be represented with cumulative percentage of variables with a scale of 0 to 100.
4. Y-axis should be represented with cumulative percentage of frequencies with a scale of 0 to 100.
5. Plot the percentage of cumulative variables against the percentages of corresponding cumulative frequencies. Join all the plotted points to get a curve.
6. Also draw a 45% straight line on the graph in order to know the extent of deviation of variables from the central value.

212 Business Statistics

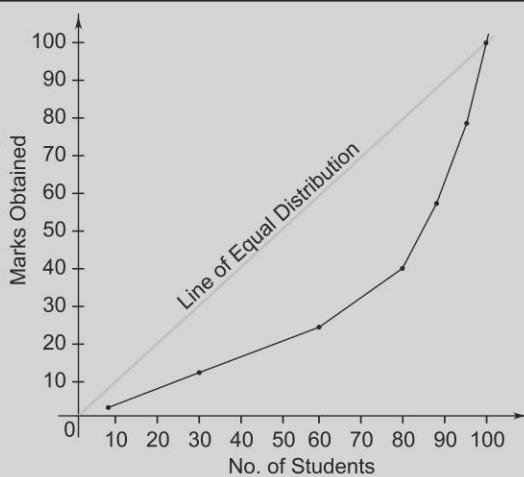
Illustration 7.22

Draw a Lorenz curve from the following data.

Marks	10	40	50	60	70	80	90
No. of Students	8	22	30	20	8	4	8

Solutions

Marks	Marks			Students		
	Marks	Cumulative Marks	Cummulative % marks	No. of students	Cumulative students	Cumulative %
10	10	2.5	8	8	8	8
40	50	12.5	22	30	30	30
50	100	25	30	60	60	60
60	160	40	20	80	80	80
70	230	57.5	8	88	88	88
80	310	77.5	4	92	92	92
90	400	100	8	100	100	100

**7.5 MISCELLANEOUS ILLUSTRATIONS****Illustration 7.23**

Calculate range from the following data.

20, 60, 80, 18, 35, 45, 65, 72, 82, 90, 101, 108

Solutions

$$\begin{aligned}\text{Range} &= L - S \\ &= 108 - 18 = 90\end{aligned}$$

Illustration 7.24

Calculate range and co-efficient of range from the data.

250, 275, 135, 152, 173, 129, 186, 157

Solutions

Calculation of range and co-efficient of range.

$$\text{Range} = L - S$$

$$\text{when, } L = 275; S = 129$$

$$\text{Range} = 275 - 129 = 146$$

$$\text{Co-efficient of Range} = \frac{L - S}{L + S} = \frac{275 - 129}{275 + 129} = \frac{146}{404} = 0.36$$

Illustration 7.25

Calculate quartile deviation and co-efficient of quartile deviation from the data given below.

Weight of scholars in k.gms. 57.5, 52.2, 63.4, 69.4, 49.5, 50.4, 56.7

Solutions

Quartile deviation and its co-efficient weight of scholars arranged in an order.

49.5, 50.4, 52.2, 56.7, 57.5, 63.4, 69.4

$$Q_1 = \text{size of } \frac{N+1}{4} \text{ th item} = \frac{7+1}{4} \text{ th item} = 2 \text{nd item}$$

Size of second item is 50.4 = Q_1

$$Q_3 = \text{size of } 3\left(\frac{N+1}{4}\right) \text{ th item} = 3\left(\frac{7+1}{4}\right) \text{ th item} = 6 \text{ th item}$$

Size of the sixth item is 63.4 = Q_3

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{63.4 - 50.4}{2} = \frac{13}{2} = 6.5$$

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63.4 - 50.4}{63.4 + 50.4} = \frac{13}{113.8} = 0.11$$

Quartile Deviation = 6.5, Co-efficient of Q.D. = 0.11

Illustration 7.26

Compute quartile deviation from the following data :

Height in inches	68	69	70	71	72	73	74	75	76
No. of students	25	30	42	45	43	32	30	20	16

(B. Com, MKU, BDU, CHU)

Solutions

Computation of quartile deviation co-efficient of quartile deviation and semi-inter quartile range.

Heights in inches	f	c.f
68	25	25
69	30	55
70	42	97
71	45	142
72	43	185
73	32	217
74	30	247
75	20	267
76	16	283

$$\begin{aligned} Q_1 &= \text{Size of } \left(\frac{N+1}{4} \right) \text{th item} \\ &= \text{Size of } \left(\frac{283+1}{4} \right) \text{th item} \end{aligned}$$

$$\frac{284}{4} = 71 \text{th item} = 70$$

$$\begin{aligned} Q_3 &= \text{Size of } \left(\frac{3N+1}{4} \right) \text{th item} \\ &= \text{Size of } \left(\frac{3(283+1)}{4} \right) \text{th item} \\ &= 213 \text{th item} = 73 \end{aligned}$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{73 - 70}{2} = \frac{3}{2} = 1.5$$

$$\text{Co-efficient Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{73 - 70}{73 + 70} = \frac{3}{143} = 0.021$$

$$\text{Inter quartile range} = Q_3 - Q_1 = 73 - 70 = 3$$

$$\text{Semi-inter quartile range} = \frac{\text{Quartile range}}{2} = \frac{3}{2} = 1.5$$

Illustration 7.27

Calculate the lower and upper quartiles, 6th and 85th percentiles from the following data.

X	27	21	48	32	35	39	46
f	24	25	26	20	13	17	12

(B. Com, BU, BDU, CHU)

Solutions

Calculation of Q_1 , Q_2 , D_6 and P_{85}

X	f	c.f.
27	24	24
21	25	49
48	26	75
32	20	95
35	13	108
39	17	125
46	12	137
N = 137		

$$Q_1 = \frac{N+1}{4} \text{ th item} = \frac{137+1}{4} = \frac{138}{4} = 34.5 \text{th item}$$

$$Q_1 = 34.5 \text{th item} = 21$$

$$Q_3 = 3X\left(\frac{N+1}{4}\right) \text{th item} = 3 \times 34.5 \text{th item} = 103.5 \text{th item} = 35$$

$$D_3 = 6\left(\frac{N+1}{10}\right) \text{th item} = 6 \left(\frac{137+1}{10}\right) = 6 \times 13.8 = 82.8 \text{th item} = 32$$

$$P_{85} = 85 \left(\frac{N+1}{100} \right) \text{th item} = 85 \left(\frac{137+1}{100} \right) = 85 \times 1.38 = 117.3 \text{th item} = 39$$

Illustration 7.28

From the following data, compute co-efficient of quartile deviation.

Size	14–17	18–20	21–23	24–26	27–29
Frequency	24	34	48	30	44

Solutions

Computation of Quartile Deviation

Size	f	c.f.
14–17	24	24
18–20	34	58
21–23	48	106
24–26	30	136
27–29	14	150

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$Q_1 = \text{size of } N/4 \text{th item} = \frac{150}{4} \text{ th item} = 37.5 \text{th item}$$

Q_1 lies in the class 18 – 20 (17.5 – 20.5)

$$Q_3 = \text{size of } 3N/4 \text{th item} = 3 \times 37.5 = 112.5 \text{th item}$$

Q_3 lies in the class 24 – 26 (23.5 – 26.5)

$$\begin{aligned} Q_1 &= L \pm \frac{N/4 - cf}{f} \times C \\ &= 18 + \frac{37.5 - 24}{34} \times 2 \\ &= 18 + 0.79 = 18.79 \end{aligned}$$

$$\begin{aligned} Q_3 &= L \pm \frac{3N/4 - cf}{f} \times C \\ &= 24 + \frac{\frac{3 \times 37.5}{4} - 24}{34} \times 2 \end{aligned}$$

$$= 24 + \frac{28.13 - 24}{34} \times 2 \\ = 24 + 0.24 = 24.24$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{24.24 - 18.79}{2} = \frac{5.45}{2} = 2.725$$

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{24.24 - 18.79}{24.24 + 18.79} = \frac{5.45}{43.03} = 0.13$$

Illustration 7.29

Compute first quartile, 3rd quartile, 45th percentile and the co-efficient of quartile deviation from the following data.

Class	0–100	100–200	200–300	300–400	400–500
Frequency	14	25	38	26	17
(B. Com, MSU, MKU, CHU)					

Solutions

Calculation of Q_1 , Q_3 , P_{45} and co-efficient of quartile deviation.

X	f	cf
0–100	14	14
100–200	25	39
200–300	38	77
300–400	26	103
400–500	17	120

$$Q_1 = N/4\text{th item} = \frac{120}{4} \text{th item} = 30\text{th item} = 100 - 200$$

$$Q_3 = \frac{3N}{4} \text{th item} = 3 \times 30 = 90\text{th item} 300 - 400$$

$$Q_1 = L + \frac{N/4 - cf}{f} \times C$$

$$Q_1 = 100 + \frac{30 - 14}{25} \times 100$$

$$= 100 + \frac{16}{25} \times 100 = 100 + 0.64 \times 100 = 100 + 64 = 164$$

$$Q_3 = L + \frac{\frac{3N}{4} - cf}{f} \times C$$

$$= 300 + \frac{90 - 77}{26} \times 100 = 300 + 50 = 350$$

$$P_{45} = 45 N / 100 \text{ th item} = 45 \times \frac{120}{100} \text{ th item} = 54 \text{th item}$$

P_{45} = Lies in the class 200 – 300

$$P_{45} = L + \frac{45 N / 100 - cf}{f} \times C$$

$$= 200 + \frac{54 - 39}{38} \times 100 = 200 + 39.47 = 239.47$$

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{350 - 164}{350 + 164} = \frac{186}{514} = 0.36$$

Illustration 7.30

Find the mean deviation and the mean co-efficient of dispersion from median of the following series :

Size	15	16	17	18	19	20
Frequency	18	22	28	18	13	11

(B.Com, MSU, MKU, CHU, BU)

Solutions

Calculation of Median and M.D.

X	f	Cf	 D (X - 17)	f D
15	18	18	-2	36
16	22	40	-1	22
17	28	68	0	0
18	18	86	1	18
19	13	99	2	26
20	11	110	3	33
N = 110			$\sum f D = 135$	

$$\text{Median} = \text{Size of } \frac{N+1}{2} \text{th item} = \text{Size of } \frac{110+1}{2} \text{th item}$$

$$= 55.5 \text{th item} = 17 \text{ size}$$

$$\text{M.D.} = \frac{\sum f|D|}{N} = \frac{135}{110} = 1.23$$

$$\text{Co-efficient of M.D.} = \frac{\text{M.D.}}{\text{Median}} = \frac{1.23}{17} = 0.072$$

Illustration 7.31

The following data are related with the daily expenses of the factories in a town. Calculate arithmetic mean and mean deviation from mean, median and mode.

Expenses (Rs)	3500	4500	5500	6500	7500	8500	9500
No. of factories	6	18	23	33	16	7	3

(B.Com, MKU, BDU, CHU)

Solutions

Calculation of mean deviation

X	f	cf	X - A A = 6500	fd
3500	6	6	-3000	-18000
4500	18	24	-2000	-36000
5500	23	47	-1000	-23000
6500	33	80	0	0
7500	16	96	1000	16000
8500	7	103	2000	14000
9500	3	106	3000	9000
N = 106			$\sum fd = -38000$	

$$\begin{aligned}\text{Mean} &= A \pm \frac{\sum fd}{N} \\ &= 6500 + \left(\frac{-38000}{106} \right) = 6500 - 358.49 = 6141.51\end{aligned}$$

$$\text{Median} = \frac{N+1}{2} \text{th item} = \frac{106+1}{2} \text{th item} = \frac{107}{2} \text{th item}$$

$$= 53.5 \text{th item} = 6500$$

$$\text{Mode} = 3 \text{ median} - 2 \text{ mean}$$

$$= (3 \times 6500) - (2 \times 6141.51)$$

$$= 19500 - 12283.02 = 7216.98$$

Calculation of mean deviation

X	f	$f(X - 6141.51)$	$f D $	$X - 6500$	$f D $	$X - 7216.98$	$f D $
				$ D $		$ D $	
3500	6	2641.51	15849.06	+ 3000	+ 18000	3716.98	22301.88
4500	18	1641.51	29547.18	+ 2000	+ 36000	2716.98	48905.64
5500	23	641.51	14754.73	+ 1000	+ 23000	1716.98	39490.54
6500	33	358.49	11830.17	0	0	0	0
7500	16	1358.49	21735.84	+ 1000	+ 16000	283.02	4528.32
8500	7	2358.49	16509.43	+ 2000	+ 14000	1283.02	8981.14
9500	3	3358.49	10075.47	+ 3000	+ 9000	2283.02	6849.06
$N = 106$		$\sum f D = 120301.88$		$\sum f D = 116000$		$\sum f D = 131056.58$	

Mean deviation from mean

$$\text{M.D.} = \frac{\sum f|D|}{N} = \frac{120301.88}{106} = 1134.92$$

Mean deviation from median

$$\text{M.D.} = \frac{\sum f|D|}{N} = \frac{116000}{106} = 1094.34$$

Mean deviation from mode

$$\text{M.D.} = \frac{\sum f|D|}{N} = \frac{131056.58}{106} = 1236.38$$

Illustration 7.32

Calculate standard deviation and co-efficient of variation from the data given below.

Class integrals	0–15	15–30	30–45	45–60	60–75	75–90	90–105	105–120
Frequency	44	56	62	75	69	42	39	32

(B.Com, MKU, BDU, CHU)

Solutions

Calculation of standard deviation

X	m	f	$d = \left[\frac{m - 67.5}{15} \right]$	fd	d^2	fd^2
0–15	7.5	44	- 4	- 176	16	704
15–30	22.5	56	- 3	- 168	9	504

Contd.

X	m	f	$d = \left[\frac{m - 67.5}{15} \right]$	fd	d^2	fd^2
30–45	37.5	62	-2	-124	4	248
45–60	52.5	75	-1	-75	1	75
60–75	67.5	69	0	0	0	0
75–90	82.5	42	1	42	1	42
90–105	97.5	39	2	78	4	156
105–120	112.5	32	3	96	9	288
		$N = 419$		$\sum fd = -327$		$\sum fd^2 = 2017$

$$\begin{aligned}\text{Standard deviation } (\sigma) &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times C \\ &= \sqrt{\frac{2017}{419} - \left(\frac{-327}{419}\right)^2} \times 15 \\ &= \sqrt{4.81 - 0.61} \times 15 \\ &= \sqrt{4.2} \times 15 = 2.05 \times 15 = 30.75\end{aligned}$$

$$\text{Co-efficient of variation} = \frac{\sigma}{\bar{x}} \times 100$$

$$\begin{aligned}\therefore \bar{x} &= A \pm \frac{\sum fd}{N} \times C \\ &= 67.5 \pm \frac{(-327)}{419} \times 15 \\ &= 67.5 - 11.71 = 55.79\end{aligned}$$

when

$$\bar{x} = 55.79; \sigma = 30.75$$

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100 = \frac{30.75}{55.79} \times 100$$

$$\text{C.V.} = 55.12$$

Illustration 7.33

You are given heights of 50 students in a class. Calculate the standard deviation and also calculate the lowest and highest heights through mean ± 3 technique.

Height (in cms)	52.5	53.0	53.5	54.0	55.0	55.5	57.5	58.0	59.5
No.of students	6	9	7	23	10	8	6	4	5

(B.Com, BU, BDU, MSU)

Solutions

Calculation of standard deviation

X	f	d = (X - 55)	d²	fd	fd²
52.5	6	-2.5	6.25	-15	37.5
53.0	9	-2	4	-18	36
53.5	7	-1.5	2.25	-10.5	15.75
54.0	23	-1	1	-23	23
55.0	10	0	0	0	0
55.5	8	0.5	0.25	4	2
57.5	6	2.5	6.25	15	37.5
58.0	4	3	9	12	36
59.5	5	4.5	20.25	22.5	101.25
N = 78			Σ fd = -13 Σ fd² = 289		

$$\bar{X} = A + \frac{\sum fd}{N} = 55 + \left(\frac{-13}{78} \right) = 55 - 0.17 = 54.83$$

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \\ &= \sqrt{\frac{289}{78} - \left(\frac{-13}{78} \right)^2} = \sqrt{3.71 - 0.03} = \sqrt{3.68} = 1.92\end{aligned}$$

Calculation of lowest and highest heights

$$\bar{X} = 54.83; \sigma = 1.92$$

$$\text{Highest Height} = \bar{X} + 3 \sigma = 54.83 + (3 \times 1.92)$$

$$= 54.83 + 5.76 = 60.59$$

$$\text{Lowest Height} = \bar{X} - 3 \sigma = 54.83 - (3 \times 1.92)$$

$$= 54.83 - 5.76 = 49.07$$

Illustration 7.34

Calculate the missing information of the following data:

	Level I	Level II	Level III	Combined
Total number	?	350	300	900
Mean	325	?	336	329
Standard deviation	7.2	6.7	?	6.25

(B.Com, MKU, BDU, CHU)

Solutions

Let N_1, N_2, N_3 denote total number in the 1st, 2nd and 3rd levels respectively.

We are given $N_1 = ? ; N_2 = 350 ; N_3 = 300$

$$N_1 + N_2 + N_3 = 900$$

$$N_2 + N_3 = 350 + 300 = 650$$

$$N_1 = 900 - (N_2 + N_3) = 900 - 650 = 250$$

Finding mean of 2nd item

Let $\bar{X}_1, \bar{X}_2, \bar{X}_3$ denotes mean of 1st, 2nd and 3rd levels respectively.

$$\bar{X}_{123} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2 + N_3 \bar{X}_3}{N_1 + N_2 + N_3}$$

$$\bar{X}_{123} = 329; N_1 + N_2 + N_3 = 900 \quad \bar{X}_1 = 325; \quad \bar{X}_2 = ?; \quad \bar{X}_3 = 336$$

$$\therefore 329 = \frac{(250 \times 325) + (350 \times \bar{X}_2) + (300 \times 336)}{250 + 350 + 300}$$

$$329 = \frac{81250 + 350 \bar{X}_2 + 100800}{900}$$

$$329 \times 900 = 182050 + 350 \bar{X}_2$$

$$296100 = 182050 + 350 \bar{X}_2$$

$$296100 - 182050 = 350 \bar{X}_2$$

$$114050 = 350 \bar{X}_2$$

$$\bar{X}_2 = \frac{114050}{350} = 325.86$$

Calculation of standard deviation

$$\sigma_{123} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_3 \sigma_3^2 + N_1 d_1^2 + N_2 d_2^2 + N_3 d_3^2}{N_1 + N_2 + N_3}}$$

$$d_1 = \bar{X}_1 - \bar{X}_{123} = 325 - 329 = -4$$

$$d_2 = \bar{X}_2 - \bar{X}_{123} = 325.86 - 329 = -3.14$$

$$d_3 = \bar{X}_3 - \bar{X}_{123} = 336 - 329 = 7$$

 σ_{123}

$$= \sqrt{\frac{250 \times (7.2)^2 + 350 \times (6.7)^2 + 300 \times \sigma_3^2 + 250 \times (-4)^2 + 350 \times (-3.14)^2 + 300 \times (7)^2}{250 + 350 + 300}}$$

$$6.25 = \sqrt{\frac{250 \times 51.84 + 350 \times 44.89 + 300 \times \sigma_3^2 + 250 \times 16 + 350 \times 9.86 + 300 \times 49}{900}}$$

$$6.25 = \sqrt{\frac{12960 + 15711.5 + 300\sigma_3^2 + 4000 + 3451 + 14700}{900}}$$

$$6.25 = \sqrt{\frac{50822.5 + 300\sigma_3^2}{900}}, (6.25)^2 = \frac{50822.5 + 300\sigma_3^2}{900}$$

$$= 39.0625 \times 900 = 50822.5 + 300\sigma_3^2$$

$$= 35156.25 - 50822.5 = 300\sigma_3^2$$

$$= 15666.25 = 300\sigma_3^2$$

$$\sigma_3^2 = \sqrt{\frac{15666.25}{300}} = \sqrt{52.220833} = 7.23$$

Illustration 7.35

The following are the run-scores of two cricketers in 12 innings. Find which batsman is more consistent in scoring.

I	16	8	20	91	56	48	36	11	70	92	6	25
II	46	41	33	56	61	71	76	40	33	25	52	56

(B.Com, MKU, MSU, BDU, CHU)

Solutions

Calculation of co-efficient of variation.

X	x (X - 40)	x ²	Y	y (Y - 49)	y ²
16	-24	576	46	-3	9
8	-32	1024	41	-8	64
20	-20	400	33	-16	256
91	51	2601	56	7	49

Contd.

X	x ($X - 40$)	x^2	Y	y ($Y - 49$)	y^2
56	16	256	61	12	144
48	8	64	71	22	484
36	-4	16	76	27	729
11	-29	841	40	-9	81
70	30	900	33	-16	256
92	52	2704	25	-24	576
6	-34	1156	51	2	4
25	-15	225	56	7	49
$\sum X = 479$		$\sum x^2 = 10763$	$\sum Y = 589$	$\sum y^2 = 2701$	

Co-efficient of variation I

$$\bar{X} = \frac{\sum X}{N} = \frac{479}{12} = 39.9167 = 40$$

$$\sigma = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{10763}{12}} = \sqrt{896.9166} = 30$$

$$C.V. = \frac{\sigma}{\bar{X}} = \frac{30}{40} \times 100 = 75$$

Co-efficient of variation II

$$\bar{Y} = \frac{\sum Y}{N} = \frac{589}{12} = 49.08$$

$$\sigma = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{2701}{12}} = \sqrt{225.0833} = 15$$

$$C.V. = \frac{\sigma}{\bar{y}} \times 100 = \frac{15}{49} \times 100 = 30.61$$

Co-efficient of variation I = 75

Co-efficient of variation II = 30.61

Since the co-efficient of variation is more in case of batsman I as compared to II, batsman II is more consistent in scoring.

Illustration 7.36

The arithmetic mean of observations is 95 and the variation is 6.4. If 6 of the observations are 45, 58, 69, 72, 62 and 36. Calculate the other two observations.

Solutions

Calculation of missing observations.

$$\begin{aligned}\bar{X} &= \frac{\sum X}{N} \\ \therefore \sum X &= N\bar{X} \\ \text{Hence, } N &= 8; \bar{X} = 95 \\ \sum X &= N\bar{X} = 8 \times 95 = 760\end{aligned}$$

Let the two missing items be X_1 and X_2

$$\begin{aligned}45 + 58 + 69 + 72 + 62 + 36 + X_1 + X_2 &= 760 \\ 342 + X_1 + X_2 &= 760 \\ X_1 + X_2 &= 760 - 342 \\ X_1 + X_2 &= 418\end{aligned}$$

Variance = 6.4

$$\begin{aligned}\sigma^2 &= \frac{\sum X^2}{N} - (\bar{X})^2 \\ 6.4 &= \frac{\sum X^2}{N} - (95)^2 \\ 6.4 &= \frac{\sum X^2}{N} - 9025 \\ 8 \times 6.4 &= \sum X^2 - 9025 \times 8 \\ 51.2 &= \sum X^2 - 72200\end{aligned}$$

$$\sum X^2 = 72200 + 51.2 = 72251.2$$

$$\sum X^2 = X_1^2 + X_2^2 + (45)^2 + (58)^2 + (69)^2 + (72)^2 + (62)^2 + (36)^2$$

$$\sum X^2 = X_1^2 + X_2^2 + 2025 + 3364 + 4761 + 5184 + 3844 + 1296$$

$$\sum X^2 = X_1^2 + x_2^2 + 20474$$

where, $\sum X^2 = 72251.2$ then

$$72251.2 = X_1^2 + X_2^2 + 20474$$

$$72251.2 - 20474 = X_1^2 + X_2^2$$

$$X_1^2 + X_2^2 = 51777.2$$

$$(X_1 + X_2)^2 = X_1^2 + X_2^2 + 2X_1 X_2$$

$$(418)^2 = 51777.2 + 2X_1 X_2$$

$$174724 = 51777.2 + 2X_1 X_2$$

$$2X_1 X_2 = 174724 - 51777.2$$

$$2X_1 X_2 = 122946.8$$

$$X_1 X_2 = \frac{122946.8}{2} = 61473.4$$

$$\begin{aligned}(X_1 - X_2)^2 &= X_1^2 + X_2^2 - 2X_1 X_2 = 51777.2 - 2 \times 61473.4 \\ &= 51777.2 - 122946.8 = -71169.6\end{aligned}$$

$$X_1 - X_2 = -267 \quad (1)$$

$$\begin{array}{r} X_1 + X_2 = 418 \\ \hline 2X_1 = 151 \end{array} \quad (2)$$

$$X_1 = \frac{151}{2} = 75.5$$

$$X_1 + X_2 = 418; X_1 = 75.5$$

$$75.5 + X_2 = 418$$

$$X_2 = 418 - 75.5 = 342.5$$

$$X_1 = 75.5 \text{ and } X_2 = 342.5$$

Illustration 7.37

The mean and standard deviation of 100 items are found to be 50 and 10 respectively. If at the time of calculation, two items were wrongly taken as 2 and 65 instead of 11 and 15, find the correct mean and standard deviation. What is the correct co-efficient of variation?

(B.Com, MKU, MSU, BDU, CHU)

Solutions

Correct mean

We are given $\bar{X} = 50$, $\sigma = 10$, $N = 100$

$$\bar{X} = \frac{\sum X}{N}$$

$$50 = \frac{\sum X}{100}$$

$$\sum X = 5000$$

But correct $\sum X = \sum X - \text{wrong items and correct items}$

$$= 5000 - 2 - 65 + 11 + 15 = 4959$$

$$\therefore \text{Correct mean } \bar{X} = \frac{\text{Correct } \sum X}{N} = \frac{4959}{100} = 49.59$$

Correct standard deviation

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

$$10 = \frac{\sum X^2}{100} - (50)^2$$

$$(10)^2 = \frac{\sum X^2}{100} - (50)^2$$

$$100 = \frac{\sum X^2}{100} - 2500$$

$$100 \times 100 = \sum X^2 - 2500 \times 100$$

$$10000 = \sum X^2 - 250000$$

$$10000 + 250000 = \sum X^2$$

$$\sum X^2 = 10000 + 250000$$

$$\sum X^2 = 260000$$

But Correct $\sum X^2$ = Incorrect $\sum X^2$ – Wrong items square + Correct items square.

$$\begin{aligned}\therefore \text{Correct } \sum X^2 &= 260000 - (2)^2 - (65)^2 + (11)^2 + (15)^2 \\ &= 260000 - 4 - 4225 + 121 + 225 \\ &= 256117\end{aligned}$$

$$\begin{aligned}\text{Again correct } \sigma &= \sqrt{\frac{\text{Correct } \sum X^2}{N} - (\text{Correct } \bar{X})^2} \\ &= \sqrt{\frac{256117}{100} - (49.59)^2} = \sqrt{2561.17 - 2459.17} \\ &= \sqrt{102} = 10.10\end{aligned}$$

$$\text{Co-efficient of variance} = \frac{\sigma}{\bar{X}} \times 100 = \frac{10.10}{49.59} \times 100 = 20.37\%$$

Illustration 7.38

Co-efficient of variations of two series are 80% and 100% respectively. Their standard deviations are 40 and 36 respectively. What are their arithmetic means?

Solutions

Let \bar{X}_1 and \bar{X}_2 be the means of first and second series. Let corresponding standard deviation be σ_1 and σ_2 . Thus,

$$\text{First series} \quad C.V. = \frac{\sigma_1}{\bar{X}_1} \times 100$$

$$\text{Co-efficient of variance} = 80 = \frac{\sigma_1}{\bar{X}_1} \times 100$$

$$80 = \frac{40}{\bar{X}_1} \times 100$$

$$80 \bar{X}_1 = 40 \times 100$$

$$80 \bar{X}_1 = 4000$$

$$\bar{X}_1 = \frac{4000}{80}$$

$$\bar{X}_1 = 50$$

$$\text{Second series} \quad C.V. = 100 = \frac{36}{\bar{X}_2} \times 100$$

$$100 \bar{X}_2 = 36 \times 100$$

$$100 \bar{X}_2 = 3600$$

$$\bar{X}_2 = \frac{3600}{100}$$

$$\bar{X}_2 = 36$$

Illustration 7.39

The mean and standard deviation of 73 children on an average test are respectively 37.6 and 17.1. To them are added a new group of 36 who have had less training and whose mean is 29.2 and the standard deviation 16.2. How will the value of combined group differ from those of the original 73 children as to (a) mean and (b) the standard deviation?

(B.Com, MKU, BDU, BU, CHU)

Solutions

Group I : $N_1 = 73$; $\bar{X}_1 = 37.6$; $\sigma_1 = 17.1$

Group II : $N_2 = 36$; $\bar{X}_2 = 29.2$; $\sigma_2 = 16.2$

Combined mean :

$$\begin{aligned}\bar{X}_{12} &= \frac{\bar{X}_1 N_1 + \bar{X}_2 N_2}{N_1 + N_2} \\ &= \frac{37.6 \times 73 + 29.2 \times 36}{73 + 36} \\ &= \frac{2744.8 + 1051.2}{109} = \frac{3796}{109} = 34.83\end{aligned}$$

Combined standard deviation:

$$\begin{aligned}\sigma_{12} &= \sqrt{\frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N_1 + N_2}} \\ &= \sqrt{\frac{73(17.1^2 + (37.6 - 34.83)^2) + 36(16.2^2 + (29.2 - 34.83)^2)}{73 + 36}} \\ &= \sqrt{\frac{73(292.41 + 7.67) + 36(262.44 + 31.70)}{109}} \\ &= \sqrt{\frac{(73 \times 300.08) + (36 \times 294.14)}{109}} \\ &= \sqrt{\frac{21905.84 + 10589.04}{109}} = \sqrt{\frac{32494.88}{109}} = \sqrt{298.12} = 17.27\end{aligned}$$

From the above calculation, it can be concluded that (a) the mean of the original 73 children decreased by 2.77, that is, 37.6–34.83 (b) the standard deviation of 73 children increased by 0.17, that is, 17.27–17.1.

Illustration 7.40

A sample of size 25 has mean 2.5 and S.D 2.0; another sample of size 20 has mean 2.7 and S.D 4.0. If the two samples are plotted together, find the mean and the standard deviation of the combined mean.

(B.Com, MKU, BDU, CHU)

Solutions

Combined mean:

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$N_1 = 25 ; \bar{X}_1 = 2.5 ; \sigma_1 = 2$$

$$N_2 = 20 ; \bar{X}_2 = 2.7 ; \sigma_2 = 4$$

$$\bar{X}_{12} = \frac{(25 \times 2.5) + (20 \times 2.7)}{25 + 20} = \frac{62.5 + 54}{45} = \frac{116.5}{45} = 2.59$$

Combined standard deviation :

$$\begin{aligned}\sigma_{12} &= \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 (\bar{X}_1 - \bar{X}_{12})^2 + N_2 (\bar{X}_2 - \bar{X}_{12})^2}{N_1 + N_2}} \\ &= \sqrt{\frac{25 \times (2)^2 + 20(4)^2 + 25(2.5 - 2.59)^2 + 20(2.7 - 2.59)^2}{25 + 20}} \\ &= \sqrt{\frac{25 \times 4 + 20 \times 16 + 25 \times 0.008 + 20 \times 0.012}{45}} \\ &= \sqrt{\frac{100 + 320 + 0.2 + 0.24}{45}} \\ &= \sqrt{\frac{420.44}{45}} = \sqrt{9.34} = 3.06\end{aligned}$$

Illustration 7.41

The numbers examined the mean weight and the S.D. in each group of examination by two medical science examiners are given below. Find the mean weight and S.D. of both the groups taken together.

Medical Science Examiner	No. of Examined	Mean Weight	S.D.
P	70	133	26.5
Q	80	140	28.2

Solutions

Combined mean

$$\bar{X}_{12} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2}$$

$$N_1 = 70 ; \bar{X}_1 = 133 ; N_2 = 80 ; \bar{X}_2 = 140$$

$$\bar{X}_{12} = \frac{70 \times 133 + 80 \times 140}{70 + 80} = \frac{9310 + 11200}{150} = \frac{20510}{150}$$

$$\bar{X}_{12} = 136.73$$

$$\text{Combined S.D } \sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$N_1 = 70; \sigma_1 = 26.5; N_2 = 80; \sigma_2 = 28.2; d_1 = \bar{X}_1 - \bar{X}_{12}$$

$$d_1 = 133 - 136.73 = 3.73$$

$$d_2 = \bar{X}_2 - \bar{X}_{12} = 140 - 136.73 = 3.27$$

$$\sigma_{12} = \sqrt{\frac{70 \times (26.5)^2 + 80 \times (28.2)^2 + 70 \times (3.73)^2 + 80 \times (3.27)^2}{70 + 80}}$$

$$= \sqrt{\frac{(70 \times 702.25) + (80 \times 795.24) + (70 \times 13.91) + (80 \times 10.69)}{150}}$$

$$= \sqrt{\frac{49157.5 + 63619.2 + 973.90 + 855.43}{150}}$$

$$= \sqrt{\frac{114606.03}{150}} = \sqrt{764.040} = 27.64$$

Illustration 7.42

Find the standard deviation and co-efficient of variation from the following data.

Wages	No.of workers
Upto Rs 100	32
Upto Rs 200	50
Upto Rs 300	85
Upto Rs 400	127
Upto Rs 500	177
Upto Rs 600	222
Upto Rs 700	242
Upto Rs 800	250

Solutions

Calculation of co-efficient of variation

Wages Rs.	M.P <i>m</i>	<i>Cf</i>	No. of workers <i>cf</i>	$\frac{m - 350}{100}$ <i>fd</i>	<i>fd</i>	<i>d</i> ²	<i>fd</i> ²
0–100	50	32	32	–3	–96	9	288
100–200	150	50	18	–2	–36	4	72
200–300	250	85	35	–1	–35	1	35
300–400	350	127	42	0	0	0	0
400–500	450	177	50	1	50	1	50
500–600	550	222	45	2	90	4	180
600–700	650	242	20	3	60	9	180
700–800	750	250	8	4	32	16	128
$N = 250$				$\sum fd = 65$		$\sum fd^2 = 933$	

$$\begin{aligned}
 \bar{x} &= A + \frac{\sum fd}{N} \times i \\
 &= 350 + \frac{65}{250} \times 100 = 376 \\
 \sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C \\
 &= \sqrt{\frac{933}{250} - \left(\frac{65}{250} \right)^2} \times 100 \\
 &= \sqrt{3.732 - (0.26)^2} \times 100 \\
 &= \sqrt{3.732 - 0.068} \times 100 = \sqrt{3.66} \times 100 \\
 \sigma &= 1.91 \times 100 = 191 \\
 C.V. &= \frac{\sigma}{\bar{x}} \times 100 \\
 \sigma &= \frac{191}{376} \times 100 = 0.51 \times 100 = 51\%
 \end{aligned}$$

Illustration 7.43

Calculate co-efficient of quartile deviation and co-efficient of variation from the following data :

Marks	Below 10	Below 20	Below 30	Below 40	Below 50
No. of students	4	18	48	68	78

(B.Com, MKU, BDU, CHU)

Solutions

Calculation of co-efficient of quartile deviation and co-efficient of variation.

Marks	M.P	f	No. of students	$d = \left(\frac{m - 25}{10} \right)$	d^2	fd	fd^2
	<i>m</i>		<i>c.f</i>				
0–10	5	4	4	−2	4	−8	16
10–20	15	14	18	−1	1	−14	14
20–30	25	30	48	0	0	0	0
30–40	35	20	68	+1	1	20	20
40–50	45	10	78	+2	4	20	40
$N = 78$				$\sum fd = 18 \quad \sum fd^2 = 90$			

Mean :

$$\begin{aligned}\bar{X} &= A + \frac{\sum fd}{N} \times C \\ &= 25 + \frac{18}{78} \times 10 \\ &= 25 + 0.23 \times 10 \\ &= 25 + 2.3 = 27.3\end{aligned}$$

Standard Deviation :

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C \\ &= \sqrt{\frac{90}{78} - \left(\frac{18}{78} \right)^2} \times 10 \\ &= \sqrt{1.15 - (0.23)^2} \times 10 \\ &= \sqrt{1.15 - 0.05} \times 10 \\ &= \sqrt{1.10} \times 10 = 1.05 \times 10 = 10.5\end{aligned}$$

$$C.V. = \frac{\sigma}{\bar{x}} \times 100 = \frac{10.5}{27.3} \times 100 = 38.46\%$$

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$Q_1 = \text{Size of } \frac{N}{4} \text{ th item}$$

Q_1 lies in the class 20 – 30

$$\begin{aligned} Q_1 &= L + \frac{\frac{N}{4} - cf}{f} \times i \\ &= 20 + \frac{19.5 - 18}{30} \times 10 \\ &= 20 + \frac{1.5}{30} \times 10 \end{aligned}$$

$$Q_1 = 20 + 0.5 = 20.5$$

$$\begin{aligned} Q_3 &= \text{Size of } \frac{3N}{4} \text{ th item} \\ &= \text{Size of } \frac{3 \times 78}{4} = \frac{234}{4} = 58.5 \text{th item} \end{aligned}$$

Q_3 lies in the class 30 – 50

$$\begin{aligned} Q_3 &= L_l + \frac{\frac{3N}{4} - cf}{f} \times i \\ &= 30 + \frac{58.5 - 48}{20} \times 10 \\ &= 30 + \frac{10.5}{20} \times 10 \\ &= 30 + 5.25 = 35.25 \end{aligned}$$

$$\begin{aligned} \text{Co-efficient of Q.D.} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\ &= \frac{35.25 - 20.5}{35.25 + 20.5} = \frac{14.75}{55.75} = 0.27 \end{aligned}$$

Illustration 7.44

The arithmetic mean and standard deviation of a series of 40 items were calculated by a student as 40 cm and 25 cm respectively. But while calculating them an item, 33 was misread 50. Find the correct mean and standard deviation.

Solutions

$$\begin{aligned} \text{Arithmetic mean} &= \frac{\sum X}{N} \\ 40 &= \frac{\sum X}{40} \end{aligned}$$

$$\sum X = 40 \times 40 = 1600$$

$$\text{Correct } \sum X = 1600 + 33 - 50 = 1583$$

$$\text{Correct mean} = \frac{1583}{80} = 39.58 \text{ cm}$$

$$\text{S.D.} = \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2}$$

$$5 = \sqrt{\frac{\sum X^2}{40} - \left(\frac{1600}{40}\right)^2}$$

$$5 = \sqrt{\frac{\sum X^2}{40} - (40)^2}$$

$$5^2 = \frac{\sum X^2}{40} - (40)^2$$

$$\frac{\sum X^2}{40} = 5^2 + 1600$$

$$\sum X^2 = 1625 \times 40 = 65000$$

$$\begin{aligned} \text{Correct } \sum X^2 &= 65000 - (50)^2 + (33)^2 \\ &= 65000 - 2500 + 1089 = 63589 \end{aligned}$$

$$\begin{aligned} \text{Correct S.D.} &= \sqrt{\frac{\sum X^2}{N} - \left(\frac{\sum X}{N}\right)^2} \\ &= \sqrt{\frac{63589}{40} - \left(\frac{1583}{40}\right)^2} \\ &= \sqrt{1589.73 - 1566.58} \\ &= \sqrt{23.15} = 4.81 \end{aligned}$$

Illustration 7.45

A purchase of 50 new glass marble, 25 of Brand A and 50 of Brand B at the same price under substantially similar deviation of glass marble, he gets the following results.

	Glass Marble A	Glass Marble B
Arithmetic	150	200
Standard Deviation	30	80

Which brand you would advise him to buy in future and why?

(B.Com, MKU, BDU, MSU, CHU)

Solutions

Glass Marble *A*

$$\begin{aligned}\text{Co-efficient of variation} &= \frac{\text{S.D.}}{\text{A. Mean}} \times 100 \\ &= \frac{30}{150} \times 100 = 20\%\end{aligned}$$

Glass Marble *B*

$$\begin{aligned}\text{Co-efficient of variation} &= \frac{\text{S.D.}}{\text{A. Mean}} \times 100 \\ &= \frac{80}{200} \times 100 = 40\%\end{aligned}$$

Since, co-efficient of variation for glass marble *B* is more, it is more variable. Hence, it is advisable to buy glass marble *A*.

Illustration 7.46

Mean of 300 items is 150 and the S.D. is 14 (V). Find the sum of square of all items.

Solutions

$$\begin{aligned}\text{A.M.} &= \frac{\sum X}{N} \\ 150 &= \frac{\sum X}{300} \\ \sum X &= 300 \times 150 = 45000 \\ V &= \frac{\sum X^2}{N} - \left(\frac{\sum X}{N} \right)^2 \\ 14 &= \sqrt{\frac{\sum X^2}{300} - \left(\frac{45000}{300} \right)^2} \\ 14 &= \frac{\sum X^2}{300} - 22500\end{aligned}$$

$$\frac{\sum X^2}{300} = 22500 + 14$$

$$\frac{\sum X^2}{300} = 22500 + 196$$

$$\sum X^2 = 22696 \times 300$$

$$\sum X^2 = 6808800$$

SUMMARY

Dispersion

It is the measure of the variations of the item from central value.

Objectives of the Measure of Dispersion

- To find out the reliability of an Average
- To control the variation of data from the central value
- To compare two or more set of data regarding their variability.

Methods of Measuring Dispersion

- Range
- Quartile Deviation
- Mean Deviation
- Standard Deviation
- Lorenz Curve

Range

Difference between the largest value and the smallest value of the variables.

Quartile Deviation

Difference between the third quartile and the first quartile divided by two.

Mean Deviation

It is the measure of deviation based on all the items in a distribution.

Standard Deviation

It is the square root of the sum of square deviations from average.

Mathematical Properties of Standard Deviation

- Combined standard deviation
- Co-efficient of variation
- Variance

Combined Standard Deviation (σ_{12})

It can be calculated by following the same method of calculating the combined mean of two or more than two groups.

Co-efficient of Variation (c.r)

The relative measure of standard deviation.

Variance (σ^2)

It is the square of standard deviation.

Lorenz Curve

Graphic method of measuring dispersion.

FORMULAE**Range (R)****Individual series, Discrete and Continuous Series**

$$R = L - S$$

R = Range

L = Largest Value

S = Smallest Value

Coefficient of Range (C.R)

Individual, Discrete and Continuous Series

$$C.R = \frac{L - S}{L + S}$$

Quartile Deviation (Q.D)

Individual, Discrete and Continuous Series

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

Q.D. = Quartile Deviation

Q_1 = Lower quartile

Q_3 = Upper quartile

Coefficient of Quartile Deviation (Co-efficient of Q.D.)

Individual, Discrete and Continuous Series

$$\text{Co-efficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation Individual Series

$$M.D. = \left\{ \begin{array}{l} \frac{\sum (x - \bar{x})}{N} \\ \frac{\sum (x - \text{Med})}{N} \\ \frac{\sum (x - z)}{N} \end{array} \right\} \frac{\sum IDI}{N}$$

M.D. = Mean Deviation

$\sum |D|$ = Total of Deviations from mean, median or mode ignoring signs.

N = Number of items

Discrete and continuous series

$$M.D. = \frac{\sum f |D|}{N}$$

When mean or median is in fraction M.D.

$$M.D. = \frac{\sum mfa - \sum fmb - (\sum fa - \sum fb) \times \bar{x}}{N}$$

$\sum f|D|$ = sum of the products of frequency and respective deviation from \bar{x} or med. Or $Z \sum fma$ and $\sum fmb$ = Totals of products of mid points and frequencies corresponding to mid points above and below the average value (or medial value) respectively.

$$M.D. = \frac{\sum mfa - \sum fmb - (\sum fa - \sum fb)}{N} \times \text{Med}$$

$\sum fma$ and $\sum fmb$ = The totals of frequencies corresponding to mid points above and below the mean value median value.

N = Total number of items

Co-efficient of mean deviation individual, discrete and continuous series.

$$\text{Co-efficient of M.D.} = \frac{\text{Mean Deviation}}{\bar{x} \text{ or med or } Z}$$

Standard Deviation (σ) individual series

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

$\sum x^2$ = Total square deviations from mean

N = Number of items

Assumed Mean

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N} \right)^2}$$

$d = X - \text{Assumed Mean}$

Step Deviation Method:

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N} \right)^2} \times C$$

Discrete and Continuous Series:

Actual mean method

$$\begin{aligned} \sigma &= \sqrt{\frac{\sum f(x - \bar{x})^2}{N}} \\ &= \sqrt{\frac{\sum fd^2}{N}} \\ d &= x - \bar{x} \end{aligned}$$

Assumed mean method

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2}$$

Step deviation method

$$\sigma = \sqrt{\frac{\sum fd^{1^2}}{N} - \left(\frac{\sum fd^1}{N} \right)^2} \times C$$

Co-efficient of variation (C.V.)

$$C.V. = \frac{\sigma}{\bar{x}} \times 100$$

C.V. = Co-efficient of variation

 σ = Standard Deviation \bar{X} = MeanCombined standard deviation (σ_{12})

$$\sigma_{12} = \sqrt{\frac{N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$$d_1 = \bar{x}_1 - \bar{x}_{12}$$

$$d_2 = \bar{x}_2 - \bar{x}_{12}$$

EXERCISES**(a) Choose the best option:**

1. The measure of the degree of scatter of data from the central value is
 - (a) Dispersion
 - (b) Skewness
 - (c) Average
2. The difference between the largest value and the smallest value of the variable is
 - (a) Range
 - (b) Mean
 - (c) Quartile Deviation
3. Quartile Deviation is otherwise called as
 - (a) Quartile Range
 - (b) Inter Quartile Range
 - (c) Intra Quartile Range

4. Mean Deviation is otherwise called as
 - (a) Average Deviation (b) Arithmetic Mean (c) Dispersion
5. In Standard Deviation method, deviations should be taken only from
 - (a) Harmonic Mean
 - (b) Arithmetic Mean
 - (c) Geometric Mean
6. The relative measure of standard deviation is called
 - (a) Variance
 - (b) Arithmetic Mean
 - (c) Co-efficient of variation
7. Variance is the square of
 - (a) Range
 - (b) Quartile deviation
 - (c) Standard deviation
8. The graphical method of measuring dispersion is
 - (a) Lorenz curve (b) Range (c) Ogive
9. Inter-quartile range is
 - (a) $Q_3 - Q_1$ (b) $Q_1 - Q_3$ (c) $Q_3 - Q_2$
10. Algebraic sum of deviations from mean is
 - (a) Positive (b) Negative (c) Zero
11. The range of a given distribution is
 - (a) Greater than Standard deviation
 - (b) Less than Standard deviation
 - (c) Equal to Standard deviation
12. Sum of squares of deviations is minimum when taken from
 - (a) Mean (b) Median (c) Mode

Answers

- | | | | | | |
|------|------|------|-------|-------|-------|
| 1. a | 2. a | 3. b | 4. a | 5. b | 6. c |
| 7. c | 8. a | 9. a | 10. c | 11. a | 12. a |

b. Theoretical Questions

1. What do you mean by dispersion?
2. What is Quartile deviation?
3. What is Range?
4. Define mean deviation. How does it differ from standard deviation?
5. What are the properties of a good measure of dispersion?
6. Define standard deviation. Why is it commonly used as a measure of dispersion?

7. Discuss briefly the absolute and relative measure of dispersion.
 8. Discuss various measure of dispersion.
 9. What is co-efficient of variation? What purpose does it serve?
 10. Distinguish between variance and co-efficient of variation.
 11. What is a Lorenz-curve? How it is useful in measuring income inequalities between two regions? Give the uses of standard deviation.
 12. What do you mean by mean deviation? Discuss its relative points over range and quartile deviation as a measure of dispersion.
 13. Compare mean deviation and standard deviation as measure of variation. Which of the two is a better measure? Why ?
 14. What are the characteristics of standard deviation?
 15. What is Lorenz-curve? How do you construct it? What is its use?

c. Practical Problems

- 16.** From the monthly income of 10 families given below, calculate:
(a) the median (b) the geometric mean (c) the co-efficient of range

S. No.:	1	2	3	4	5	6	7	8	9	10
Income in Rs.:	145	367	268	73	185	619	280	115	870	315

17. Find the value of third quartile if the values of first quartile and quartile deviation are:

Wages:	0–10	10–20	20–30	30–40	40–50
No. of workers:	22	38	46	35	20

Answer mode = 24.21; median = 24.46; $Q_1 = 14.803$; $Q_3 = 24.21$; co-eff. of Q.D. = 0.396 **(B.Com, MSU, MKU, BDU)**

- 18.** Compute the co-efficient of quartile deviation of the following data:

Size:	4–8	8–12	12–16	16–20	20–24	24–28	28–32	32–36	36–40
Frequency:	6	10	18	30	15	12	10	6	2

Answer $Q_1 = 14.5$; $Q_3 = 24.92$; co-eff. of Q.D. = 0.2643

19. Calculate the quartile deviation for the following frequency distribution.

x:	60	62	64	66	68	70	72
Frequency:	12	16	18	20	15	13	9

Answer Q.D. = 31

20. Find (i) inter-quartile range (ii) semi-inter-quartile range and (iii) co-efficient of quartile deviation from the following frequency distribution.

26. Calculate the mean deviation from the mean for the following data.

Marks:	0–10	10–20	20–30	30–40	40–50	50–60	60–70
No. of Students	6	5	8	15	7	6	3

Answer Mean = 33.4; M.D. about mean = 13.184;

27. Calculate the mean deviation about the mean for the following data.

x:	5	15	25	35	45	55	65
f:	8	12	10	8	3	2	7

Also find the M.D. about median and comment on the results obtained in (a) and (b)

Answer Mean = 29; M.D. about mean = 16;

Median = 22; M.D. about median = 15.8.

(B.Com, MSU, MKU, BDU)

28. The following distribution gives the difference in age between husband and wife in a particular community.

Difference:	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40
in years								

Frequency:	449	705	507	281	109	52	16	4
-------------------	-----	-----	-----	-----	-----	----	----	---

Calculate mean deviation about median from these data. What light does it throw on the social conditions of a community?

Answer M. D. about median = 5.24.

29. Find the median and mean deviation of the following data.

Size:	0–10	10–20	20–30	30–40	40–50	50–60	60–70
Frequency:	7	12	18	25	16	14	8

Answer Median = 35.2; M.D. = 13.148.

(B.Com, MSU, MKU, BDU)

30. Compute the mean deviation from the median and from mean for the following distribution of the scores of 50 college student.

Score:	140–150	150–160	160–170	170–180	180–190	190–200
---------------	---------	---------	---------	---------	---------	---------

Frequency:	4	6	10	10	9	3
-------------------	---	---	----	----	---	---

Answer 10.24; 10.56.

(B.Com, MSU, MKU, CHU)

31. Find the mean deviation around the median.

Size	Frequency
1 and upto 10	1
1 and upto 20	3
1 and upto 30	6

1 and upto 40	8
1 and upto 50	10

Answer 10.334

32. From the following information, find the standard deviation of x and y variables.

$$\sum x = 235; \sum y = 250; \sum x^2 = 6750; \sum y^2 = 6840; N = 10$$

Answer $\sigma_x = 11.08$; $\sigma_y = 7.68$

33. The following table relates to the profits and losses of 200 firms.

Profits (Rs.)	No. of Firms
5000 to 6000	16
4000 to 5000	24
3000 to 4000	60
2000 to 3000	20
1000 to 2000	10
0 to 1000	10
-1000 to 0	12
-2000 to -1000	13
-3000 to 2000	18
-4000 to 3000	14
Total	200

Calculate the standard deviation of profits.

- Answer S.D. = Rs. 2791.49. (B.Com, MSU, MKU, BU, BDU)**
34. In the following data, two class frequencies are missing.

Class Interval	Frequency
100 – 110	4
110 – 120	7
120 – 130	15
130 – 140	–
140 – 150	40
150 – 160	–
160 – 170	16
170 – 180	10
180 – 190	6
190 – 200	3

However, it was possible to ascertain that the total number of frequencies was 150 and that the median has been correctly found out as 146.25. You are required to find with the help of information given:

- (i) The two missing frequencies
- (ii) Having found the missing frequencies, calculate arithmetic mean and standard deviation
- (iii) Without using the direct formula, find the value of mode.

Answer (i) 24, 25 (ii) A.M. 147.33, S.D. 19.2; (iii) Mode = 144.09.

- 35.** Calculate the arithmetic mean and the standard deviation of the values of the world's annual gold output (in millions of pounds) for 20 different years:

94	95	96	93	87	79	73	69	68	67
78	82	83	89	95	103	108	117	130	97

Also calculate the percentage of cases lying outside the mean at distances $\pm \sigma$, $\pm 2\sigma$, $\pm 3\sigma$, where σ denotes the standard deviation.

Hint: $\mu \pm \sigma = 106.14$; 74.16. The number of observation lying outside these limits is 7 out of 20. Hence, required percentage $= 7/20 \times 100 = 35$.

Answer Mean = 90.15; S.D. = 15.99 approx.
Percentage lying outside $\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$ is given by $7/20 \times 100$, $1/20 \times 100$, $0/20 \times 100$, i.e., 35%, 5% and 0% respectively.

- 36.** The number of employees, wages per employee and the variance of the wages per employee for two factories are given below:

	Factory A	Factory B
No. of Employees	50	100
Average wages per employee per month (in Rs.)	120	85
Variance of the wages per employee per month (in Rs.)	9	16

In which factory is there greater variation in the distribution of wages per employee?

Answer In factory B = C.V. (A) = 2.5, C.V. (B) = 4.7

(B.Com, MSU, MKU, BDU, CHU)

- 37.** Two workers on the same job show the following results over a long period of time:

	Worker A	Worker B
Mean time of completing the job (minutes)	30	25
Standard deviation (minutes)	6	4
(i) Which worker appears to be more consistent in the time he requires to complete the job?		

- (ii) Which worker appears to be faster in completing the job?
Explain.

Answer (i) B , (ii) C : $\bar{x}_B < \bar{x}_A$ (B.Com, MKU, BDU, CHU)

38. The following table shows that monthly expenditures of 80 students of a university on morning breakfast:

Expenditure	No. of students
78–82	2
73–77	6
68–72	7
63–67	12
58–62	18
53–57	13
48–52	9
43–47	7
38–42	4
33–37	2

Calculate arithmetic mean, standard deviation and co-efficient of variation of the above data.

Answer $\bar{x} = \text{Rs. } 58.38$; $\sigma = \text{Rs. } 10.36$; C.V. = 17.75

39. The shareholders research centre of India has conducted recently a research study on price behaviour of three leading industrial shares A , B and C for the period 1979 to 1985, the results of which are published as follows in its quarterly journal.

Share	Average Price	Standard Deviation	Current Selling Price
A	18.2	5.4	36.00
B	22.5	4.5	34.75
C	24.0	6.0	39.00

The above figures are given in Rs.

- (a) Which share in your opinion appears to be more stable in value?
(b) If you are the holder of all the type shares, which one would you like to dispose of at present and why?

Answer (a) C.V. (A) = 30; C.V. (B) = 20; C.V. (C) = 25 share B is more stable in value. (b) share A , because it has maximum variation.

40. The runs scored by two batsmen A and B in 10 innings are as follows:

By A:	10	115	5	73	7	120	36	84	29	19
By B:	45	12	76	42	4	50	37	48	13	0

Who is better run-getter? Who is more consistent?

Answer C.V. (A) = 75.92; C.V. (B) = 64.17; B is more consistent.

- 41.** During 10 weeks of a session, the marks obtained by two candidates, Ramesh and Suresh taking the computer programme course are given below:

Ramesh:	58	59	60	54	65	66	52	75	69	52
----------------	----	----	----	----	----	----	----	----	----	----

Suresh:	87	89	78	71	73	84	65	66	56	46
----------------	----	----	----	----	----	----	----	----	----	----

- (i) Who has the better scores—Ramesh or Suresh?
- (ii) Who is more consistent?

Answer Ramesh : $\bar{x}_1 = 7.25$; C.V. = 11.89;

Suresh : $\bar{x}_2 = 71.5$; $\sigma_2 = 13.08$; C.V. = 18.29

- (i) Suresh is better scorer ($\bar{x}_2 = 71.5$; $\sigma_2 = 13.08$; C.V. = 18.29)
- (ii) Ramesh is more consistent.

- 42.** Samples of polythene bags from two manufacturers, *A* and *B* are tested by a prospective buyer for bursting pressure and the results are as given in the table.

Which set of bags has more uniform pressure? If prices are the same, which manufacturers bags would be preferred by the buyer? Why?

Bursting Pressure (lb)	No. of Bags	
	A	B
5.0 – 9.9	2	9
10.0 – 14.9	9	11
15.0 – 19.9	29	18
20.0 – 24.9	54	32
25.0 – 29.9	11	27
30.0 – 34.9	5	13

Answer $\bar{x}_a = 21$; $\sigma_a = 4.878$; C.V. (*A*) = 23.23; Suresh: $\bar{x}_b = 21.81$; $\sigma_b = 70.75$; C.V. (*B*) = 32.44; Set *A*.

- 43.** Two brands of tyres are tested with the following results:

Life (in, '000 miles)	No. of tyres of Brand	
	x	y
20 – 25	1	0
25 – 30	22	24
30 – 35	64	76
35 – 40	10	0
40 – 45	3	0

- (a) Which brand of tyres have greater average life?
- (b) Compare the variability and state which brand of tyres would you use on your feet of trucks.

Answer $\bar{x} = 32.1$ thousand miles

$\sigma_x = 3.137$ thousand miles. C.V. (x) = 9.77

$\bar{y} = 31.3$ thousand miles

$\sigma_y = 0.912$ thousand miles. C.V. (y) = 2.914

(a) brand x ; (b) Brand x tyres are more variable; Brand y .

44. The mean and standard deviation of the marks obtained by two groups of students consisting of 50 each are given below. Calculate the mean and standard deviation of the marks obtained by all the 100 students:

Group	Mean	Standard deviation
1	60	8
2	55	7

Answer $\bar{x}_{12} = 57.5$; $\sigma_{12} = 7.92$ (B.Com, MKU, BDU, BU)

45. Calculate the standard deviation of the combined group of 500 items from the following data:

	Group I	Group II	Group III
No. of items	100	150	250
Arithmetic mean	50	55	60
Variance	100	121	144

Answer $\sigma_{123} = 11.98$

46. For two firms A and B , the following details are available:

	A	B
Number of employees	100	200
Average salary (Rs.)	1600	1800
Standard deviation of salary (Rs.)	16	18

- (i) Which firm pays large package of salary?
- (ii) Which firm shows greater variability in the distribution of salary?
- (iii) Compute the combined average salary and combined variance of both the firms.

Answer (i) B ; (ii) C.V. (A) = 1; C.V. (B) = 1. Both firms show equal variability.

(iii) $\bar{x}_{12} = \text{Rs. } 1,733.33$ and $\sigma_{12}^2 = \text{Rs. } 9190.22$

(B.Com, CHU, MSU, MKU, BDU)

47. If the mean deviation of a moderately skewed distribution is 7.2 unit, find the standard deviation as well as quartile deviation.

Answer S.D. = 5/4 M.D. = 9; Q.D. = 5/6 M.D. = 6.0

- 48.** For a series, the value of mean deviation is 15. Find the most likely value of its quartile deviation.

Answer Q.D. = $5/6$ M.D. = 12.5

- 49.** From the following table giving data regarding income of employees in two factories. Draw a graph (Lorenz curve) to show which factory has greater inequalities of income.

income ('00 Rs.):	Below 200	200–500	500–1000	1000–2000	2000–3000
Factory A:	7000	1000	1200	800	500
Factory B:	800	1200	1500	400	200

- 50.** The frequency distribution of marks obtained in Mathematics (M) and English (E) are as follows:

Mid-value of Marks:	5	15	25	35	45	55	65	75	85	95
No. of Students (m):	10	12	13	14	22	27	20	12	11	9
No. of Students (f):	1	2	26	50	59	40	10	8	3	1

Analyse the data by drawing the Lorenz curves on the same diagram and describe the main features you observe.

8

CHAPTER

SKEWNESS, KURTOSIS AND MOMENTS

8.1 INTRODUCTION

The measures of central tendency tell us about the concentration of the observation about the middle of the distribution. The measures of dispersion give us an idea about the spread or scatter of the observations about some measures of central tendency. These two measures namely, central tendency and dispersion are inadequate to characterise a distribution completely and they must be supported and supplemented by three more measures namely, skewness, kurtosis and moments.

8.1.1 Skewness

Skewness means lack of symmetry. It means asymmetrical distribution. It is the study of shape of the curve of the frequency distribution.

8.2 DEFINITIONS

Some definitions of Skewness are given below:

When a series is not symmetrical it is said to be asymmetrical or skewed.

Croxton and Cowden

Skewness is lack of symmetry. When a frequency distribution is plotted on a chart, skewness present in the items tends to disperse chart more on one side of the mean than on the other.

Riggleman and Frishee

A distribution is said to be skewed, when the mean and median fall at different points in the distribution, and the balance (or centre of gravity) is shifted to one side or the other—to the left or right.

Garrett

Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution, the mean, median and mode are identical. The more the mean moves away from the mode, the larger the asymmetry or skewness.

Simpson and Kafka

It is clear from the above definitions that the word skewness refers to the lack of symmetry. If a distribution is normal, there would be no skewness in it and the curve drawn from the distribution would be symmetrical. In case of skewed distributions, the curve drawn would be tilted either to the left or to the right.

The following three figures would give an idea about the shape of symmetrical and asymmetrical curves.

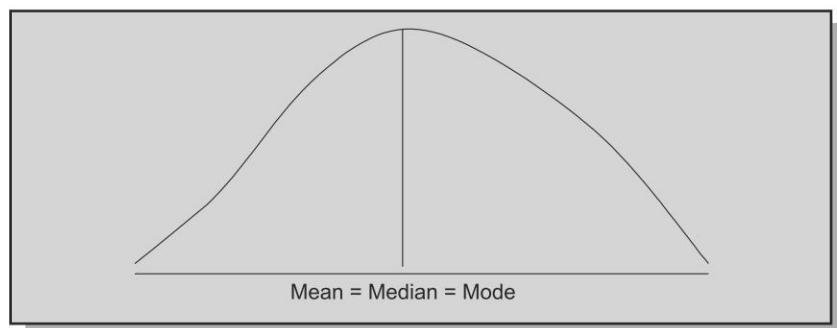


Fig. 8.1 Symmetrical Distribution

Figure 8.1 gives the shape of an ideal symmetrical curve. It is bell shaped and there is no skewness in it. The value of mean, median and mode in such a curve would coincide i.e., $\text{mean} = \text{median} = \text{mode}$

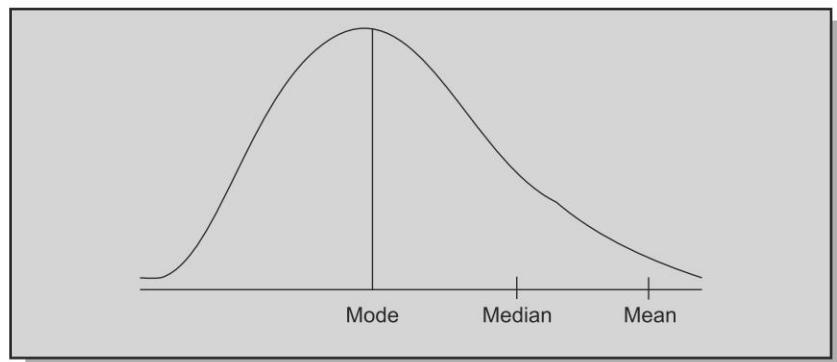


Fig. 8.2 Positively Skewed Distribution

Figure 8.2 gives the shape of a moderately skewed curve to the right. In it, the value of mean would be more than the values of median and mode. Median would have a higher value than that of the mode, that is, $\text{mean} > \text{median} > \text{mode}$. Such curves are called positively skewed.

Figure 8.3 represents the shape of a moderately skewed curve. This curve is skewed to the left and in it the value of mode would be greater than the value of

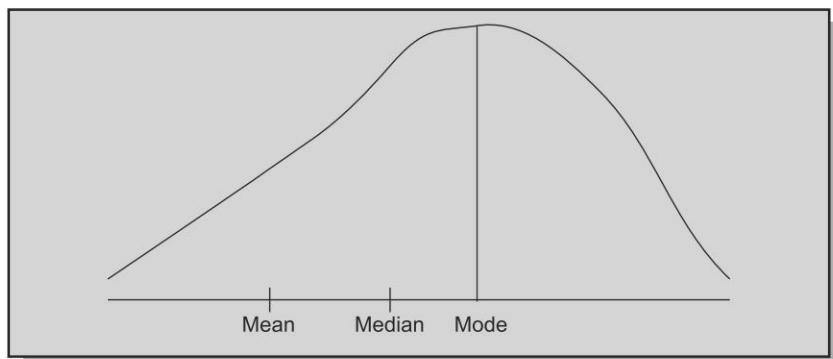


Fig. 8.3 Negatively Skewed Distribution

the median and the value of the median would be greater than the value of the mean, that is, $\text{Mean} < \text{Median} < \text{Mode}$. Such curves are called negatively skewed. Thus,

- (i) In a symmetrical distribution, $\text{Mean} = \text{Median} = \text{Mode}$
- (ii) In a positively skewed distribution, $\text{Mean} > \text{Median} > \text{Mode}$
- (iii) In a negatively skewed distribution, $\text{Mean} < \text{Median} < \text{Mode}$.

Difference between Dispersion and Skewness

Dispersion	Skewness
1. It deals with the spread of values around central value.	1. It deals with symmetry of distribution around central value.
2. It deals with the amount of variation.	2. It deals with the direction of variation.
3. It judges the truthfulness of the central values.	3. It judges the differences between the central values.

8.3 MEASURES OF SKEWNESS

The measures of asymmetry are usually called measures of skewness. The absolute measures are also known as measures of skewness. Measures of skewness indicate not only the extent of skewness, but also the direction. The absolute measure tells us the extent of asymmetry, whether it is positive or negative.

$$\text{Absolute Skewness } \} = \text{Mean} - \text{Mode}$$

If the answer is positive, then it means that the distribution is positively skewed. If the value of mean is less than the value of mode, it means the distribution is negatively skewed.

Relative Measures of Skewness

The following are the important relative measures of skewness:

1. Karl Pearson's co-efficient of Skewness.

2. Bowley's co-efficient of Skewness.
3. Kelly's co-efficient of Skewness.
4. Measures of Skewness based on moments.

I. Karl Pearson's Co-efficient of Skewness According to Karl Pearson, the co-efficient of Skewness should be measured through the difference of mean and mode divided by standard deviation. The formula for Karl Pearson's co-efficient of Skewness is—

$$\text{Co-efficient of Skewness } (Sk_p) = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

The result would be between ± 1 . If the answer is zero, then the distribution is symmetrical. If the answer is positive, then it is called positively skewed and if it is negative, then it is called negatively skewed.

In case, the mode is ill-defined, the co-efficient can be determined by the changed formula,

$$\text{Co-efficient of Skewness } (Sk_p) = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

The result of the Co-efficient of Skewness through this formula, may vary between ± 3 . But mostly it lies between ± 1 .

Illustration 8.1

Calculate Karl Pearson's co-efficient of Skewness for the following data:

45	25	43	60	47
----	----	----	----	----

Solutions

Calculation of mean and standard deviation

x	d (x - 43)	d ²
45	2	4
25	-18	324
43	0	0
60	17	289
47	4	16
$\Sigma d = 5$		$\Sigma d^2 = 633$

$$\begin{aligned}\text{Mean} &= A \pm \frac{\Sigma d}{N} \\ &= 43 + \frac{5}{5} = 43 + 1 = 44\end{aligned}$$

$$\text{S.D.} = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N} \right)^2}$$

$$\begin{aligned}
 &= \sqrt{\frac{633}{5} - \left(\frac{5^2}{5}\right)} \\
 &= \sqrt{126.6 - 1} \\
 &= \sqrt{125.6} \\
 &= 11.21
 \end{aligned}$$

Ascending order of the data – 25, 43, 45, 47, 60

$$\begin{aligned}
 \text{Median} &= N + 1/2 \text{th item} \\
 &= 5 + 1/2 \text{th item} \\
 &= 3 \text{rd item i.e., } = 45
 \end{aligned}$$

Since it is difficult to identify the mode, the following may be applied for calculating the Co-efficient of Skewness.

Co-efficient of Skewness

$$\begin{aligned}
 Sk_p &= \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} \\
 &= \frac{3(44 - 45)}{11.21} = \frac{3x - 1}{11.21} = \frac{-3}{11.21} \\
 &= -0.27
 \end{aligned}$$

Illustration 8.2

The marks obtained by 100 students of a class are given below.

Marks	40	42	47	52	65
No. of students	19	21	27	10	13

Calculate Karl Pearson's co-efficient of Skewness.

Solutions

Calculation of Mean and Standard Deviation

x	f	d = (x - a) = (x - 45)	d²	fd	fd²
40	19	-5	25	-95	475
42	21	-3	9	-63	189
47	27	2	4	54	108
52	10	7	49	70	490
65	13	20	400	260	5200
90				$\Sigma fd = 226$	$\Sigma fd^2 = 6462$

$$\text{Mean} = A \pm \frac{\sum fd}{N}$$

$$= 45 + \frac{226}{90} \\ = 45 + 2.51 = 47.51$$

Mode = 47.51

$$\text{S.D.} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \\ = \sqrt{\frac{6462}{90} - \left(\frac{226}{90}\right)^2} \\ = \sqrt{71.8 - 6.30} \\ = \sqrt{65.5} = 8.09$$

$$\text{Karl Pearson's co-efficient of Skewness } (Sk_p) = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} \\ = \frac{47.51 - 47}{8.09} \\ = \frac{0.51}{8.09} \\ = 0.06$$

Illustration 8.3

Calculate Karl Pearson's co-efficient of Skewness from the following data.

Marks	0–20	20–40	40–60	60–80	80–100
No. of students	14	21	25	16	20

Solutions

Calculation of mean, mode and standard deviation

x	f	c.f.	Mid-value	d' $(m - a)/c$	fd	fd ²
				$\frac{m - 50}{20}$		
0–20	14	14	10	-2	-28	56
20–40	21	35	30	-1	-21	21
40–60	25	60	50	0	0	0
60–80	16	76	70	1	16	16
80–100	20	96	90	2	40	80
N = 96				$\Sigma fd = 7$ $\Sigma fd^2 = 173$		

$$\begin{aligned}\text{Mean} &= A + \frac{\sum fd}{N} \times C \\ &= 50 + \frac{7}{96} \times 20 \\ &= 50 + 1.46 \\ &= 51.46\end{aligned}$$

Calculation of Mode

Modal class : 40 – 60

Hence, $L = 40$

$$\begin{aligned}\Delta 1 &= f_1 - f_0 = 25 - 21 = 4 \\ \Delta 2 &= f_1 - f_2 = 25 - 16 = 9\end{aligned}$$

$$\begin{aligned}\text{Mode} &= A + \frac{\Delta 1}{\Delta 1 + \Delta 2} \times 1 \\ &= 40 + \frac{4}{4+9} \times 20 \\ &= 40 + 6.15\end{aligned}$$

Mode = 46.15

$$\begin{aligned}\text{S.D.} &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C \\ &= \sqrt{\frac{173}{96} - \left(\frac{7}{96} \right)^2} \times 20 \\ &= \sqrt{1.80 - 0.005} \times 20 \\ &= \sqrt{1.795} \times 20 \\ &= 1.34 \times 20\end{aligned}$$

S.D. = 26.8

$$\begin{aligned}Sk_p &= \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} \\ &= \frac{51.46 - 46.15}{26.8} = \frac{5.31}{26.8} \\ &= 0.20\end{aligned}$$

Illustration 8.4

From the marks secured by 110 students in section *A*, 110 students in section *B* of a class, the following measures are obtained.

$$\text{Section } A \quad \bar{x} = 36.83 \quad \text{S.D.} = 14.8 \quad \text{Mode} = 41.67$$

$$\text{Section } B \quad \bar{x} = 37.83 \quad \text{S.D.} = 14.8 \quad \text{Mode} = 37.07$$

Determine which distribution of marks is more skewed.

Solutions

We have to compute the Co-efficient of Skewness for both the sections, *A* and *B*.

Section A :

$$\begin{aligned} Sk_p &= \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation} (\sigma)} \\ &= \frac{36.83 - 41.67}{14.8} \\ &= \frac{-4.84}{14.8} \\ &= -0.327 \end{aligned}$$

Section B:

$$\begin{aligned} Sk_p &= \frac{\text{Mean} - \text{Mode}}{\sigma} \\ &= \frac{37.83 - 37.07}{14.8} \\ &= \frac{0.76}{14.8} \\ &= 0.0574 \end{aligned}$$

Thus, we find that distribution of marks of section *A* is more skewed.

Illustration 8.5

From a moderately skewed distribution of retail price for women's chappals, it is found that the mean price is Rs. 40 and the median price is Rs. 37. If the co-efficient of variation is 40%, find the Pearsonian co-efficient of Skewness of the distribution.

$$Sk = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Mean = 40 and median = 37 are given in the problem. To find the Co-efficient of Skewness, we need standard deviation.

$$\text{C.V.} = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$40 = \frac{\sigma}{40} \times 100$$

$$2.5\sigma = 40$$

$$\sigma = 40/2.5 = 16$$

$$\begin{aligned} Sk &= \frac{3(40 - 37)}{16} \\ &= \frac{3 \times 3}{16} = \frac{9}{16} \\ &= 0.56 \end{aligned}$$

2. Bowley's Co-efficient of Skewness Bowley's co-efficient of Skewness is based upon the quartiles of the distribution. This measure is otherwise called as quartile measure of Skewness. The formula under this method is stated as below.

$$Sk_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 + Q_2) - (Q_2 + Q_1)}$$

$$\text{or } Sk_B = \frac{Q_3 - Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

where Q_3 = third quartiles

Q_2 = second quartile or median and

Q_1 = first quartile

The value of this co-efficient of skewness varies between ± 1 .

Illustration 8.6

Calculate Bowley's co-efficient of Skewness for the data given below.

Monthly Income (Rs.'000)	35	45	50	55	60	65	70
No. of Workers	17	19	19	22	20	19	15

Solutions

Calculation of Bowley's co-efficient of Skewness

Income (Rs. '000)	No. of workers f	c.f.
35	17	17
45	19	36
50	19	55
55	22	77
60	20	97
65	19	116
70	15	131

$$\text{Median } Q_2 = \frac{N+1}{2} \text{ th item} = \frac{131+1}{2} \text{ th item} = 66 \text{th item}$$

i.e., Rs. 55.

$$Q_3 = \frac{3(N+1)}{4} \text{ th item} = \frac{3(131+1)}{4} \text{ th item} = 99 \text{th item}$$

i.e., Rs. 65

Bowley's co-efficient of Skewness

$$\begin{aligned} Sk_B &= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} \\ &= \frac{65 + 45 - (2 \times 55)}{65 - 45} \\ &= \frac{110 - 110}{20} \\ &= 0/20 = 0 \end{aligned}$$

Illustration 8.7

Calculate Bowley's co-efficient of Skewness from the following data

Expenses (Rs)	0–10	10–20	20–30	30–40	40–50
No. of families	4	11	8	17	22

Solutions

Calculation of co-efficient of Skewness

Expenses (Rs)	No. of families f	Available No. of families c.f.
0–10	4	4
10–20	11	15
20–30	8	23
30–40	17	40
40–50	22	62

262 Business Statistics

Median = $N/2$ th item = 62/2th item = 31st item.

It lies in class interval 30–40

Hence, $L = 30$, $cf = 23$, $f = 17$, $i = 10$

$$\begin{aligned} &= L + \frac{N/2 - cf}{f} \times C \\ &= 30 + \frac{62/2 - 23}{17} \times 10 \\ &= 30 + \frac{31 - 23}{17} \times 10 \\ &= 30 + 4.71 \end{aligned}$$

$$\text{Median} = 34.71$$

$Q_1 = N/4$ th item = 62/4th item = 15.5th item.

It lies in class interval 20–30

Hence, $L = 20$, $cf = 15$, $f = 8$, $i = 10$ (30–20)

$$\begin{aligned} Q_1 &= L + \frac{N/4 - cf}{f} \times i \\ &= 20 + \frac{62/4 - 15}{8} \times 10 \\ &= 20 + \frac{15.5 - 15}{8} \times 10 \\ &= 20 + 0.63 \end{aligned}$$

$$Q_1 = 20.63$$

$Q_3 = 3N/4$ th item = $(3 \times 62/4)$ th item = 46.5th item.

It lies in class interval 40–50

Hence, $L = 40$, $cf = 40$, $f = 22$, $i = 10$ (50–40)

$$\begin{aligned} Q_3 &= L + \frac{3N/4 - cf}{f} \times i \\ &= 40 + \frac{3 \times 62/4 - 40}{22} \times 10 \\ &= 40 + 2.95 \end{aligned}$$

$$Q_3 = 42.95$$

Bowley's co-efficient of Skewness

$$\begin{aligned} Sk_B &= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} \\ &= \frac{42.95 + 20.63 - (2 \times 34.71)}{42.95 - 20.63} \end{aligned}$$

$$\begin{aligned}
 &= \frac{63.58 - 69.42}{22.32} \\
 &= \frac{-5.84}{22.32} \\
 &= -0.26
 \end{aligned}$$

3. Kelly's Co-efficient of Skewness It is based upon the percentiles or deciles of the distribution. It covers the beginning and end of the distribution. Here, the distribution is analysed on the basis of the 10th and 90th percentiles.

The formula is:

$$\begin{aligned}
 Sk_k &= \frac{p_{90} + p_{10} - 2p_{50}}{p_{90} - p_{10}} \\
 \text{or} \quad &= \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}
 \end{aligned}$$

Where, p_{90} is 90th percentile; p_{10} is 10th percentile; p_{50} is 50th percentile or median; D_9 is Decile 1 and D_5 is Decile 5 or median.

Illustration 8.8

Calculate Kelly's co-efficient of Skewness from the following data.

x	110	135	160	185	210	235	260
f	22	28	35	40	20	30	25

Solutions

Calculation of Kelly's co-efficient of Skewness

x	f	cf
110	22	22
135	28	50
160	35	85
185	40	125
210	20	145
235	30	175
260	25	200
N = 200		

$$\begin{aligned}
 \text{Median } (p_{50}) &= \frac{N+1}{2} \text{th item} = \frac{200+1}{2} \text{th item} \\
 &= 100.5 \text{th item i.e., } = 185
 \end{aligned}$$

$$p_{90} = \frac{90(N+1)}{100} \text{th item}$$

$$\begin{aligned}
 &= \frac{90(200+1)}{100} = \frac{18090}{100} \text{th item} \\
 &= 180.9 \text{th item} \\
 &= 260 \\
 p_{10} &= \frac{10(N+1)}{100} \text{th item} \\
 &= \frac{10(200+1)}{100} = \frac{2010}{100} \text{th item} \\
 &= 20.1 \text{th item} \\
 \text{i.e., } &= 110
 \end{aligned}$$

Kelly's Co-efficient of Skewness

$$\begin{aligned}
 Sk_k &= \frac{p_{90} + p_{10} - 2p_{50}}{p_{90} - p_{10}} \\
 &= \frac{260 + 110 - (2 \times 185)}{260 - 110} \\
 &= \frac{370 - 370}{150} = \frac{0}{150} = 0
 \end{aligned}$$

Illustration 8.9

Compute Kelly's co-efficient of Skewness from the following data.

Monthly Expenses	No. of families
0–30	25
30–60	29
60–90	37
90–120	50
120–150	41

Solutions

Calculation of Kelly's co-efficient of Skewness

Monthly Expenses (Rs) x	No. of families f	Cumulative No. of families cf
0–30	25	25
30–60	29	54
60–90	37	91
90–120	50	141
120–150	41	182
N = 182		

Median = $N/2$ th item = 182/2th item = 91th item.

It lies in class interval 60–90

$$\begin{aligned} &= L + \frac{N/2 - cf}{F} \times i \\ &= 60 + \frac{91 - 54}{37} \times 30 \\ &= 60 + 30 \end{aligned}$$

Median = 90

$$\begin{aligned} p_{90} &= \frac{90 N}{100} \text{th item} \\ &= \frac{90 \times 182}{100} \\ &= 163.8 \text{th item} \end{aligned}$$

It lies in class interval 120–150.

Hence, $L = 120$, $cf = 141$, $f = 41$, $i = 50$ ($150 - 120$)

$$\begin{aligned} p_{90} &= L + \frac{90 N/100 - cf}{f} \times i \\ &= 120 + \frac{90 \times 182/100 - 141}{41} \times 30 \\ p_{90} &= 120 + 16.68 = 136.68 \\ p_{10} &= \frac{10 N}{100} \text{th item} \\ &= \frac{10 \times 182}{100} \text{th item} \\ &= 18.2 \text{th item} \end{aligned}$$

It lies in class interval 0 – 30

Hence, $L = 0$, $cf = 0$, $f = 25$, $i = 30$ ($30 - 0$)

$$\begin{aligned} p_{10} &= L + \frac{10 N/100 - cf}{100} \times i \\ &= 0 + \frac{18.2 - 0}{25} \times 30 \\ &= 0 + 21.84 \\ &= 21.84 \end{aligned}$$

Kelly's co-efficient of Skewness

$$\begin{aligned} Sk_k &= \frac{p_{90} + p_{10} - 2 \text{ Median}}{p_{90} - p_{10}} \\ &= \frac{136.68 + 21.84 - (2 \times 90)}{136.68 - 21.84} \\ &= \frac{158.52 - 180}{114.84} = \frac{21.48}{114.84} = -0.19 \end{aligned}$$

8.4 MOMENTS

The term moment is generally used in physics, mechanics which means “movement of a force” that turn a pivoted lever. When it is applied in Statistics, it describes the various characteristics of frequency distribution, namely, central tendency, dispersion, skewness and kurtosis. Moments can be defined as the arithmetic mean of various powers of deviation taken from the mean of a distribution. The mean of the first power is called first moment; mean of the second power is called second moment and so on. Moment is denoted by Greek letter μ (pronounced as mu). Hence, the first moment is called μ_1 , second moment as μ_2 and so on.

8.4.1 Calculation of Moments for Individual Observation

$$\begin{aligned}\mu_1 &= \frac{\sum(x - \bar{x})}{N} \quad \text{or} \quad \frac{\sum x}{N} \\ \mu_2 &= \frac{\sum(x - \bar{x})^2}{N} \quad \text{or} \quad \frac{\sum x^2}{N} \\ \mu_3 &= \frac{\sum(x - \bar{x})^3}{N} \quad \text{or} \quad \frac{\sum x^3}{N} \\ \mu_4 &= \frac{\sum(x - \bar{x})^4}{N} \quad \text{or} \quad \frac{\sum x^4}{N}\end{aligned}$$

where, $x(x - \bar{x})\mu$, will always be equal to zero since the sum of deviations of the data from arithmetic mean is always zero.

8.4.2 Calculation of Moments for Frequency Distribution

(i.e., for discrete series and continuous series)

$$\begin{aligned}\mu_1 &= \frac{\sum f(x - \bar{x})}{N} \quad \text{or} \quad \frac{\sum fx}{N} \\ \mu_2 &= \frac{\sum f(x - \bar{x})^2}{N} \quad \text{or} \quad \frac{\sum f x^2}{N} \\ \mu_3 &= \frac{\sum f(x - \bar{x})^3}{N} \quad \text{or} \quad \frac{\sum f x^3}{N} \\ \mu_4 &= \frac{\sum f(x - \bar{x})^4}{N} \quad \text{or} \quad \frac{\sum f x^4}{N}\end{aligned}$$

where, $x(x - \bar{x})$ for the continuous series, mid-value m should be taken instead of x .

8.4.3 Calculation of Skewness and Kurtosis from Moments

$$1. \text{ Skewness: } \beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

(β_1 is called Beta one which indicates Skewness)

$$2. \text{ Kurtosis: } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

(β_2 is called Beta two which indicates Kurtosis)

8.4.4 Moments based on Arbitrary Origin

The Assumed mean can also be used for calculating moments, instead of Actual mean. The moments calculated through assumed mean is called raw moments which can be denoted as μ'_1 . It should be converted into real moments through other formula. The details are given below:

For individual observation:

$$\mu'_1 = \frac{\sum(x - A)}{N} \quad \text{or} \quad \frac{\sum d}{N}$$

$$\mu'_2 = \frac{\sum(x - A)^2}{N} \quad \text{or} \quad \frac{\sum d^2}{N}$$

$$\mu'_3 = \frac{\sum(x - A)^3}{N} \quad \text{or} \quad \frac{\sum d^3}{N}$$

$$\mu'_4 = \frac{\sum(x - A)^4}{N} \quad \text{or} \quad \frac{\sum d^4}{N}$$

For frequency distribution,

$$\mu'_1 = \frac{\sum f(x - A)}{N} \quad \text{or} \quad \frac{\sum fd}{N} \quad \text{or} \quad \frac{\sum fd'}{N} \times C$$

$$\mu'_2 = \frac{\sum f(x - A)^2}{N} \quad \text{or} \quad \frac{\sum fd^2}{N} \quad \text{or} \quad \frac{\sum fd'^2}{N} \times C$$

$$\mu'_3 = \frac{\sum f(x - A)^3}{N} \quad \text{or} \quad \frac{\sum fd^3}{N} \quad \text{or} \quad \frac{\sum fd'^3}{N} \times C$$

$$\mu'_4 = \frac{\sum f(x - A)^4}{N} \quad \text{or} \quad \frac{\sum fd^4}{N} \quad \text{or} \quad \frac{\sum fd'^4}{N} \times C$$

where, $d' = \frac{x - A}{C}$, C = Common factor

For continuous series, mid-value, (' m) should be used instead of x . Conversion of moments based on arbitrary origin (assumed mean) into moments based on arithmetic mean.

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2$$

$$\mu_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2(\mu'_2) - 3(\mu'_1)^4$$

Illustration 8.10

Calculate the first four moments from the data given below.

26	28	30	50	5	6	9
----	----	----	----	---	---	---

Solutions

Calculation of moments

x	$d = (x - \bar{x})$	$(x - \bar{x})$	$(x - \bar{x})^2 d^2$	$(x - \bar{x})^3 d^3$	$(x - \bar{x})^4 d^4$
26	4		16	64	256
28	6		36	216	1296
30	8		64	512	4096
50	28		784	21952	614656
5	-17		289	-4913	83521
6	-16		256	-4096	65536
9	-13		169	-2197	28561
$\Sigma x = 154$		$\Sigma d = 0$	$\Sigma d^2 = 1614$	$\Sigma d^3 = 11538$	$\Sigma d^4 = 797922$

$$\text{Arithmetic mean} = \frac{\sum x}{N} = \frac{154}{7} = 22$$

$$\mu_1 = \frac{\sum d}{N} = \frac{9}{7} = 0$$

$$\mu_2 = \frac{\sum d^2}{N} = \frac{1614}{7} = 230.6$$

$$\mu_3 = \frac{\sum d^3}{N} = \frac{11538}{7} = 1648.3$$

$$\mu_4 = \frac{\sum d^4}{N} = \frac{797922}{7} = 113988.9$$

Illustration 8.11

Calculate the first four moments from the particulars given below.

x	5	10	15	20	25
f	2	3	4	5	1

Solutions

Calculation of Moments

x	f	fx	d(x - 15)	fd	fd²	fd³	fd⁴
5	2	10	-10	-20	200	2000	20000
10	3	30	-5	-15	75	375	1875
15	4	60	0	0	0	0	0
20	5	100	5	25	125	625	3125
25	1	25	10	10	100	1000	10000
$\Sigma x = 154$	$N = 15$	$\Sigma fx = 225$		$\Sigma fd = 0$	$\Sigma fd^2 = 500$	$\Sigma fd^3 = 1625$	$\Sigma fd^4 = 35000$

$$\text{Arithmetic mean} = \frac{\sum fx}{N} = \frac{225}{15} = 15$$

$$\mu_1 = \frac{\sum fd}{N} = \frac{0}{15} = 0$$

$$\mu_2 = \frac{\sum fd^2}{N} = \frac{500}{15} = 33.33$$

$$\mu_3 = \frac{\sum fd^3}{N} = \frac{1625}{15} = 108.33$$

$$\mu_4 = \frac{\sum fd^4}{N} = \frac{35000}{15} = 2333.33$$

Illustration 8.12

Compute the first four moments from the data given below.

x	0–20	20–40	40–60	60–80	80–100
f	3	5	4	1	2

Solutions

Calculation of moments

x	f	m	d(m - 42)	fd	fd²	fd³	fd⁴
0–20	3	10	-32	-96	3072	-98304	3145728
20–40	5	30	-12	-60	720	-8640	103680

x	f	m	d(m - 42)	fd	fd ²	fd ³	fd ⁴
40–60	4	50	8	32	256	2048	16384
60–80	1	70	28	28	784	21952	614656
80–100	2	90	48	96	4608	221184	10616832
N =	$\Sigma fm =$			$\Sigma fd =$	$\Sigma fd^2 =$	$\Sigma fd^3 =$	$\Sigma fd^4 =$
15	630			0	9440	138240	14497280

$$\text{Arithmetic mean} = \frac{\sum fm}{N} = \frac{630}{15} = 42$$

$$\mu_1 = \frac{\sum fd}{N} = \frac{0}{15} = 0$$

$$\mu_2 = \frac{\sum fd^2}{N} = \frac{9440}{15} = 629.33$$

$$\mu_3 = \frac{\sum fd^3}{N} = \frac{138240}{15} = 9216$$

$$\mu_4 = \frac{\sum fd^4}{N} = \frac{14497280}{15} = 966485.33$$

Illustration 8.13

Calculate the first four moments from the data given below.

x	10–20	20–30	30–40	40–50	50–60
f	5	4	2	6	8

Solutions

Calculation of moments through assumed mean

x	f	m	$d'(m - 35)/10$	fd'	fd^2	fd^3	fd^4
10–20	5	15	-2	-10	20	-40	80
20–30	4	25	-1	-4	4	-4	4
30–40	2	35	0	0	0	0	0
40–50	6	45	1	6	6	6	6
50–60	8	55	2	16	32	64	128
$\Sigma f = 25$				$\Sigma fd' =$	$\Sigma fd^2 =$	$\Sigma fd^3 =$	$\Sigma fd^4 =$
				8	62	26	218

Moments based on Arbitrary Mean:

$$\mu'_1 = \frac{\sum fd}{N} \times i = \frac{8}{25} \times 10 = 3.2$$

$$\mu'_2 = \frac{\sum fd^2}{N} \times i^2 = \frac{62}{25} \times 100 = 248$$

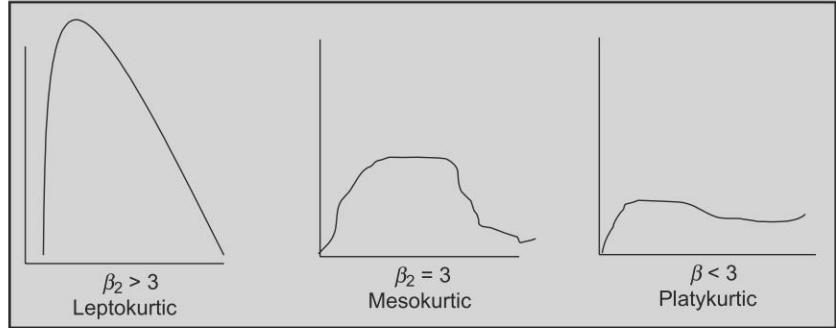
$$\begin{aligned}
 \mu'_3 &= \frac{\sum fd^3}{N} \times i^3 = \frac{26}{25} \times 1000 = 1040 \\
 \mu'_4 &= \frac{\sum fd^4}{N} \times i^4 = \frac{218}{25} \times 10000 = 87200 \\
 \mu_1 &= \mu'_1 - \mu'_1 = 3.2 - 3.2 = 0 \\
 \mu_2 &= \mu'_2 - (\mu'_1)^2 = 248 - 10.24 = 237.76 \\
 \mu_3 &= \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3 \\
 &= 1040 - 3 \times 3.2 \times 248 + 2(3.2)^3 \\
 &= 1040 - 2380.8 + 65.536 \\
 &= -1275.264 \\
 \mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2(\mu'_2) - 3(\mu'_1)^4 \\
 &= 87200 - 4(3.2 \times 1040) 6(3.2)^2 (248) - 3(3.2)^4 \\
 &= 87200 - 13312 + 15237.12 - 314.573 \\
 &= 88810.547
 \end{aligned}$$

8.5 KURTOSIS

Kurtosis is a Greek word, which means bulginess. It refers to the measure of the peakness or flatness of the distribution of data. The extent of the peakness of the distribution from the normal distribution can be measured through Kurtosis.

- If the curve is more peaked than the normal curve, then it is called ‘Leptokurtic’.
- If the curve is more flat typed than the normal curve, then it is called ‘Platykurtic’.
- If the normal curve is called ‘Mesokurtic’.

The following diagrams are representation of Kurtosis:



The Kurtosis can be calculated through the following formula:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

272 Business Statistics

where, β_2 means Kurtosis; μ_4 indicates 4th moment; μ_2 indicates 2nd moment.

If β_2 is equal to 3 for a normal curve, it is Mesokurtic. If value of $\beta_2 > 3$, then the curve is more peaked than the normal curve which is called Leptokurtic. If the value of $\beta_2 < 3$, then the curve is less peaked than the normal curve, which is called Platykurtic.

Illustration 8.14

Calculate Kurtosis from the following data:

16	8	6	12	3	5	6
----	---	---	----	---	---	---

Solutions

Calculation of Kurtosis

x	$d = X - A$	d^2	d^4
16	8	64	4096
8	0	0	0
6	-2	4	16
12	4	16	256
3	-5	25	625
5	-3	9	81
6	-2	4	16
$\sum x = 56$		$\sum d^2 = 122$	$\sum d^4 = 5090$

$$\text{Arithmetic mean } (\bar{x}) = \frac{\sum x}{N} = \frac{56}{7} = 8$$

$$\mu_2 = \frac{\sum d^2}{N} = \frac{122}{7} = 17.43$$

$$\mu_4 = \frac{\sum d^4}{N} = \frac{5090}{7} = 727.14$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{727.14}{(17.43)^2} = \frac{727.14}{303.80}$$

$$\beta_2 = 2.39$$

Since $\beta_2 < 3$, it is Platykurtic.

Illustration 8.15

Calculate Kurtosis from the following data.

x	15	8	12	10	7	25
f	4	2	3	5	1	6

Solutions

Calculation of Kurtosis

<i>x</i>	<i>f</i>	<i>d(x - 12)</i>	<i>fd</i>	<i>fd²</i>	<i>fd³</i>	<i>fd⁴</i>
15	4	3	12	36	108	324
8	2	-4	-8	32	-128	512
12	3	0	0	0	0	0
10	5	-2	-10	20	-40	80
7	1	-5	-5	25	-125	625
25	6	13	78	1014	13182	171366
<i>N = 21</i>		$\sum fd = 67$		$\sum d^2 = 1127$	$\sum fd^3 = 12997$	$\sum fd^4 = 172907$

$$\mu'_1 = \frac{\sum fd}{N} = \frac{67}{21} = 3.19$$

$$\mu'_2 = \frac{\sum fd^2}{N} = \frac{1127}{21} = 53.67$$

$$\mu'_3 = \frac{\sum fd^3}{N} = \frac{12997}{21} = 618.90$$

$$\mu'_4 = \frac{\sum fd^4}{N} = \frac{172907}{21} = 8233.67$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 53.67 - (3.19)^2$$

$$= 53.67 - 10.18 = 43.49$$

$$\begin{aligned}\mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2(\mu'_2) - 3(\mu'_1)^4 \\ &= 8233.67 - 4(3.19)(618.90) + (1040)6(3.19)^2(53.67) - 3(3.19)^4 \\ &= 8233.67 - 7897.16 + 3276.91 - 310.66 \\ &= 3302.76\end{aligned}$$

$$\text{Kurtosis, } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3302.76}{(43.49)^2}$$

$$= \frac{3302.76}{1891.38} = 1.75$$

Since $\beta_2 < 3$, it is called Platykurtic.

8.6 MISCELLANEOUS ILLUSTRATIONS

Illustration 8.16

Calculate Skewness from the following distribution.

x	1	2	3	4	5	6	7
Y	10	14	5	3	8	14	16

Solutions

Calculation of Skewness

x	f	d = x - 4	fd	fd²
1	10	-3	-30	90
2	14	-2	-28	56
3	5	-1	-5	5
4	3	0	0	0
5	8	1	8	8
6	14	2	28	56
7	16	3	48	144
N = 70		$\sum fd = 21$ $\sum fd^2 = 359$		

$$\text{Mean} = A + \frac{\sum fd}{N}$$

$$= 4 + \frac{21}{70} \\ = 4 + 0.30 \\ = 4.30$$

$$\text{S.D.} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2}$$

$$\text{S.D.} = \sqrt{\frac{359}{10} - \left(\frac{21}{70} \right)^2} \\ = \sqrt{5.13 - 0.09} \\ = \sqrt{5.04}$$

$$\sigma = 2.04$$

The highest frequency is 16. Mode is 7.

$$S_k = \frac{\text{Mean} - \text{Mode}}{\text{S.D.}}$$

$$= \frac{4.30 - 7}{2.04}$$

$$= -1.32$$

Illustration 8.17

Calculate Skewness for the following:

Weekly wages (Rs)	150	200	250	300	350	400	450
No. of earners	9	26	18	16	13	0	8

Solutions

Calculation of Skewness

Weekly Wages	$d = \frac{x - 300}{50}$	Wages of Earners f	d^2	fd	fd^2
150	-3	9	9	-27	81
200	-2	26	4	-52	104
250	-1	18	1	-18	18
300	0	16	0	0	0
350	1	13	1	13	13
400	2	10	4	20	40
450	3	8	9	24	72
$N = 100$			$\sum fd = -40$		
			$\sum fd^2 = 328$		

$$\text{Mean} = A \pm \frac{\sum fd}{N} \times C$$

$$A = 300 \quad \sum fd = -40 \quad C = 50$$

$$= 300 - \frac{-40}{100}$$

$$= 300 - 20$$

$$\bar{X} = 280$$

$$\text{S.D.} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C$$

$$\text{S.D.} = \sqrt{\frac{328}{100} - \left(\frac{-18}{100} \right)^2} \times 50$$

$$= \sqrt{3.28 - 0.16} \times 50$$

$$\sigma = 1.77 \times 50 = 88.5$$

By Inspection method, mode is 200.

$$\text{Co-efficient of Skewness} = \frac{\bar{X} - \text{Mode}}{\sigma}$$

$$\text{where, } \bar{X} = 280 \quad \text{Mode} = 200 \quad \sigma = 88.5$$

$$= \frac{280 - 200}{88.5} = \frac{80}{88.5} \\ = 0.90$$

Illustration 8.18

Calculate the co-efficient of Skewness for the following data.

Height in inches	152	153	154	155	156	157	158
No. of persons	9	20	32	36	27	18	6

Solutions

Calculation of Skewness

Weekly Wages	$d = x - 155$	d^2	f	fd	fd^2
152	-3	9	9	-27	81
153	-2	4	20	-40	80
154	-1	1	32	-32	32
155	0	0	36	0	0
156	1	1	27	27	27
157	2	4	18	36	72
158	3	9	6	18	54
$N = 148$			$\sum fd = -18$	$\sum fd^2 = 346$	

$$\begin{aligned}\text{Mean} &= A + \frac{\sum fd}{N} \\ &= 155 + \frac{-18}{148} \\ &= 155 - 0.12\end{aligned}$$

$$\bar{X} = 154.88$$

$$\text{S.D.} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2}$$

$$\begin{aligned} \text{S.D.} &= \sqrt{\frac{346}{148} - \left(\frac{-18}{148}\right)^2} \\ &= \sqrt{.34 - 0.01} = \sqrt{2.33} \\ \sigma &= 1.52 \end{aligned}$$

The highest frequency is 36. Mode is 155.

$$\begin{aligned} \text{Co-efficient of Skewness} &= \frac{\text{Mean} - \text{Mode}}{\sigma} \\ &= \frac{154.88 - 155}{1.52} = \frac{-0.12}{1.52} \\ &= -0.08 \end{aligned}$$

Illustration 8.19

Calculate Karl Pearson's co-efficient of skewness for the following data.

28 15 38 26 28 22 20 24 18

Solutions

Calculation of Skewness

Data x	$d = x - 28$	d^2
28	0	0
15	-13	169
38	10	100
26	-2	4
28	0	0
22	-6	36
20	-8	64
24	-4	16
18	-10	100
$\sum d = -33$		$\sum d^2 = 489$

$$\begin{aligned} \text{Mean} &= A + \frac{\sum d}{N} \\ &= 28 + \frac{-33}{9} \\ &= 28 - 3.67 \\ \bar{X} &= -24.33 \end{aligned}$$

$$\text{S.D.} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$\text{S.D.} = \sqrt{\frac{346}{148} - \left(\frac{-18}{148}\right)^2}$$

$$= \sqrt{54.33 - 13.20} = \sqrt{41.13}$$

$$\sigma = 6.41$$

Mode = 28

$$\text{Karl Pearson's co-efficient of Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$= \frac{24.33 - 28}{6.43} = \frac{-3.67}{6.43}$$

$$S_k = -0.57$$

Illustration 8.20

Calculate the Pearson's measures on the basis of mean, mode and standard deviation.

x	14.5	15.5	16.5	17.5	18.5	19.5	20.5	21.5
y	70	80	96	200	250	174	86	44

Solutions

Calculation of Skewness

x	f	d = x - 18.5	d²	fd	fd²
14.5	70	-4	16	-280	1120
15.5	80	-3	9	-240	720
16.5	96	-2	4	-192	384
17.5	200	-1	1	-200	200
18.5	250	0	0	0	0
19.5	174	1	1	174	174
20.5	86	2	4	172	344
21.5	44	3	9	132	396
N = 1000				$\sum fd = -434$	$\sum fd^2 = 3338$

$$\text{Mean} = A + \frac{\sum fd}{N}$$

$$\begin{aligned}
 &= 18.5 + \frac{-434}{1000} \\
 &= 18.5 - 0.434 \\
 \bar{X} &= 18.066 \\
 \text{S.D.} &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \\
 \text{S.D.} &= \sqrt{\frac{3338}{1000} - \left(\frac{-434}{1000}\right)^2} \\
 &= \sqrt{3.338 - 0.188} = \sqrt{3.15} \\
 \sigma &= 1.77
 \end{aligned}$$

Mode = 18.5. Because, the highest frequency is 250.

$$\begin{aligned}
 \text{Co-efficient of Skewness} &= \frac{\text{Mean} - \text{Mode}}{\sigma} \\
 &= \frac{18.66 - 18.5}{1.77} = \frac{-0.434}{1.77} \\
 S_k &= -0.245
 \end{aligned}$$

Illustration 8.21

Calculate the Pearson's co-efficient of Skewness from the following data.

Marks	05	10	20	30	40	50	60	70	80
No. of students	75	70	50	40	40	35	15	7	0

Solutions

Calculation of Skewness

Marks x	m	d = m - 45/10	d²	f	fd	fd²	c.f
0 – 10	5	-4	16	5	-20	80	5
10 – 20	15	-3	9	20	-60	180	25
20 – 30	25	-2	4	10	-20	40	35
30 – 40	35	-1	1	0	0	0	35
40 – 50	45	0	0	5	0	0	40
50 – 60	55	1	1	20	20	20	60
60 – 70	65	2	4	8	16	32	68
70 – 80	75	3	9	7	21	63	75
N = 75				$\sum fd = -43$	$\sum fd^2 = 415$		

$$\text{Mean} = A + \frac{\sum fd}{N} \times C$$

$$= 45 + \frac{-43}{1000}$$

$$= 45 - 5.7$$

$$\bar{X} = 39.3$$

$$\text{S.D.} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C$$

$$\text{S.D.} = \sqrt{\frac{415}{75} - \left(\frac{-43}{75} \right)^2}$$

$$= \sqrt{5.53 - 0.32} \times 10 = \sqrt{5.21} \times 10$$

$$= 2.28 \times 10$$

$$\sigma = 22.8$$

Mode is ill-defined; therefore, we must use median to calculate the co-efficient of skewness.

Median = $N/2 = 75/2 = 37.5$ th item which lies in 40 – 50

$$\text{Median} = L_1 + \frac{N/2 - cf}{f} \times i$$

$$L_1 = 40 \quad cf = 35 \quad f = 5 \quad i = 10$$

$$= 40 + \frac{75/2 - 35}{5} \times 10$$

$$= 40 + \frac{37.5 - 35}{5} \times 10$$

$$= 40 + 5 = 45$$

$$\text{Co-efficient of Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$$

$$= \frac{3(39.3 - 45)}{22.8}$$

$$= \frac{3 \times -5.7}{22.8}$$

$$= \frac{-17.1}{22.8}$$

$$S_k = -0.75$$

Illustration 8.22

Calculate the Pearson's co-efficient of Skewness from the following data.

Life time (Hours)	No. of Tubes
200 – 300	18
300 – 400	50
400 – 500	62
500 – 600	86
600 – 700	74
700 – 800	64
800 – 900	52
900 – 1000	28
1000 – 1100	16
	450

Solutions

Calculation of Skewness

Life time	m	$d = m - 650/100$	d^2	No. of Tubes	f	fd	fd^2
200 – 300	250	-4	16	18	-72	288	
300 – 400	350	-3	9	50	-150	450	
400 – 500	450	-2	4	62	-124	248	
500 – 600	550	-1	1	86	-86	86	
600 – 700	650	0	0	74	0	0	
700 – 800	750	1	1	64	64	64	
800 – 900	850	2	4	52	104	208	
900 – 1000	950	3	9	28	84	252	
1000 – 1100	1050	4	16	16	64	256	
				$N = 450$	$\sum fd = -116$	$\sum fd^2 = 1852$	

$$\text{Mean} = A + \frac{\sum fd}{N} \times C$$

$$= 650 + \frac{-116}{450} \times 100$$

$$= 650 - 0.26 \times 100$$

$$= 650 - 26$$

$$\bar{X} = 624$$

The highest frequency is 86.

The mode lies in 500 – 600.

$$\begin{aligned}
 Z &= L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \\
 L_1 &= 500 \quad f_1 = 86 \quad f_0 = 62 \quad f_2 = 74 \quad I = 100 \\
 &= 500 + \frac{86 - 62}{(2 \times 86) - 62 - 74} \times 100 \\
 &= 500 + \frac{24}{172 - 62 - 74} \times 100 \\
 &= 500 + \frac{24}{36} \times 100
 \end{aligned}$$

$$\text{Mode} = 500 + 66.67 = 566.67$$

$$\begin{aligned}
 \text{S.D.} &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times C \\
 \text{S.D.} &= \sqrt{\frac{1852}{450} - \left(\frac{-116}{450}\right)^2} \times 100 \\
 &= \sqrt{4.12 - 0.07} \times 100 = \sqrt{4.05} \times 100
 \end{aligned}$$

$$\sigma = 2.0125 \times 100 = 201.25$$

$$\begin{aligned}
 \text{Co-efficient of Skewness} &= \frac{\bar{X} - \text{Mode}}{\sigma} \\
 &= \frac{-566.67}{201.25} = \frac{57.33}{201.25}
 \end{aligned}$$

$$S_k = 0.28$$

Illustration 8.23

Calculate mean, dispersion and skewness for the following data.

Diff. in years	Frequency
0–5	249
5–10	607
10–15	501
15–20	271
20–25	121
25–30	53
30–35	14

Solutions

Calculation of mean, S.D. and Skewness

Diff. in years	m	$d =$ $m - 17.5/5$	d^2	f	fd	fd^2
0–5	2.5	-3	9	249	-747	2241
5–10	7.5	-2	4	607	-1214	2428
10–15	12.5	-1	1	501	-501	501
15–20	17.5	0	0	271	0	0
20–25	22.5	1	1	121	121	121
25–30	27.5	2	4	53	106	212
30–35	32.5	3	9	14	42	126
		$N = \sum fd = 1816 - 2193$		$\sum fd^2 = 5629$		

$$\begin{aligned} \text{Mean} &= A \pm \frac{\sum fd}{N} \times C \\ &= 17.5 \frac{-2193}{1816} \times 5 \\ &= 17.5 - 1.21 \times 5 \\ &= 17.5 - 6.05 \end{aligned}$$

$$\bar{X} = 11.45$$

The highest frequency is 607. The mode lies in 5 – 10.

$$\begin{aligned} Z &= L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C \\ &= 5 + \frac{607 - 249}{2 \times 607 - 249 - 501} \times C \\ &= 5 + \frac{358}{1214 - 249 - 501} \times C = 5 + 3.85 \\ Z &= 8.85 \end{aligned}$$

$$\begin{aligned} \text{S.D.} &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C \\ \text{S.D.} &= \sqrt{\frac{5629}{41816} - \left(\frac{-2193}{1816} \right)^2} \times 100 \\ &= \sqrt{3.099 - (-1.21)^2} \times 5 = \sqrt{1.639 \times 5} \end{aligned}$$

$$\sigma = 1.28 \times 5 = 6.4$$

$$\text{Co-efficient of Skewness} = \frac{\bar{X} - \text{Mode}}{\sigma}$$

$$S_k = \frac{11.45 - 8.85}{\sqrt{6.4}} = \frac{2.6}{\sqrt{6.4}}$$

$$S_k = 0.406$$

Illustration 8.24

Calculate the co-efficient of skewness based on mean, median and standard deviation from the following data.

Variables x	Frequency
100–110	6
110–120	18
120–130	36
130–140	54
140–150	66
150–160	52
160–170	31
170–180	7

Solutions

Calculation of Skewness, based on mean, median and S.D.

Variable x	m	d = m – 135/10	d ²	f	fd	fd ²	c.f
100–110	105	-3	9	6	-18	54	6
110–120	115	-2	4	18	-36	72	24
120–130	125	-1	1	36	-36	36	60
130–140	135	0	0	54	0	0	114
140–150	145	1	1	66	66	66	180
150–160	155	2	4	52	104	208	232
160–170	165	3	9	31	93	279	263
170–180	175	4	16	7	28	112	270
<i>N = 270</i>				$\sum fd = \frac{201}{201}$			
$\sum fd^2 = \frac{827}{201}$							

$$\text{Mean} = A \pm \frac{\sum fd}{N} \times C$$

$$= 135 + \frac{201}{270} \times 10$$

$$= 135 + 7.44$$

$$\bar{X} = 142.44$$

Median = size of $N/2$ th item

$$\begin{aligned} &= \text{size of } 270/2\text{th item} \\ &= \text{size of 135th item} \end{aligned}$$

Which lies in the class interval 40–50

$$\begin{aligned} \text{Median} &= L_1 + \frac{N/2 - cf}{f} \times i \\ \text{where, } L_1 &= 140 \quad cf = 180 \quad f = 66 \quad i = 10 \\ &= 140 + \frac{270/2 - 114}{66} \times 10 \\ &= 140 + \frac{135 - 114}{66} \times 10 \\ &= 140 + \frac{21}{66} \times 10 = 140 + 3.18 \\ \text{Median} &= 143.18 \end{aligned}$$

$$\begin{aligned} \text{S.D.} &= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C \\ \text{S.D.} &= \sqrt{\frac{827}{270} - \left(\frac{-201}{270} \right)^2} \times 100 \\ &= \sqrt{3.06 - 0.55} \times 10 = \sqrt{2.51} \times 10 \\ \sigma &= 1.584 \times 10 = 15.84 \end{aligned}$$

$$\begin{aligned} \text{Co-efficient of Skewness} &= \frac{3(\text{Mean} - \text{Median})}{\sigma} \\ &= \frac{3(142.44 - 143.18)}{15.84} \\ &= \frac{3 \times -0.74}{15.84} \\ &= \frac{-2.22}{15.84} \\ S_k &= -0.14 \end{aligned}$$

Illustration 8.25

For the frequency distribution given below, calculate the co-efficient of skewness based on the quartiles.

Class limits	Frequency
10–19	10
20–29	18
30–39	28

Class limits	Frequency
40–49	40
50–59	42
60–69	30
70–79	18
80–89	10
90–99	4

Solutions

Calculation of Median

Class Interval	f	cf
9.5–19.5	10	10
19.5–29.5	18	28
29.5–39.5	28	56
39.5–49.5	40	96
49.5–59.5	42	138
59.5–69.5	30	168
69.5–79.5	18	186
79.5–89.5	10	196
89.5–99.5	4	200
N = 200		

$$\begin{aligned}\text{Median} &= \text{size of } N/2\text{th item} \\ &= \text{size of } 200/2\text{th item} \\ &= 100\text{th item which lies in the class } 49.5–59.5\end{aligned}$$

$$\text{Median} = L_1 + \frac{N/2 - cf}{f} \times i$$

$$\begin{aligned}L_1 &= 49.5 \quad cf = 96 \quad f = 42 \quad i = 10 \quad N/2 = 100 \\ &= 49.5 + \frac{100 - 96}{42} \times 10 \\ &= 49.5 + 0.095 \times 10 \\ &= 49.5 + 0.9\end{aligned}$$

$$\text{Median} = 50.45$$

$$\begin{aligned}Q_1 &= \text{size of } N/4\text{th item} \\ &= \text{size of } 200/4\text{th item} \\ &= 50\text{th item which lies in the class } 29.5 – 39.5\end{aligned}$$

$$Q_1 = L_1 + \frac{N/4 - cf}{f} \times i$$

$$L_1 = 29.5 \quad cf = 28 \quad f = 28 \quad i = 10 \quad N/4 = 50$$

$$\begin{aligned}
 &= 29.5 + \frac{50 - 28}{28} \times 10 \\
 &= 29.5 + 0.79 \times 10 \\
 &= 29.5 + 7.9
 \end{aligned}$$

$$Q_1 = 37.4$$

Q_3 = size of $3N/4$ th item

= size of $3 \times 200/4$ th item

= 150th item which lies in the class $59.5 - 69.5$

$$Q_3 = L_1 + \frac{3N/4 - cf}{f} \times i$$

where, $L_1 = 59.5$ $cf = 138$ $f = 30$ $i = 10$ $3N/4 = 150$

$$\begin{aligned}
 &= 59.5 + \frac{150 - 138}{30} \times 10 \\
 &= 59.5 + 0.4 \times 10 \\
 &= 59.5 + 4
 \end{aligned}$$

$$Q_3 = 63.5$$

Bowley's co-efficient of Skewness:

$$\begin{aligned}
 S_k &= \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1} \\
 &= \frac{63.5 + 37.4 - 2 \times 50.45}{63.5 - 37.4} \\
 &= \frac{100.9 - 100.9}{26.1} = 0/26.1 = 0
 \end{aligned}$$

Illustration 8.26

Find the Standard deviation and co-efficient of Skewness for the following distribution.

Variable	Frequency
0–5	5
5–10	12
10–15	18
15–20	24
20–25	16
25–30	7

Solutions

Calculation of mean, S.D. and Skewness

Variable	M	$d = m - 12.5/5$	d^2	f	fd	fd^2
0–5	2.5	-2	4	5	-10	20
5–10	7.5	-1	1	12	-12	1228
10–15	12.5	0	0	18	-0	0
15–20	17.5	1	1	24	24	24
20–25	22.5	2	4	16	32	64
25–30	27.5	3	9	7	21	63
			$N = 82$	$\sum fd = 55$	$\sum fd^2 = 143$	

Section A

$$\begin{aligned} \text{Mean} &= A \pm \frac{\sum fd}{N} \times C \\ &= 12.5 + \frac{55}{82} \times 5 \\ &= 12.5 + 0.67 \times 5 \\ &= 12.5 + 3.35 \end{aligned}$$

$$\bar{X} = 15.85$$

By Inspection method, mode lies in the group of 15–20

$$\text{Mode } Z = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where, $L_1 = 15$ $f_1 = 24$ $f_0 = 18$ $f_2 = 16$ $i = 5$

$$\begin{aligned} &= 15 + \frac{24 - 18}{2 \times 24 - 18 - 16} \times 5 \\ &= 15 + \frac{6}{14} \times 5 \\ &= 15 + 2.15 \end{aligned}$$

$$Z = 17.15$$

$$\text{S.D.} = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N} \right)^2} \times C$$

$$\text{S.D.} = \sqrt{\frac{143}{82} - \left(\frac{55}{82} \right)^2} \times 100 = \sqrt{1.14 \times 5}$$

$$\sigma = 5.7$$

$$\text{Co-efficient of skewness} = \frac{\bar{X} - \text{Mode}}{\sigma}$$

$$= \frac{15.85 - 17.15}{5.7} = \frac{-1.3}{5.7}$$

Sk = - 0.228

Section B

Median = size of $N/2$ th item
 = size of $100/2$ th item
 = 50th item which lies in the class 61–64

$$Q_1 = L_1 + \frac{N/2 - cf}{f} \times i$$

Where, $L_1 = 61$ $c.f. = 47$ $f = 28$ $i = 3$ $N/2 = 50$

$$= 61 + \frac{50 - 47}{28} \times 3$$

$$= 61 + 0.33$$

Median = 61.33

Q_1 = size of $N/4$ th item
 = size of $100/4$ th item
 = 25th item which lies in the class 58–61

$$Q_1 = L_1 + \frac{N/4 - cf}{f} \times i$$

where, $L_1 = 58$ $c.f. = 22$ $f = 25$ $i = 3$ $N/4 = 25$

$$= 58 + \frac{25 - 22}{25} \times 3$$

$$= 58 + 0.36$$

$Q_1 = 58.36$

Q_3 = size of $3N/4$ th item
 = size of $3 \times 100/4$ th item
 = 75th item which lies in the class 61–64

$$Q_3 = L_1 + \frac{3N/4 - cf}{f} \times i$$

$L_1 = 61$ $c.f. = 47$ $f = 28$ $i = 3$ $3N/4 = 75$

$$= 61 + \frac{75 - 47}{28} \times 3$$

$$= 61 + 3$$

$Q_3 = 64$

Bowley's Co-efficient of Skewness

$$S_k = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

$$\begin{aligned}
 &= \frac{64 + 58.36 - 2 \times 61.33}{64 - 58.36} \\
 &= \frac{-0.3}{5.64} \\
 &= -0.053
 \end{aligned}$$

Illustration 8.27

Find out the co-efficient of Skewness from the following data and show which section is more skewed.

Income (in Rs)	Section A	Section B
55–58	14	22
58–61	19	25
61–64	26	28
64–67	18	14
67–70	13	11

Solutions

Income	Section A		Section B	
	F	cf	f	cf
55–58	14	14	22	22
58–61	19	33	25	47
61–64	26	59	28	75
64–67	18	77	14	89
67–70	13	90	11	100

Section A

Median = size of $N/2$ th item
 = size of $90/2$ th item
 = 45th item which lies in the class 61–64

$$\text{Median} = L_1 + \frac{N/2 - cf}{f} \times i$$

where, $L_1 = 61$ $cf = 33$ $f = 26$ $i = 3$ $N/2 = 45$

$$\begin{aligned}
 &= 61 + \frac{45 - 33}{26} \times 3 \\
 &= 61 + 0.46 \times 3 \\
 &= 61 + 1.38
 \end{aligned}$$

Median = 62.38

Q_1 = size of $N/4$ th item

= size of $90/4$ th item
 = 22.5th item which lies in the class 58–61

$$Q_1 = L_1 + \frac{N/4 - cf}{f} \times i$$

where, $L_1 = 58$ $cf = 14$ $f = 19$ $i = 3$ $N/4 = 22.5$

$$= 58 + \frac{22.5 - 14}{19} \times 3$$

$$= 58 + 1.35$$

$$Q_1 = 59.35$$

Q_3 = size of $3 N/4$ th item

= size of $3 \times 100/4$ th item
 = 75th item which lies in the class 61–64

$$Q_3 = L_1 + \frac{3N/4 - cf}{f} \times i$$

$L_1 = 64$ $cf = 59$ $f = 18$ $i = 3$ $3N/4 = 67.5$

$$= 64 + \frac{67.5 - 59}{18} \times 3$$

$$= 64 + 1.41$$

$$Q_3 = 65.41$$

Bowley's Co-efficient of Skewness

$$S_k = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

$$= \frac{65.41 + 59.35 - 2 \times 62.33}{65.41 - 59.35}$$

$$= \frac{124.76 - 124.76}{6.06}$$

$$= \frac{0}{6.06}$$

$$= 0$$

Section B

Median = size of $N/2$ th item
 = size of $100/2$ th item
 = 50th item which lies in the class 61–64

$$\text{Median} = L_1 + \frac{N/2 - cf}{f} \times i$$

where, $L_1 = 61$ $cf = 47$ $f = 28$ $i = 3$ $N/2 = 50$

$$= 61 + \frac{50 - 47}{28} \times 3$$

$$= 61 + 0.33$$

$$\text{Median} = 61.33$$

Q_1 = size of $N/4$ th item
 = size of $100/4$ th item
 = 25th item which lies in the class 58–61

$$Q_1 = L_1 + \frac{N/4 - cf}{f} \times i$$

where, $L_1 = 58$ $cf = 22$ $f = 25$ $i = 3$ $N/4 = 25$

$$= 58 + \frac{25 - 22}{25} \times 3 \\ = 58 + 0.36$$

$Q_1 = 58.36$
 Q_3 = size of $3N/4$ th item
 = size of $3 \times 100/4$ th item
 = 75th item which lies in the class 61–64

$$Q_3 = L_1 + \frac{3N/4 - cf}{f} \times i$$

where, $L_1 = 61$ $cf = 47$ $f = 28$ $i = 3$ $3N/4 = 75$

$$= 61 + \frac{75 - 47}{28} \times 3 \\ = 61 + 3$$

$$Q_3 = 64$$

Bowley's Co-efficient of Skewness

$$S_k = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} \\ = \frac{64 + 58.36 - 2 \times 61.33}{64 - 58.36} \\ = \frac{-0.3}{5.64} \\ = -0.053$$

Illustration 8.28

Calculate measures of Skewness based on quantities and median from the following data.

Variable	Frequency
10–20	358
20–30	2417
30–40	976
40–50	129

Variable	Frequency
50–60	62
60–70	18
70–80	10

Solutions

Variable	f	c.f.
10–20	358	358
20–30	2417	2775
30–40	976	3751
40–50	129	3880
50–60	62	3942
60–70	18	3960
70–80	10	3970

Median = size of $N/2$ th item
 = size of 3970/2th item
 = 1985th item which lies in the class 20–30

$$\text{Median} = L_1 + \frac{N/2 - cf}{f} \times i$$

where, $L_1 = 20$ $cf = 358$ $f = 2417$ $i = 10$ $N/2 = 1985$

$$= 20 + \frac{1985 - 358}{2417} \times 10 \\ = 20 + 6.73$$

Median = 26.73

Q_1 = size of $N/4$ th item
 = size of 3970/4th item
 = 992.5th item which lies in the class 20–30

$$Q_1 = L_1 + \frac{N/4 - cf}{f} \times i$$

where, $L_1 = 20$ $cf = 358$ $f = 2417$ $i = 10$ $N/4 = 992.5$

$$= 20 + \frac{992.5 - 358}{2417} \times 3 \\ = 20 + 2.63$$

$Q_1 = 22.63$

Q_3 = size of $3N/4$ th item
 = size of $3 \times 3970/4$ th item
 = 2977.5th item which lies in the class 30–40

$$Q_3 = L_1 + \frac{3N/4 - cf}{f} \times i$$

where, $L_1 = 30$ $cf = 2775$ $f = 976$ $i = 10$ $3N/4 = 2977.5$

$$= 30 + \frac{2977.5 - 2775}{976} \times 10$$

$$= 30 + 2.07$$

$$Q_3 = 32.07$$

Bowley's Co-efficient of Skewness

$$S_k = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

where, $Q_3 = 32.07$ $Q_1 = 22.63$ Median = 26.73

$$= \frac{32.07 + 22.63 - 2 \times 26.73}{32.07 - 22.63}$$

$$= \frac{1.24}{9.44}$$

$$S_k = 0.13$$

Illustration 8.29

Calculate the appropriate measures of Skewness from the following income distributions.

Monthly Income (Rs)	Frequency
Upto 1000	9
1001–1500	51
1501–2000	120
2001–2500	240
2501–3000	136
3001–5000	33
5001–7500	9
7501–10000	2
	600

Solutions

Monthly Income (Rs)	Frequency	cf
Upto 1000	9	9
1001–1500	51	60
1501–2000	120	180
2001–2500	240	420
2501–3000	136	556
3001–5000	33	589

Monthly Income (Rs)	Frequency	<i>cf</i>
5001–7500	9	598
7501–10000	2	600
	600	

Median = size of $N/2$ th item
 = size of 600/2th item
 = 300th item which lies in the class 2001–2500

$$\text{Median} = L_1 + \frac{N/2 - cf}{f} \times i$$

$$L_1 = 200 \quad cf = 180 \quad f = 240 \quad i = 500 \quad N/2 = 300$$

$$= 2001 + \frac{300 - 180}{240} \times 500 \\ = 2001 + 250$$

$$\text{Median} = 2251$$

Illustration 8.30

From the information given below, calculate Karl Pearson's co-efficient of Skewness and also Quartile, co-efficient of Skewness.

Measures	Firm A	Firm B
Mean	150	140
Median	142	155
Std. Deviation	30	55
Third Quartile	195	260
First Quartile	62	80

Solutions

Karl Pearson's co-efficient of Skewness

$$\text{Co-efficient of Skewness} = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

Firm A

$$\bar{X} = 150; \text{Median} = 142; \sigma = 30$$

$$= \frac{3(150 - 142)}{30}$$

$$= \frac{24}{30}$$

$$S_k = 0.6$$

Firm B

$$\bar{X} = 140; \text{Median} = 155; \sigma = 55$$

$$= \frac{3(140 - 155)}{55}$$

$$= \frac{-45}{55}$$

$$= -0.82$$

Quartile Co-efficient of Skewness

$$S_k = \frac{Q_3 - Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

Firm A

$$Q_3 = 195 \quad Q_1 = 62 \quad \text{Median} = 142$$

$$\begin{aligned} S_k &= \frac{195 + 62 - 2 \times 142}{195 - 62} \\ &= \frac{-27}{133} \\ S_k &= -0.2 \end{aligned}$$

Firm B

$$Q_3 = 263 \quad Q_1 = 80 \quad \text{Median} = 155$$

$$\begin{aligned} &= \frac{260 + 80 - 2 \times 155}{260 - 80} \\ &= \frac{30}{180} \\ &= 0.17 \end{aligned}$$

Illustration 8.31

Consider the following distribution

	Distribution A	Distribution B
Mean	100	90
Median	90	80
Std. Deviation	10	10

- (a) Distribution A has the same degree of variator as distribution B.
- (b) "Both distributions have the same degree of Skewness." Comment.

Solutions

$$\begin{aligned} \text{(a) C.V. (for distribution A)} &= \frac{\sigma_A}{X_A} \times 100 \\ \sigma &= 10 \quad \bar{X} = 100 \\ \text{C.V.} &= 10/100 \times 100 \\ &= 10\% \end{aligned}$$

$$\text{(b) C.V. (for distribution B)}$$

$$\begin{aligned} \sigma &= 10 \quad \bar{X} = 90 \\ \text{C.V.} &= 10/90 \times 100 \\ &= 11.1\% \end{aligned}$$

False, the co-efficients of variation of distribution A and B are different.
Degree of variation is different.

Co-efficient of Distribution A

Co-efficient of Distribution B

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

$$\begin{aligned}\sigma &= 10, \bar{X} = 100, \text{Median} = 90 \\ &= \frac{3(100 - 90)}{10} = \frac{30}{10} \\ &= 3\end{aligned}$$

$$\begin{aligned}\sigma &= 10, \bar{X} = 100, \text{Median} = 90 \\ &= \frac{3(90 - 80)}{10} = \frac{30}{10} \\ &= 3\end{aligned}$$

True, the distributions A and B have the same degree of Skewness.

Illustration 8.32

In a distribution, Mean = 65, Median = 70, Co-efficient of Skewness = -0.6. Find
 (i) Mode (ii) Co-efficient of Variation. **(C.A.)**

Solutions

$$\begin{aligned}S_k &= \frac{3(\text{Mean} - \text{Median})}{\sigma} \\ 0.6 &= \frac{3(65 - 70)}{\sigma} \\ 0.6\sigma &= \frac{-15}{-0.6} = 25 \\ &= 0.6\sigma = -15 \\ \sigma &= \frac{-15}{0.6} = 25 \\ S_k &= \frac{(\text{Mean} - \text{Median})}{\sigma} \\ -0.6 &= \frac{65 - \text{Mode}}{25} \\ -15 &= 65 - \text{Mode} \\ -0.6 \times 25 &= 65 - \text{Mode} \\ \text{Mode} &= 65 + 15 \\ &= 80 \\ \text{C.V.} &= \frac{\sigma}{X} \times 100 \\ &= 25/65 \times 100 \\ &= 38.46 \%\end{aligned}$$

Illustration 8.33

For a distribution, Bowley's co-efficient of Skewness is -0.36, $Q_1 = 8.6$ and Median = 12.3. What is the Quartile Co-efficient of dispersion?

Solutions

Bowley's co-efficient of Skewness

$$\begin{aligned}
 &= \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1} \\
 S_k = -0.36 \quad Q_1 = 8.6 \quad \text{Median} = 12.3 \\
 -0.36 &= \frac{Q_3 + 8.6 - 2 \times 12.3}{Q_3 - 8.6} \\
 -0.36 Q_3 + 3.096 &= Q_3 + 8.6 - 2 \times 12.3 \\
 -0.36 Q_3 - Q_3 &= 8.6 - 24.6 - 3.096 \\
 -1.36 Q_3 &= 8.6 - 27.696 \\
 Q_3 &= \frac{-19.096}{-1.36} \\
 Q_3 &= 14.04
 \end{aligned}$$

Co-efficient of Quartile Deviation

$$\begin{aligned}
 &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\
 &= \frac{14.04 - 8.6}{14.04 + 8.6} \\
 &= \frac{5.44}{22.64} \\
 &= 0.24
 \end{aligned}$$

Illustration 8.34

For a moderately Skewed distribution, arithmetic mean = 160, mode = 157 and Standard deviation = 50. Find (a) co-efficient of Variation, (b) Pearsonian co-efficient of Skewness and (c) Median.

Solutions

(a) Co-efficient of Variation

$$\begin{aligned}
 \text{C.V.} &= \frac{\sigma}{X} \times 100 \\
 &= 50/160 \times 100 \\
 &= 31.25\%
 \end{aligned}$$

(b) Pearsonian co-efficient of Skewness

$$= \frac{\bar{X} - \text{Mode}}{\sigma}$$

$$\begin{aligned}\bar{X} &= 160, \text{Mode} = 157, \sigma = 50 \\ &= \frac{160 - 157}{50} = \frac{3}{50} \\ &= 0.06\end{aligned}$$

$$\begin{aligned}\text{(c) Median} &= 1/3(2 \text{ mean} + \text{mode}) \\ &= 1/3 (2 \times 160 + 157) \\ &= 1/3 (320 + 157) \\ &= 1/3 \times 477 \\ &= 159\end{aligned}$$

Illustration 8.35

For a group of 10 items, $\sum X = 452$, $\sum X^2 = 24,270$ and mode = 43.7. Find the Pearsonian co-efficient of Skewness.

Solutions

$$\sum X = 452, N = 10;$$

$$\begin{aligned}x &= \frac{\sum X}{N} = \frac{452}{10} = 45.2 \\ \text{Standard deviation} &= \sqrt{\frac{\sum x^2}{N} - \frac{(\sum x)^2}{N}}\end{aligned}$$

$$\begin{aligned}\sum X^2 &= 24,270 \quad N = 10, \quad \sum X = 452 \\ \sigma &= \sqrt{\frac{24270}{10} - \frac{(452)^2}{10}} \\ &= \sqrt{2427 - (45.2)^2} \\ &= \sqrt{2427 - 2043.04} \\ &= \sqrt{383.96} \\ &= 19.59\end{aligned}$$

Karl Pearson's co-efficient of Skewness:
Co-efficient of Skewness

$$= \frac{\bar{X} - \text{Mode}}{\sigma}$$

$$\bar{X} = 45.2, \text{Mode} = 43.7, \sigma = 19.59$$

$$= \frac{45.2 - 43.7}{19.59} = \frac{1.5}{19.59}$$

$$S_k = 0.077$$

Moments**Illustration 8.36**

The first four moments of a distribution about the value 5 are 2, 20, 40 and 50. Obtain, as far as possible, the various characteristics of the distribution on the basis of the information given. Comment upon the nature of the distribution.

Solutions

$$A = 5, \mu_1 = 2, \mu_2 = 20, \mu_3 = 40, \mu = 40$$

1. Mean = $A + \mu_1$
 $= 5 + 2 = 7$
2. Variance = $\mu_2 - (\mu_1)^2$
 $= 20 - (2)^2$
 $= 16$
3. $\mu_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3$
 $= 40 - 3 \times 2 \times 20 + 2 \times (2)^3$
 $= 40 - 120 + 16$
 $= -64$
4. $\mu_4 = \mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4$
 $= 50 - 4 \times 2 \times 40 + 6 \times (2)^2 \times 20 - 3 \times (2)^4$
 $= 50 - 320 + 480 - 48$
 $= 162$

$$\beta_1 = \frac{\mu_3^2}{\mu_2}$$

$$= \frac{(-64)^2}{(16)^3} = \frac{4096}{4096} = 1$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$$= \frac{162}{(16)^2} = \frac{162}{256} = 0.63$$

Since, β_2 is less than 3, the curve is Platykurtic.

Illustration 8.37

Find the first four moments for the following distribution.

Size	Frequency
0–5	1
5–10	2
10–15	6
15–20	7
20–25	4

Solutions

Calculation of first four moments

Size	m	f	$d = m - 12.5/5$	fd	fd^2	fd^3	fd^4			
0–5	2.5	1	-2	-2	4	-8	16			
5–10	7.5	2	-1	-2	2	-2	2			
10–15	12.5	6	0	0	0	0	0			
15–20	17.5	7	1	7	7	7	7			
20–25	22.5	4	2	8	16	32	6.4			
$N = 20$			$\Sigma fd = 11$		$\Sigma fd^2 = 29$		$\Sigma fd^3 = 29$		$\Sigma fd^4 = 89$	

$$\mu_1 = \frac{\sum fd}{N} \times C = \frac{11}{20} \times 5 = 2.75$$

$$\mu_2 = \frac{\sum fd^2}{N} \times C^2 = \frac{29}{20} \times (5)^2 = 36.25$$

$$\mu_3 = \frac{\sum fd^3}{N} \times C^3 = \frac{29}{20} \times (5)^3 = 181.25$$

$$\mu_4 = \frac{\sum fd^4}{N} \times C^4 = \frac{29}{20} \times (5)^4 = 2781.25$$

$$\mu_1 = 0$$

$$\begin{aligned}\mu_2 &= \mu_2 - (\mu_1)^2 \\ &= 36.25 - (2.75)^2 \\ &= 36.25 - 7.5625 \\ &= 28.6875\end{aligned}$$

$$\begin{aligned}\mu_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 \\ &= 181.25 - 3(2.75)(36.25) + 2(2.75)^3\end{aligned}$$

$$\begin{aligned}
 &= 181.25 - 299.0625 + 41.5938 \\
 &= -76.2187. \\
 \mu_4 &= \mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4 \\
 &= 2781.25 - 4(2.75)(181.25) + 6(2.75)^2(36.25) - 3(2.75)^4 \\
 &= 2781.25 - 1993.75 + 1644.8438 - 171.5742 \\
 &= 4426.0938 - 2165.3242 \\
 &= 2260.7696
 \end{aligned}$$

Illustration 8.38

Calculate first four moments from the following data and find out β_1 and β_2 .

x	f
0	5
1	10
2	15
3	20
4	25
5	20
6	15
7	10
8	5

Solutions

Calculation of first four moments

Size	f	fx	d = x - 4	fd	fd²	fd³	fd⁴
0	5	0	-4	-20	80	-320	1280
1	10	10	-3	-30	90	-270	810
2	15	30	-2	-30	60	-120	240
3	20	60	-1	-20	20	-20	20
4	25	100	0	0	0	0	0
5	20	100	1	20	20	20	20
6	15	90	2	30	60	120	240
7	10	70	3	30	90	270	810
8	5	40	4	30	80	320	1280
N =	125	$\sum fx = 500$	$\sum d = 0$	$\sum fd = 0$	$\sum fd^2 = 500$	$\sum fd^3 = 0$	$\sum fd^4 = 4700$

$$x = \frac{\sum fx}{N} = \frac{500}{125} = 4$$

$$\mu_1 = \frac{\sum f(x-\bar{x})}{N} = \frac{0}{125} = 0$$

$$\begin{aligned}\mu_2 &= \frac{\sum fd^2}{N} = \frac{500}{125} = 4 \\ \mu_3 &= \frac{\sum fd^3}{N} = \frac{0}{125} = 0 \\ \mu_4 &= \frac{\sum fd^4}{N} = \frac{4700}{125} = 37.6 \\ \beta_1 &= \frac{\mu_3^2}{\mu_2^3} \\ &= \frac{0^2}{4^3} = \frac{0^2}{4^3} = 0 \\ \beta_2 &= \frac{\mu_4}{\mu_2^2} \\ &= \frac{37.6}{(4)^2} = \frac{37.6}{16} = 2.35\end{aligned}$$

The value of $\beta_2 < 3$. Therefore, the curve is Platykurtic.

Kurtosis

Illustration 8.39

The four moments of a frequency distribution about an arbitrary origin are:

$$\mu_1 = -2; \mu_2 = 14; \mu_3 = -20; \mu_4 = 50$$

Find the values of β_1 and β_2 .

Solutions

$$\begin{aligned}\mu_2 &= \mu_2 - (\mu_1)^2 \\ &= 14 - (-2)^2 = 10 \\ \mu_3 &= \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3 \\ &= -20 - 3(-2)(14) + 2(-2)^3 \\ &= -20 + 84 - 16 \\ &= 84 - 36 = 48 \\ \mu_4 &= \mu_4 - 4\mu_1\mu_3 + 6(\mu_1)^2\mu_2 - 3(\mu_1)^4 \\ &= 50 - 4(-2)(-20) + 6(-2)^2(14) - 3(-2)^4 \\ &= 50 - 160 + 336 - 48 \\ &= 386 - 208 \\ &= 178\end{aligned}$$

304 Business Statistics

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(48)^2}{(10)^3} = \frac{2304}{1000} = 2.304$$

$$\begin{aligned}\beta_2 &= \frac{\mu_4}{\mu_2^2} \\ &= \frac{178}{(10)^2} = \frac{178}{100} = 1.78\end{aligned}$$

Illustration 8.40

The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the Skewness and Kurtosis of the distribution.

Solutions

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = \frac{0.49}{15.625} = + 0.031$$

Therefore, the distribution is Skewed.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = + 3$$

Therefore, the distribution is Symmetrical.

Illustration 8.41

The following data are given to an economist for the purpose of economic analysis. The data refer to the length of the life of a sample of Good Year Tyres. Is the distribution Platykurtic?

$$N = 100; \sum f dx = 50; \sum f dx^2 = 1967.2; \sum f dx^3 = 2925.8; \sum f dx^4 = 86650.2$$

Solutions

$$\mu_1 = \frac{\sum f dx}{N} = \frac{50}{100} = 0.5$$

$$\mu_2 = \frac{\sum f dx^2}{N} = \frac{1967.2}{100} = 19.672$$

$$\mu_3 = \frac{\sum f dx^3}{N} = \frac{2925.8}{100} = 29.258$$

$$\mu_4 = \frac{\sum f dx^4}{N} = \frac{86650.2}{100} = 866.502$$

$$\begin{aligned}
 \mu_1 &= 0 \\
 \mu_2 &= \mu_2 - (\mu_1)^2 \\
 &= 19.672 - (0.5)^2 \\
 &= 19.672 - 0.25 \\
 &= 19.422 \\
 \mu_3 &= \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3 \\
 &= 29.258 - 3(0.5)(19.672) + 2(0.5)^3 \\
 &= 29.258 - 29.508 + 0.25 \\
 &= 0 \\
 \mu_4 &= \mu_4 - 4\mu_1\mu_3 + 6(\mu_1)^2\mu_2 - 3(\mu_1)^4 \\
 &= 866.502 - 4(0.5)(29.258) + 6(0.5)^2(19.672) - 3(0.5)^4 \\
 &= 866.502 - 58.516 + 29.508 - 0.1875 \\
 &= 837.3065 \\
 \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{837.3065}{(19.422)^2} \\
 &= \frac{837.3065}{377.214} \\
 &= 2.219
 \end{aligned}$$

Since β_2 is less than 3, the curve is Platykurtic.

SUMMARY

Skewness

Skewness means asymmetrical distribution.

Positive skewness

When the value of mean is more than the mode.

Negative skewness

When the value of mode is greater than mean.

Measures of skewness

- Absolute measures of skewness
- Relative measures of skewness

Absolute measures of skewness

It may be measured by taking the difference between mode and mean.

Relative measures of skewness

- Karl Pearson's coefficient of skewness
- Bowley's coefficient of skewness
- Kelly's coefficient of skewness
- Measures of skewness based on moments

Karl Pearson's coefficient of skewness

Difference of means and mode divided by standard deviation.

Bowley's coefficient of skewness

It is based upon the quartiles of the distribution.

Kelly's coefficient of skewness

It is based upon the percentiles or deciles of the distribution. It covers the beginning and end of the distribution.

Moments (μ)

It is related with the deviation of distribution from the arithmetical mean.

Kurtosis

Greek word, which means bulginess. It refers to the measure of the peakness or flatness of the distribution of data.

Leptokurtic

If the curve is more peaked than the normal curve.

Platykurtic

If the curve is more flattopped than the normal curve.

Mesokurtic

It is the normal curve.

FORMULAE

Absolute skewness (Karl Pearson)

$$\text{Skp} = \bar{x}; 2$$

$$\text{Bowley} = \text{SKB} = Q_3 + Q_1 - 2 \text{ Med.}$$

\bar{x} = Mean

Z = Mode

Q_3 = Upper Quartile

Q_1 = Lower Quartile

Med – Median

Co-efficient of skewness

1. Karl Pearson co-efficient of skewness

$$(a) Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

Sk_p = Karl Pearson's co-efficient of Skewness

If mode is ill defined

$$(b) Sk_p = \frac{3(\bar{x} - \text{med})}{\sigma}$$

\bar{X} = Mean

σ = standard deviation

2. Bowly's co-efficient of Skewness

$$SK_B = \frac{(Q_3 + Q_1 - 2 \text{ Med})}{Q_3 + Q_1}$$

SK_B = Bowley's co-efficient of Skewness.

3. Kelly's co-efficient of Skewness

$$SK_k = \frac{P_{10} + P_{90} - 2 \text{ med}}{P_{90} + P_{10}}$$

$$SK_k = \frac{D_1 + D_9 - 2 \text{ med}}{D_9 + D_1}$$

SK_k = Kelly's co-efficient of Skewness

P_{10} and P_{90} = Percentiles

D_1 and D_9 = Deciles

MOMENTS

Moment about mean – individual series

$$M_1 = \frac{\sum(x - \bar{x})}{N} \quad \text{or} \quad \frac{\Sigma d}{N} = 0$$

$$M_2 = \frac{\sum d^2}{N} = \sigma^2$$

$$M_3 = \frac{\sum d^3}{N}$$

$$M_4 = \frac{\sum d^4}{N}$$

Frequency Distribution

$$M_1 = \frac{\sum f(x - \bar{x})}{N} \quad \text{or} \quad \frac{\sum fd}{N} = 0$$

$$M_2 = \frac{\sum fd^2}{N} = \sigma^2$$

$$M_3 = \frac{\sum fd^3}{N}$$

$$M_4 = \frac{\sum fd^4}{N}$$

Moments about arbitrary origin individual series

$$M'_1 = \frac{\sum(x - A)}{N} \quad \text{or} \quad \frac{\Sigma d}{N}$$

$$M'_2 = \frac{\sum d^2}{N}$$

$$M'_3 = \frac{\sum d^3}{N}$$

$$M'_4 = \frac{\sum d^4}{N}$$

Frequency Distribution

$$M'_1 = \frac{2f(x-A)}{N} \quad \text{or} \quad \frac{\sum fd}{N} \quad \text{or} \quad \frac{\sum fd^1}{N} \times C$$

$$M'_2 = \frac{\sum fd^2}{N} \quad \text{or} \quad \frac{\sum fd^2}{N} \times C$$

$$M'_3 = \frac{\sum fd^3}{N} \quad \text{or} \quad \frac{\sum fd^3}{N} \times C$$

$$M'_4 = \frac{\sum fd^4}{N} \quad \text{or} \quad \frac{\sum fd^4}{N} \times C$$

Converting moments about \bar{x}

$$M_1 = \mu'_1 - \mu'_1 = 0$$

$$M_2 = \mu'_2 - (\mu'_1)^2$$

$$M_3 = \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3$$

$$M_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2(\mu'_1)^2 - 3(\mu'_1)^4$$

Kurtosis

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}; \beta_2 = \frac{\mu_4^2}{\mu_2^2}$$

EXERCISES

(a) Choose the best option.

1. Skewness means
 - (a) Symmetrical distribution
 - (b) Asymmetrical distribution
 - (c) Measures of central tendency.
2. In a symmetrical distribution mean, median and mode are
 - (a) Equal
 - (b) Not Equal
 - (c) Greater than one

3. When the value of mean is more than the mode it is called
 - (a) Positive skewness
 - (b) Negative skewness
 - (c) kurtosis
4. When the value of mode is greater than mean, it is called
 - (a) Positive skewness
 - (b) Negative skewness
 - (c) kurtosis
5. The extent of the peakness of the distribution from the normal distribution can be measured through
 - (a) kurtosis (b) skewness (c) Moments
6. If the curve is more peaked than the normal curve, than it is
 - (a) Leptokurtic (b) Platykurtic (c) Mesokurtic
7. If the curve is more flattopped than the normal curve then it is
 - (a) Leptokurtic (b) Platykurtic (c) Mesokurtic
8. The normal curve is called
 - (a) Platykurtic (b) Leptokurtic (c) Mesokurtic
9. Bowley's coefficient of skewness lies between
 - (a) -1 and 1 (b) 1 and -1 (c) 0 and 1
10. $\beta_1 = 0$ implies the distribution is
 - (a) symmetrical (b) asymmetrical (c) Normal
11. If $\beta_1 > 0$, the distribution is
 - (a) Positively skewed (b) Negatively skewed (c) Normal
12. If $\beta_1 < 0$, the distribution is
 - (a) Positively skewed (b) Negatively skewed (c) Normal
13. Relative measure of skewness based on mean, standard deviation and mode is
 - (a) $\frac{m - m_0}{S.D.}$
 - (b) $\frac{m_0 - m}{S.D.}$
 - (c) $\frac{m_1 - m}{S.D.}$
14. Relative measure of kurtosis is
 - (a) β_2
 - (b) β
 - (c) β_1
15. Relative measure of skewness in terms of moments is
 - (a) β_1
 - (b) β_2
 - (c) β_3

Answers.

- | | | | | |
|--------|--------|--------|--------|--------|
| (1) b | (2) a | (3) a | (4) b | (5) a |
| (6) a | (7) b | (8) c | (9) a | (10) a |
| (11) a | (12) b | (13) a | (14) a | (15) a |

(b) Theoretical Questions

1. What is Skewness?
2. Explain the main types of Skewed Curves.
3. What do you mean by Kurtosis?
4. Distinguish between Skewness and Kurtosis.
5. What is Skewness? How does it differ for dispersion?

(B.Com., MKU, MSU, BDU)

6. Define “moments”. Explain the relation between the moments almost the mean in terms of moments about any arbitrary point.

(B.Com., MKU, MSU, BU)

7. Explain briefly the different methods of measuring Skewness.
8. Define and discuss the Quartiles of a distribution. How are they used for measuring variations and Skewness?
9. Briefly mention the tests which can be applied to determine the presence of skewness.
10. What are the moments? Explain the procedure of calculating the first four moments above the mean.
11. Prove that the moments co-efficient of Kurtosis is not affected by the change of scale.
12. Differentiate between Bowley’s measure and Karl Pearson’s measure of Skewness.
13. Explain the briefly the different methods of measuring Skewness.

(B.Com., MKU, MSU, BDU)

14. Calculate Karl Pearson’s co-efficient of Skewness for the following data of marks obtained by five students.

12	18	35	22	18
----	----	----	----	----

Answer. 0.3896

15. Find out the co-efficient of Skewness from the data given below:

Size	3	4	5	6	7	8	9	10
Frequency	7	10	14	35	102	136	43	8

Answer. -0.49

(B.Com., MKU, MSU, BU)

16. Calculate Karl Pearson’s co-efficient of Skewness from the data given below:

Income (Rs)	400–500	500–600	600–700	700–800	800–900
-------------	---------	---------	---------	---------	---------

No. of employees	8	16	20	17	3
------------------	---	----	----	----	---

Answer. -0.195

(B.Com., CHU, MSU, BU)

17. Assume that a firm has selected a random sample of 100 from its production line and has obtained the data shown in table below:

Class interval	Frequency
130–134	3
135–139	12
140–144	21
145–149	28
150–154	19
155–159	12
160–164	5

Compute the following

- (a) the arithmetic mean
- (b) the standard deviation
- (c) Karl Pearson's co-efficient of Skewness.

Answer. Mean 147.2, S.D. 7.21, $S_k = -0.071$

(B.Com., MKU, MSU, BU)

18. Find out Bowley's co-efficient of Skewness from the following data.

Marks in Statistics	15	25	35	45	55	65	75	85
not more than								
No. of students	3	10	40	55	65	72	75	80

Answer. – 30.73

(B.Com., MKU, MSU, BU)

19. Calculate co-efficient of Skewness by Quartile and median from the following data.

x	0–10	10–20	20–30	30–40	40–50	50–60
y	7	8	10	5	9	6

Answer. 0.15

(B.Com., MKU, MSU, CHU)

20. Compute co-efficient of Skewness by Kelly's method from the following data.

x	7	14	21	28	35	42
y	3	6	9	5	4	3

Answer. 0.07

21. Find out the co-efficient of Skewness through percentiles method for the following data.

Marks	35–45	45–55	55–65	65–75	75–85	85–95
No. of students	2	4	18	16	7	3

Answer. 0.11

22. Consider the following distribution.

	Distribution A	Distribution B
Mean	100	90
Median	90	80
Std. deviation	10	10

Show that

- (a) Distribution A has the same degree of variation as distribution B .
- (b) Both Distribution have the same degree of Skewness.

Answer. (a) The co-efficient of variation of distribution A and B are different as $C.V_A = 10\%$ and $C.V_B = 11.1\%$. (b) Co-ef. $Sk_A = 3$, Co-eff. $Sk_B = 3$

(B.Com., MSU, BU, CHU)

23. In a certain distribution, the following results were obtained.

$$\bar{X} = 45, \text{ median} = 48, \text{ Co-eff. Skewness} = 0.4.$$

The person who gave you the data failed to give the value of standard deviation and you are required to estimate it units the help of the available information.

Answer. S.D. = 22.5

24. Find the first four moments for the set of the number 2, 4, 6, 8.

Answer. 0, 5, 0, 41

25. Find the first four moments from the following data

X	0	1	2	3	4	5	6	7	8
F	5	10	15	20	25	20	15	10	5

(Answer. 0, 4, 0, 37.6, $\beta_1 = 0$, $\beta_2 = 2.35$. (β_1 is zero, so the distribution is symmetrical. β_2 is less than 3; So, the distribution is Platykurtic.)

26. Compute the four moments from the data given below.

X	0–10	10–20	20–30	30–40	40–50	50–60	60–70
F	5	3	2	5	4	4	2

Answer. 0, 384, -528, 260544

(B.Com., MKU, MSU)

27. Calculate Kurtosis from the following data.

9	18	7	11	4	6	8
---	----	---	----	---	---	---

Answer. B_2 is 3.33, Since $B_2 > 3$, it is Leptokurtic

28. Calculate Kurtosis from the following data.

X	0–10	10–20	20–30	30–40	40–50	50–60
F	11	7	9	20	2	4

Answer. B_2 is 4.4, Since $B_2 > 3$, it is called Leptokurtic

29. From the following data, calculate the four moments

- (a) above the value 15
- (b) about mean
- (c) Skewness based on moments
- (d) Kurtosis.

X	0–10	10–20	20–30	30–40
F	1	3	4	2

Answer. $M_1 = 7$, $M'_2 = 130$, $M'_3 = 1900$, $M'_4 = 37,000$, $M_1 = 0$, $M_2 = 81$, $M_3 = -144$, $M_4 = 14817$ moments based on mean: $M_1 = 0$, $M_2 = 81$, $M_3 = 7 = -144$, $M_4 = 14812$, $B_1 = 0.04$, $B_2 = 2.26$

30. For a moderately Skewed distribution, arithmetic mean = 160, mode = 157, and Standard deviation = 50. Find (i) Co-efficient of Variation (ii) Pearson's Co-efficient of Skewness and (iii) median.

Answer. C.V. = 31.25%, Sk 0.6, median = 159

(B.Com., MKU, BDU, CHU)

31. In a distribution, mean = 67, median = 62, Standard deviation = 6, find Skewness.

Answer. 2.5

32. In a moderately symmetrical distribution, the mode and mean are 32.1 and 35.4 respectively. Calculate the median.

Answer. 34.3

33. Pearson's co-efficient of Skewness for a distribution is 0.5 and the co-efficient of variation is 40%. Its mode is 80. Find the mean and median of the distribution.

Answer. mean = 100, median = 93.33

(B.Com., MKU, MSU, CHU)

9

CHAPTER

CORRELATION ANALYSIS

9.1 MEANING

It measures the degree of relationship between two or more variables. The degree of correlation states about the extent of the closeness of two sets of variables. For example, there exists some relationship between price and demand, heights of husbands and wives, heights of fathers and daughters, wages and price index, yield and rainfall, sales and advertisement, results and teaching. Thus, correlation is a statistical tool which helps us in analysing the co-variation of two or more variables.

9.2 DEFINITIONS

The important definitions are given below:

Correlation analysis attempts to determine the degree of relationship between variables. —**Ya Lun Chov**

Correlation analysis deals with the association between two or more variables. —**Simpson & Kafka**

Correlation is an analysis of the co-variation between two or more variables. —**A.M. Tuttle**

When the relationship is of a Quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation. —**Croxton and Cowdon**

9.3 UTILITY OF MEASURING CORRELATION (USEFULNESS)

The study of correlation analysis is useful both in physical and social sciences. The following are the usefulness of correlation in the field of business and economics.

- (i) Correlation analysis helps in quantifying precisely the degree and direction of such relationships between variables.

- (ii) It is very useful for making forecasts in economic and business activities. In business, it enables us to estimate cost, sales on the basis of observation recorded on functionally related variables with these dependent variables.
- (iii) It contributes to the understanding of economic behaviour, aids in locating the critically important variables, on which other depend, may reveal to the economist.
- (iv) The effect of correlation is to reduce the range of uncertainty of our prediction. The prediction based on correlation analysis will be more reliable and near to reality.
- (v) Correlation is a powerful statistical tool that provides quantitative expressions of the manner or extent of which events are related mathematically.

9.4 CORRELATION AND CAUSATION

While correlation is a measure of the degree of relationship between two or more variables, it gives no indication of the kind of cause and effect relationships that exist among the variables. A high degree of correlation in two variables implies that there must be a reason for such close relationship, but the cause and effect relationship can be revealed specifically only by other knowledge of the factors involved being brought to bear on the situation factors involved.

- (i) One variable may act upon the other.
- (ii) Difficulty in assigning the cause and the effect variables.
- (iii) Chance of coincidence.
- (iv) Influence of a third variable.

9.5 TYPES OF CORRELATION

There are three important types of correlation. They are:

1. Positive and Negative correlation
2. Simple, Partial and Multiple correlation
3. Linear and Non-linear correlation.

9.5.1 Positive and Negative Correlation

Correlation is classified according to the direction of change in the two variables. In this regard, the correlation may either be positive or negative.

Positive correlation refers to the change (movement) of variables in the same direction. Both the variables are increased or decreased in the same direction, it is called positive correlation. It is otherwise called as direct correlation. For example, a positive correlation exists between ages of husband and wife, height and weight of a group of individuals, increase in rainfall and production of paddy, increase in the offer and sales.

Negative correlation refers to the change (movement) of variables in the opposite direction. In other words, an increase (decrease) in the value of one

316 Business Statistics

variable is followed by a decrease (increase) in the value of the other is said to be negative correlation. It is otherwise called inverse correlation. For example, a negative correlation exists between price and demand, yield of crop and price.

The following examples illustrate the concept of positive correlation and negative correlation:

Positive Correlation

X	5	7	9	11	16	20	28	X	46	37	28	19	10
Y	26	28	33	36	47	48	52	Y	28	23	16	11	8

Negative Correlation

X	14	17	23	35	46	X	16	13	10	9	7
Y	12	10	9	8	4	Y	13	16	19	24	27

9.5.2 Simple, Partial and Multiple Correlation

Simple correlation means study of only two variables. It is the association of only two variables. For example, height and weight of B.Com-II students of a college.

In partial correlation, we identify two or more variables, not consider only two variables relationship, keeping other variables constant. For example, when we study the relationship between yield of paddy per acre and amount of rainfall, fertility of land, types of fertilizers used. Here, we want to study the relationship between yield of paddy and types of fertilizers used, keeping all other variables constant.

In multiple correlation, three or more variables are studied simultaneously. For example, when we study the relationship between the results of a student and method of teaching, nature of the subject, students' skill, types of institution, it is a problem of multiple correlation. Here, we are calculating correlation for all the identified variables simultaneously.

9.5.3 Linear and Non-linear Correlation

If the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable then the correlation is said to be linear. For example,

X	5	10	15	20	25
Y	80	160	240	320	400

On the other hand, if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable is said to be non-linear or curvilinear. For example,

X	5	10	15	20	25
Y	8	10	11	9	14

9.6 METHODS OF STUDYING CORRELATION

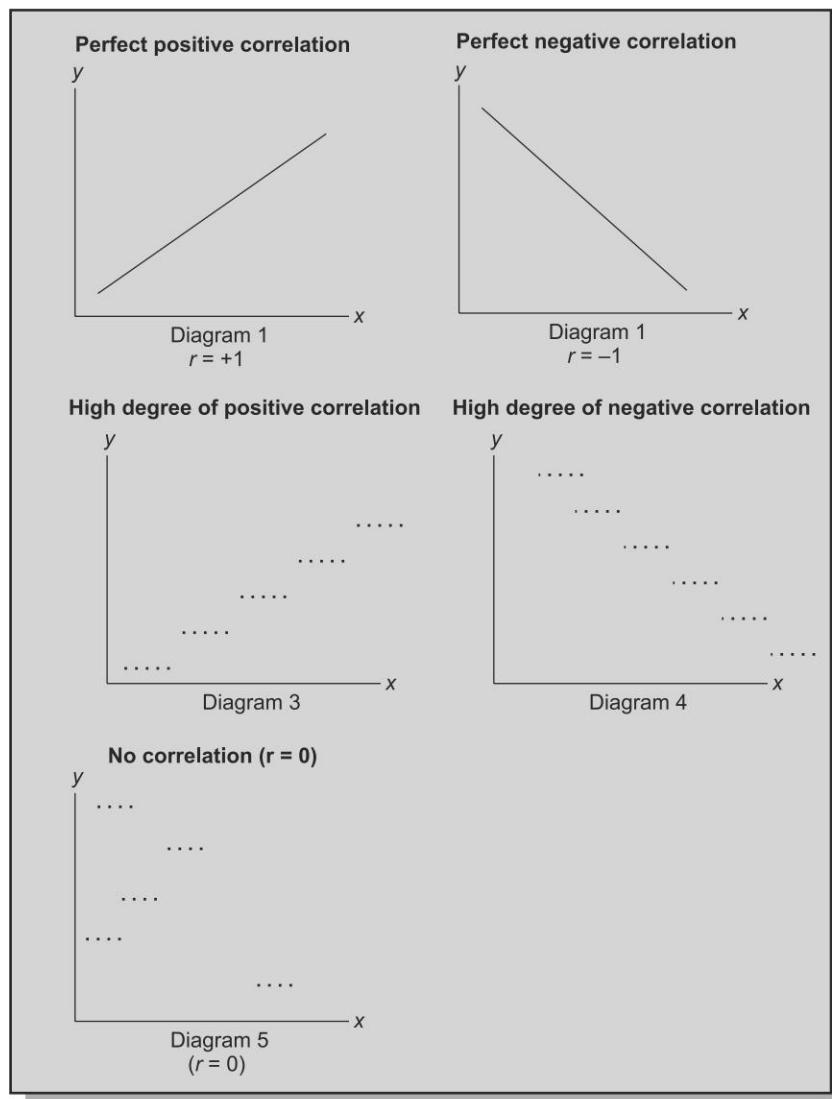
The following are the important methods of finding out the correlation:

- (a) Scatter Diagram

- (b) Graphic method
- (c) Karl Pearson's co-efficient of correlation
- (d) Rank co-efficient of correlation or Rank correlation co-efficient
- (e) Concurrent deviation method

9.6.1 Scatter Diagram

It is simple and attractive method of diagrammatic representation. In this method, the given data are plotted on a graph sheet in the form of dots. The x variables are plotted on the horizontal axis and y variables on the vertical axis. Now we can know the scatter or concentration of the various points. This will show the type of correlation.



Merits

1. It is a simple method of finding out the nature of correlation between two variables.
2. It does not require any mathematical calculations.
3. It gives a quick idea about the nature of correlation (positive/negative).
4. It is not influenced by extreme items.

Demerits

1. It gives only a rough idea about the correlation.
2. The correct value of correlation is not possible under this method.

9.6.2 Graphic Method

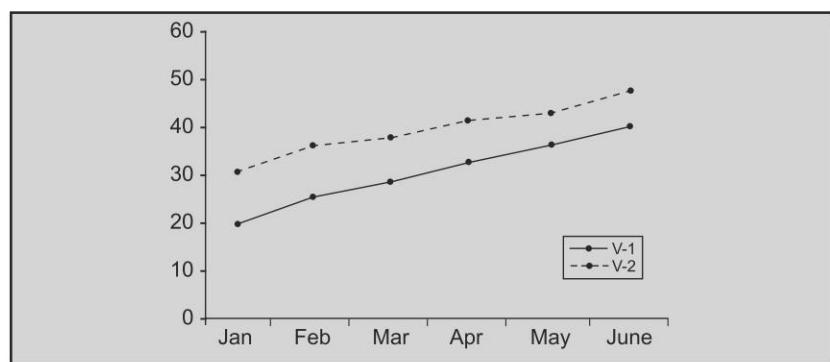
Under this method, the individual values of the two variables are plotted on the graph paper. Thus, we obtain two curves, one for x variables and another for y variables. The curves reveal the direction and closeness of the two curves and also reveal that the variables are related or not, if the both curves move in the same direction, is parallel, upward or downward, the correlation is said to be positive. On the other hand, if they move in opposite directions, then the correlation is said to be negative. The following example shall illustrate this method.

Illustration 9.1

Draw a correlation graph from the following data:

Period	Jan.	Feb.	Mar.	Apr.	May	June
Advt.(Rs. '000)	20	25	28	32	36	40
Sales(Rs. '0000)	30	36	37	41	43	48

Solutions



This method is used in the case of time series analysis. It does not reveal the extent to which the variables are related.

9.6.3 Karl Pearson's Co-efficient of Correlation

Karl Pearson (1867–1936), the British biometrician suggested this method. It is popularly known as Pearson's co-efficient of correlation. It is a mathematical method for measuring the magnitude of linear relationship between two variables. The co-efficient of correlation can be calculated as follows:

(a) Arithmetic Mean Method

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

where, $x = x - \bar{x}$, $y = y - \bar{y}$

$X = x$ series, $y = y$ series, $r = \text{co-efficient of correlation}$.

Steps

- Calculate Arithmetic mean for x series and y series.
- Find out the deviations (x) for x series from the mean is $x = (x - \bar{x})$.
- Find out the deviations (y) for y series from the mean is $y = (y - \bar{y})$.
- Square these deviations for x series and y series and obtain Σx^2 and Σy^2 .
- Multiply the deviations of x and deviation of y series and obtain Σxy .

Substitute the value of Σx^2 and Σy^2 and Σxy in the formula.

Illustration 9.2

Calculate Karl Pearson's co-efficient of correlation from the following data:

Roll No. of students	101	102	103	104	105
Marks in Statistics	56	76	93	87	88
Marks in Costing	46	55	74	68	57

Solutions

Let the marks in Statistics be denoted by x and marks in Costing by y .

Roll No.	x	$x(x - x)$ $x - 80$	x^2	y	$y(y - y)$ $y - 60$	y^2	xy
101	56	-24	576	46	-14	196	336
102	76	-4	16	55	-5	25	20
103	93	13	169	74	14	196	182
104	87	7	49	68	8	64	56
105	88	8	640	57	-3	9	-24
$\Sigma x =$		$\Sigma x =$	$\Sigma x^2 =$	$\Sigma y =$	$\Sigma y =$	$\Sigma y^2 =$	$\Sigma xy =$
		400	874	300	0	490	570

$$x = \frac{\sum x}{N} = \frac{400}{5} = 80$$

$$y = \frac{\sum y}{N} = \frac{300}{5} = 60$$

$$\sum x^2 = 874$$

$$\sum y^2 = 490$$

$$\sum xy = 570$$

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

$$= \frac{570}{\sqrt{874} \cdot \sqrt{490}}$$

$$= \frac{570}{29.563 \times 22.135}$$

$$= \frac{570}{654.38} \quad r = 0.87$$

(b) Assumed Mean Method

$$r = \frac{N \sum dxdy - \sum dx \cdot \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \sqrt{N \sum dy^2 - (\sum dy)^2}}$$

where, dx = deviation of x series from an assumed mean ($x - \bar{x}$)

dy = deviation of y series from an assumed mean ($y - \bar{y}$)

$\sum dx$ = sum of deviation of x series

$\sum dy$ = sum of deviation of y series

$\sum dxdy$ = sum of the product of the deviations of x and y series.

$\sum dx^2$ = sum of the squares of the deviation of x series.

$\sum dy^2$ = sum of the squares of the deviation of y series.

N = number of paired observations.

Steps

1. Take the deviations of x series from an assumed mean and denote these deviations as dx and obtain Σdx .
2. Take the deviations of y series from an assumed mean and denote these deviations as dy and obtain Σdy .

3. Square dx and dy and obtain $\sum dx^2$ and $\sum dy^2$.
4. Multiply dx and dy and obtain $\sum dxdy$.
5. Substitute the values in the formula,

i.e.,
$$r = \frac{N \sum dxdy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \sqrt{N \sum dy^2 - (\sum dy)^2}}$$

Illustration 9.3

Find out the co-efficient of correlation for the following data:

Height of Father (cm)	160	162	163	164	165	167	168
Height of daughter (cm)	150	152	148	149	154	160	165

Solutions

x	$dx = -164$	dx^2	y	$dy = -154$	dy^2	$dxdy$
160	-4	16	150	-4	16	16
162	-2	4	152	-2	4	4
163	-1	1	148	-6	36	6
164	0	0	149	-5	25	0
165	+1	1	154	0	0	0
167	+3	9	160	+6	36	18
168	+4	16	165	+9	81	36
$\Sigma dx = 1$ $\Sigma dx^2 = 47$			$\Sigma dy = -2$		$\Sigma dy^2 = 198$	$\Sigma dxdy = 1950$

$$\begin{aligned}
 r &= \frac{N \sum dxdy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \sqrt{N \sum dy^2 - (\sum dy)^2}} \\
 &= \frac{7 \times 80 - (1 \times -1)}{\sqrt{7 \times 47 - (1)^2} \cdot \sqrt{7 \times 198 - (-10)^2}} \\
 &= \frac{560 - (-1)}{\sqrt{299 - 1} \cdot \sqrt{1386 - 1}} = \frac{561}{\sqrt{298} \cdot \sqrt{1386 - 1}} \\
 &= \frac{561}{17.26 \times 37.22} \\
 &= \frac{561}{64642} = +0.87 \\
 r &= +0.87
 \end{aligned}$$

Illustration 9.4

Calculate the co-efficient of correlation between age of cars and annual maintenance cost and comment.

Age of Cars in Year	Annual Maintenance cost in Rs.
2	1600
4	1500
6	1800
7	1900
8	1700
10	2100
12	2000

Solutions

X	$dx = x - A$ $x - 7$	dx^2	y	$dy = y - A =$ $y - 1900/100$	dy^2	$dxdy$
2	-5	25	1600	-3	9	15
4	-3	9	1500	-4	16	12
6	-1	1	1800	-1	1	1
7	0	0	1900	0	0	0
8	+1	1	1700	-2	4	-2
10	+3	9	2100	+2	4	6
12	+5	25	2000	+1	1	5
$\Sigma dx = 0$		$\Sigma dx^2 = 70$		$\Sigma dy = -7$	$\Sigma dy^2 = 35$	$\Sigma dxdy = 7$

$$\begin{aligned}
 r &= \frac{N \sum dxdy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \sqrt{N \sum dy^2 - (\sum dy)^2}} \\
 &= \frac{7 \times 37 - (0 \times -7)}{\sqrt{7 \times 70 - (0)^2} \cdot \sqrt{7 \times 35 - (-7)^2}} \\
 &= \frac{259 - 0}{\sqrt{490} \cdot \sqrt{245 - 49}} = \frac{259}{22.14 \times 14} \\
 &= \frac{259}{309.96} = +0.84
 \end{aligned}$$

$$r = +0.84$$

Hence, there is a high degree positive correlation between age of cars and annual maintenance cost.

Direct Method Under this method, correlation co-efficient can be calculated without taking actual mean and assumed mean. The formula is

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

Steps

1. Find Σx and Σy .
2. Take the Square of variables x and obtain Σx^2 .
3. Take the Square of variables y and obtain Σy^2 .
4. Calculate the product of x corresponding x variables and y variables is xy and obtain Σxy .
5. Substitute the values in the formula.

Illustration 9.5

Find out the correlation co-efficient between height of mothers and height of daughters:

Height of mother (in cm)	150	152	156	157	160
Height of daughter (in cm)	148	145	160	156	150

Solutions

Calculation of correlation co-efficient

x	x²	y	y²	xy
150	22500	148	21904	22200
152	23104	145	21025	22040
156	24336	160	25600	24960
157	24649	156	24336	24492
160	25600	150	22500	24000
$\Sigma x = 775$	$\Sigma x^2 = 120189$	$\Sigma y = 759$	$\Sigma y^2 = 112365$	$\Sigma xy = 117692$

$$\begin{aligned}
 r &= \frac{N \sum xy - (\sum x) \times (\sum y)}{\sqrt{N \sum x^2 - (\sum x)^2} \sqrt{N \sum y^2 - (\sum y)^2}} \\
 &= \frac{5 \times 1179692 - (775 \times 759)}{\sqrt{7 \times 120189 - (775)^2} \sqrt{7 \times 115365 - (759)^2}} \\
 &= \frac{588460 - 588225}{\sqrt{600945 - 600625} \sqrt{576825 - 576081}} \\
 &= \frac{235}{\sqrt{320} \sqrt{744}} = \frac{235}{17.89 \times 27.28}
 \end{aligned}$$

$$= \frac{235}{488.04} = 0.48$$

$$r = 0.48$$

Change of scale in the calculation of r If the scale of x or y series is changed, the value of r remains unaffected. Thus, if the value of x series are 2000, 4000, 6000 and if they are divided by a common factor 1000 and we take the values as 2, 4, 6 then the value of r would not be affected. Similarly, if the values of 4 series are 0.2, 0.3, 0.4 they can be multiplied by 10 and we take the values as 2, 3 and 4 without affecting our result.

Illustration 9.6

Calculate co-efficient of correlation from the following data:

X	0.1	0.2	0.3	0.4	0.5	0.6	0.7
y	15000	25000	30000	40000	50000	55000	65000

Solutions

We can multiply x series by 10 and divide y series by 5000, so that the values of x and y would be

Calculation of co-efficient of correlation

x	$dx = x - 4$	dx^2	y	$dy = y - 8$	dy^2	$dxdy$
1	-3	9	3	-5	25	15
2	-2	4	5	-3	9	6
3	-1	1	6	-2	4	2
4	0	0	8	0	0	0
5	+1	1	10	2	4	2
6	+2	4	11	3	9	6
7	+3	9	13	5	25	15
$\Sigma dx = 0$		$\Sigma dx^2 = 28$	$\Sigma dy = 10$		$\Sigma dy^2 = 76$	$\Sigma dxdy = 46$

$$r = \frac{N \sum dxdy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \cdot \sqrt{N \sum dy^2 - (\sum dy)^2}}$$

$$= \frac{7 \times 46 - 0 \times 0}{\sqrt{7 \times 28 - (0)^2} \sqrt{7 \times 76 - (0)^2}}$$

$$= \frac{252}{\sqrt{196} \cdot \sqrt{532}}$$

$$= \frac{252}{14 \times 23.07}$$

$$= \frac{252}{322.98} = + 0.78$$

$$r = + 0.78$$

If the original values of x and y were used, the result would be the same.

(d) Calculation of Co-efficient of Correlation in Grouped Data If the values of two variables are grouped and the frequencies of different groups are given, double tabulation is necessary for finding out the co-efficient of correlation. The class intervals for y are in the column headings and for x , in the stubs. The formula is:

$$r = \frac{N \sum f dx dy - \sum f dx \cdot \sum f dy}{\sqrt{N \sum f dx^2 - (\sum f dx)^2} \sqrt{N \sum f dy^2 - (\sum f dy)^2}}$$

This formula is the same as the previous formula, which was discussed for assumed mean. The only difference is that the deviations are multiplied by the respective frequencies.

Steps

1. Take the step deviations of x and y series from assumed means and denote by dx and dy .
2. Multiply dx and dy and the respective frequency of each cell and write the figure obtained in right hand upper corner for each cell.
3. Add all the values of $f dx dy$ and obtain $\sum f dx dy$.
4. Multiply the frequencies of the variable x by the deviation of x and obtain $\sum f dx$.
5. Take the squares of the deviations of variable x and multiply by the respective frequencies and obtain $\sum f dx^2$.
6. Multiply the frequencies of the variable y by the deviations of y and obtain $\sum f dy$.
7. Take squares of the deviation of variable y and multiply by the respective frequencies and obtain $\sum f dy^2$.
8. Substitute the values in the formula and obtain the value of r .

Illustration 9.7

Y	X				
	200–300	300–400	400–500	500–600	600–700
20 – 25	–	–	–	5	5
25 – 30	–	6	7	3	4
30 – 35	6	8	10	7	–
35 – 40	4	12	17	6	–

Calculate the correlation co-efficient.

Solutions

Calculation of correlation co-efficient

X	Y	200–300	300–400	400–500	500–600	600–700	f	fdy	fdy^2	$fdxy$
X	M	250	350	450	550	650				
	dx/dy	-2	-1	0	1	2				
20–25	22.5	-1	$-$	$-$	$-$	5	$\underline{-5}$	10	-10	10
25–30	27.5	0	$-$	6	0	3	$\underline{0}$	20	0	0
30–35	32.5	1	$\underline{-12}$	$\underline{-8}$	$\underline{0}$	7	$\underline{7}$	$-$	31	31
35–40	37.5	2	$\underline{-16}$	$\underline{-24}$	$\underline{0}$	6	$\underline{12}$	$-$	39	78
	f	10	26	34	21	9	$N = \frac{99}{100}$	99	197	-56
	fdx	-20	-26	0	21	18	Σfdy	Σfdy^2	Σfdx	Σfdy
	fdx^2	40	26	0	21	36	123	Σfdy^2	Σfdx	Σfdx^2
	$fdxdy$	-28	-32	0	14	-10	-56			

$$\Sigma f dx = -7; \quad \Sigma f dx^2 = 123; \quad \Sigma f dx dy = -56$$

$$\Sigma f dy = 99; \quad \Sigma f dy^2 = 197; \quad N = 100$$

$$\begin{aligned}
r &= \frac{N \sum f dx dy - \sum f dx \cdot \sum f dy}{\sqrt{N \sum f dx^2 - (\sum f dx)^2} \sqrt{N \sum f dy^2 - (\sum f dy)^2}} \\
&= \frac{100(-56) - (-7) \cdot (99)}{\sqrt{100 \times 123 - (-7)^2} \sqrt{100 \times 197 - (99)^2}} \\
&= \frac{-5600 - (-693)}{\sqrt{12300 - 49} \sqrt{19700 - 9801}} \\
&= \frac{-4907}{\sqrt{12251} \cdot \sqrt{9899}} = \frac{-4907}{110.68 \times 99.49} \\
&= \frac{-4907}{11011.55} = -0.45
\end{aligned}$$

$$r = -0.45$$

9.6.1 Rank Correlation Co-efficient

This method is developed by a British psychologist, Charles Edward Spearman in the year 1904. It is applied where the population is not normal and the shape of distribution is not known. It requires no assumption about the parameter of the population. If a set of items is ranked according to two different attributes, then the co-efficient of correlation between the two attributes is called the co-efficient of rank correlation. But it cannot be applied to the variable frequency distribution. The formula given by Spearman to calculate rank correlation co-efficient is given below.

$$R = 1 - \frac{6 \sum D^2}{N^3 - N}$$

where, R stands for Rank correlation co-efficient

D stands for the difference of ranks between two variables

N stands for the total number of the pairs.

(a) Calculation of Rank Correlation Co-efficient when the Ranks are given

Steps

- Find out the difference of two ranks, that is, D for two variables.

328 Business Statistics

2. Take the square of these ranks in D^2 and find out ΣD^2 .
3. Substitute the value in the formula.

$$R = 1 - \frac{6 \sum D^2}{N^3 - N}$$

Illustration 9.8

Calculate Rank correlation co-efficient from the following data:

Applicant	A	B	C	D	E	F	G	H	I	J
Ranking by Interview Procedure, R_1	1	2	3	4	5	6	7	8	9	10
Ranking by Psychological Tests, R_2	3	4	10	7	8	5	1	2	6	9

Solutions

Calculation of Rank correlation co-efficient

Applicant	R_1	R_2	$D = R_1 - R_2$	D^2
A	1	3	-2	4
B	2	4	-2	4
C	3	10	-7	49
D	4	7	-3	9
E	5	8	-3	9
F	6	5	1	1
G	7	1	6	36
H	8	2	6	36
I	9	6	3	9
J	10	9	1	1
$\Sigma D^2 = 158$				

$$\begin{aligned}
 R &= 1 - \frac{6 \sum D^2}{N^3 - N} \\
 &= 1 - \frac{6 \times 158}{10^3 - 10} \\
 &= 1 - \frac{948}{1000 - 10} = 1 - \frac{948}{990} \\
 &= 1 - 0.96 = + 0.04
 \end{aligned}$$

(b) Calculation of Rank Correlation Co-efficient when Ranks are not given

When the ranks are not given for the data, ranks should be calculated for both the variables. The ranks may be given either ascending order or descending order, that is, either highest value or lowest value may be given as rank 1. But the same technique should be followed for both the variables.

Illustration 9.9

From the following data, calculate Rank correlation co-efficient:

x	80	30	60	40	20	66	96	70
y	30	24	28	68	55	43	38	40

Solutions

Starting the ranking process from the highest value for the both the variables as shown below:

Calculation of Rank correlation co-efficient

X	Y	R₁	R₂	D = R₁ - R₂	D²
80	30	2	6	-4	16
30	24	7	8	-1	1
60	28	5	7	-2	4
40	68	6	1	5	25
20	55	8	2	6	36
66	43	4	3	1	1
96	38	1	5	-4	16
70	40	3	4	1	1
$\Sigma D^2 = 100$					

$$\begin{aligned}
 R &= 1 - \frac{6 \sum D^2}{N^3 - N} \\
 &= 1 - \frac{6 \times 100}{8^3 - 8} \\
 &= 1 - \frac{600}{512 - 8}
 \end{aligned}$$

$$\begin{aligned}
 &= 1 - \frac{600}{504} \\
 &= 1 - 1.19 \\
 &= -0.19
 \end{aligned}$$

(c) When Equal Ranks

In some cases, there may be chances of obtaining same rank for two or more observations. In such a situation, it is essential to give equal rank for all such observation. For example, if two observations are ranked equal at fourth place, they are each given the rank $4 + 5/2$, that is, 4.5 which if three are ranked equal at fourth place, they are given the rank $4 + 5 + 6/3 = 6$.

The following formula is used to calculate rank correlation co-efficient when the equal ranks are obtained.

$$R = 1 - \frac{6 \left[\sum D^2 + 1/2(m^3 - 3) + 1/12(m^3 - 3) + 1/12(m^3 - 3) \right]}{N^3 - N}$$

where,

D stands for the difference of ranks between the two observations

N stands for the total number of the pairs.

M stands for the number of items whose ranks are common.

If there are more than one such group of items units' common rank, this value is added many times as the number of such groups.

For example: If rank 5 repeated 2 times,

Rank 6 repeated 3 times and the rank 9 repeated 3 times then $(m^3 - m)$ would be $(2^3 - 2)(3m^3 - 3m)(3^3 - 3)$.

Illustration 9.10

From the following data relate to the marks obtained by 10 students of a class in statistics and costing.

Statistics	30	38	28	27	28	23	30	33	28	35
Costing	29	27	22	29	20	29	18	21	27	22

Calculate Spearman's rank correlation co-efficient.

Solutions

Calculation of rank correlation co-efficient

Statistics	Rank R_1	Costing	Rank R_2	$D^2 = R_1 - R_2$	D^2
30	3.5	29	2	1.5	2.25
38	1	27	4.5	3.5	10.25
28	5	22	6.5	1.5	2.25
27	8	29	2	6	36
28	5	20	9	4	16
23	10	29	2	8	64
30	3.5	18	10	6.5	12.25
33	2	21	8	6	36
28	5	27	4.5	0.5	0.25
35	9	22	6.5	2.5	6.25
$\Sigma D^2 = 149.50$					

$$\begin{aligned}
 R &= 1 - \frac{6 \left[\sum D^2 + 1/12(m^3 - m) + 1/12(m^3 - m) + 1/12(m^3 - m) \right]}{N^3 - N} \\
 &= \frac{6 \times 149.50 + 1/12(2^3 - 2) + 1/12(3^3 - 3) + 1/12(3^3 - 3) + 1/12(2^3 - 2) + 1/12(2^3 - 2)}{10^3 - 10} \\
 &= 1 - \frac{6 \times 149.50 + 1/2 + 2 + 2 + 1/2 + 1/2}{1000 - 10} \\
 &= 1 - \frac{6 \times 149.50 + 5.5}{990} \\
 &= 1 - \frac{6 \times 155}{990} \\
 &= 1 - \frac{930}{990} \\
 &= 1 - 0.939 \\
 R &= + 0.061
 \end{aligned}$$

Illustration 9.11

Compute the rank correlation co-efficient from the following data.

Series x	115	109	112	87	98	98	120	100	98	118
Series y	75	73	85	70	76	65	82	73	68	80

Solutions

Calculation of rank correlation co-efficient

Series x	Rank R_1	Series y	Rank R_2	D	$R_1 - R_2$	D^2
115	8	75	6	2	2	4
109	6	73	4.5	1.5	1.5	2.25
112	7	85	10	-3	-3	9
87	1	70	3	-2	-2	4
98	3	76	7	-4	-4	16
98	3	65	1	2	2	4
120	10	82	9	1	1	1
100	5	73	4.5	0.5	0.5	0.25
98	3	68	2	1	1	1
118	9	80	8	1	1	1
$\Sigma D^2 = 42.50$						

$$= 1 - \frac{6 \times 42.50 + 1/12(3^2 - 3) + 1/12(2^2 - 2)}{10^2 - 10}$$

$$= 1 - \frac{6 \times 42.50 + 2 + 0.167}{990} = 0.741$$

Illustration 9.12

The competitors in a beauty contest are ranked by three judges in the following order:

I Judge	1	5	4	8	9	6	10	7	3	2
II Judge	4	8	7	6	5	9	10	3	2	1
III Judge	6	7	8	1	5	10	9	2	3	4

Use the rank correlation co-efficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

Solutions

Here, we shall find the rank correlation co-efficient of each possible pair of judges, i.e., 1 – 2, 2 – 3, 1 – 3.

Computation of Rank correlation co-efficient

R_1	R_2	R_3	$R_1 - R_2$	d_1^2	$R_1 - R_3$	d_2^2	$R_2 - R_3$	d_3^2
			d_1	d_2	d_3			
1	4	6	-3	9	-5	25	-2	4
5	8	7	-3	9	-2	4	1	1
4	7	8	-3	9	-4	16	-1	1
8	6	1	2	4	7	49	-5	25
9	5	5	4	16	4	16	0	0
6	9	10	-3	9	-4	16	-1	1
10	10	9	0	0	1	1	1	1
7	3	2	4	16	5	25	1	1
3	2	3	1	1	0	0	-1	1
2	1	4	1	1	-2	4	-3	9
			$\Sigma D_1^2 = 74$			$\Sigma D_2^2 = 166$		
						$\Sigma D_3^2 = 44$		

$R_1 - R_2:$

$$\begin{aligned} R_1 &= 1 - \frac{6 \sum D_1^2}{(N^3 - N)} = 1 - \frac{6 \times 74}{10^3 - 10} \\ &= 1 - \frac{444}{990} \\ &= 1 - 0.448 = +0.552 \end{aligned}$$

$R_1 - R_3:$

$$\begin{aligned} R &= 1 - \frac{6 \sum D_2^2}{(N^3 - N)} = 1 - \frac{6 \times 166}{10^3 - 10} \\ &= 1 - \frac{996}{990} = 1 - 1.006 = -0.006 \end{aligned}$$

$R_2 - R_3:$

$$\begin{aligned} r &= 1 - \frac{6 \sum D_3^2}{(N^3 - N)} = 1 - \frac{6 \times 44}{10^3 - 10} \\ &= 1 - \frac{264}{990} \\ &= 1 - 0.266 \\ &= 0.734 \end{aligned}$$

So, judges 2 and 3 are nearest approach to common tastes in beauty.

Merits of Rank Correlation Co-efficient

- (i) It is simpler to understand and easy to calculate as compared to Karl Pearson's method.

334 Business Statistics

- (ii) It is a use for qualitative data such as honesty, heavy, efficiency.
- (iii) It is a useful method when the ranks are given without the actual data.

Demerits

- (i) It cannot be useful when grouped for frequencies.
- (ii) It is not as accurate as Karl Pearson's co-efficient of correlation.

9.6.5 Concurrent Deviation Method

This method is based on the increase or decrease of the variables. The deviation of change of x variable and y variable should be ascertained. The formula is

$$r = \pm \frac{\sqrt{2C - n}}{n}$$

where, r stands for co-efficient of correlation

C stands for number of concurrent deviations or number of positive signs after multiplying dx and dy .

n stands for number of paired observations compared.

Steps

1. Find out the direction of change of x variable, that is, as compared with first value, if the second value is increasing or decreasing. If it is increasing, put + sign, if it is decreasing, put – sign; if it is constant, put 0. Continue this process for all the values and it is denoted by dx .
2. Find out the direction of change in y variable as in the case of x variable and denoted by dy .
3. Multiply dx with dy and determine the value of c , which is the number of positive signs.
4. Apply the formula which is

$$rc = \pm \frac{\sqrt{2C - n}}{n}$$

Illustration 9.13

Calculate the co-efficient of concurrent deviation from the following.

x	y
30	45
40	40
55	37

x	y
37	65
33	24
48	66
53	76

Solutions

Calculation of correlation by concurrent deviation method

x	dx	y	dy	Dxdy
30		45		
40	+	40	-	-
55	+	37	-	-
37	-	65	+	-
33	-	24	-	+
48	+	66	+	+
53	+	76	+	+
C = 3				

$$\begin{aligned}
 r &= \pm \frac{\sqrt{2C - n}}{n} \\
 &= \pm \frac{\sqrt{2 \times 3 - 6}}{6} \\
 &= \pm \frac{\sqrt{0}}{6} = 0
 \end{aligned}$$

Illustration 9.14

Calculate the co-efficient of correlation through concurrent deviation method.

Months	Income (Rs)	Expenditure (Rs)
1	800	540
2	900	300
3	950	450
4	880	600
5	900	650
6	950	675
7	925	550
8	1000	700
9	1050	675
10	1000	650
11	1100	680
12	1150	700

Solutions

Calculation of correlation by concurrent deviation method

Months	Income (Rs)	dx	Expenditure (Rs)	Dy	$Dxdy$
1	800		540		
2	900	+	300	-	-
3	950	+	450	+	+
4	880	-	600	+	-
5	900	+	650	+	+
6	950	+	675	+	+
7	925	-	550	-	+
8	1000	+	700	+	+
9	1050	+	675	-	-
10	1000	-	650	-	+
11	1100	+	680	+	+
12	1150	+	700	+	+

$$C = 8$$

$$\begin{aligned}
 R &= \pm \frac{\sqrt{2C - n}}{n} \\
 &= \pm \frac{\sqrt{2 \times 8 - 11}}{11} \\
 &= \pm \frac{\sqrt{16 - 11}}{11} \\
 &= \pm \frac{\sqrt{5}}{11} \\
 &= 0.67
 \end{aligned}$$

Illustration 9.15

Calculate the co-efficient of concurrent deviation from the following data.

Months	Sales (Rs)	Advt. Expenses (Rs)
Jan.	12000	3000
Feb.	14000	3200
Mar.	13000	3100
Apr.	16000	3300
May	16200	2900
June	15800	3150

Solutions

Calculation of concurrent co-efficient of correlation

Months	Sales (Rs)	dx	Advt. Expenses (Rs)	Dy	$Dxdy$
Jan.	12000		3000		
Feb.	14000	+	3200	+	+
Mar.	13000	-	3100	-	+
Apr.	16000	+	3300	+	+
May	16200	+	2900	-	-
June	15800	-	3150	+	-
C = 3					

$$\begin{aligned}
 r_c &= \pm \frac{\sqrt{2C-n}}{n} \\
 &= \pm \frac{\sqrt{2 \times 3 - 5}}{5} \\
 &= \pm \frac{\sqrt{6-5}}{5} \\
 &= \pm \frac{\sqrt{1}}{5} \\
 &= 0.447 \\
 &= 0.45
 \end{aligned}$$

9.7 LAG AND LEAD IN CORRELATION

It is important to consider lag and lead in the relationship of variables, which do not show simultaneous changes. One may find that there is some time gap between a cause and effect relationship is established. This difference in the period is known as lag.

For example,

1. The advertisement expenses spent in one month may have the effect on sales after two months.
2. The supply of raw material may increase today, but it may not have immediate effect on the price.
3. The effect of income has an impact on expenditure in the next month.
4. The boom in agricultural products may act reflected in industrial output after a gap of time.

Illustration 9.16

The following are the monthly figures of advertising expenditure and sales of a firm. It is generally found that the advertising expenditure has its impact on sales

338 Business Statistics

generally after two months, allowing for this time lag, calculate co-efficient of correlation.

Months	Advt. Expenses (Rs)	Sales (Rs)
Jan.	5000	120000
Feb.	6000	150000
Mar.	7000	160000
Apr.	9000	200000
May	12000	220000
June	15000	250000
July	14000	240000
Aug.	16000	260000
September	17000	280000
October	19000	290000
November	20000	310000
December	25000	390000

Solutions

Calculation of correlation co-efficient

Months	Advt. Expenses	$x = \frac{x - \bar{x}}{1000}$	x^2	Sales	$y = \frac{y - \bar{y}}{1000}$	y^2	xy
Jan.	5000	-7	49	160000	-100	10000	700
Feb.	6000	-6	36	200000	-60	3600	360
Mar.	7000	-5	25	220000	-40	1600	200
Apr.	9000	-3	9	250000	-10	100	30
May	12000	0	0	240000	-20	400	0
June	15000	3	9	260000	0	0	0
July	14000	2	4	280000	20	400	40
Aug.	16000	4	16	290000	30	900	120
Sept.	17000	5	25	310000	50	2500	250
Oct.	19000	7	49	390000	130	16900	910
	$\Sigma x = 0$	$\Sigma x^2 = 222$		$\Sigma y = 0$	$\Sigma y^2 = 36400$	$\Sigma xy = 2610$	

$$\bar{x} = \frac{120000}{10} = 12000$$

$$\bar{y} = \frac{2600000}{10} = 260000$$

$$\begin{aligned}\Sigma dx &= 0; & \Sigma dy &= 0; & N &= 10 \\ \Sigma dx^2 &= 222; & \Sigma dy^2 &= 36400; & \Sigma dxdy &= 2610\end{aligned}$$

$$\begin{aligned}
 r &= \frac{N \sum dxdy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \cdot \sqrt{N \sum dy^2 - (\sum dy)^2}} \\
 &= \frac{10 \times 2610 - (0)(0)}{\sqrt{10 \times 222 - (0)^2} \sqrt{10 \times 36400 - (0)^2}} \\
 &= \frac{26100}{\sqrt{2220} \sqrt{36400}} = \frac{26100}{47.1 \times 603.3} \\
 r &= \frac{26100}{31025.43} = 0.84 \\
 r &= 0.84
 \end{aligned}$$

So, there is a high degree of positive correlation between advertisement expenditure and sales.

9.7.1 Calculation of Correlation in Time Series

Data arranged on the basis of time is called time series. The variables based on time may be classified into:

- (i) Long term period.
- (ii) Short term period.

(i) Calculation of Correlation Co-efficient for Long Term Period To know the correlation co-efficient for observations based on long term, trend value for all the observations should be ascertained. The trend value for the observations can be ascertained either through moving average method or through least square method. The trend value so obtained should be utilised to calculate the deviations from the mean of the trends. Then the following common formula for co-efficient of correlation should be applied to ascertain the value of correlation co-efficients.

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

(ii) Calculation of Correlation Co-efficient for Short Term Period Co-efficient of correlation for short period is also based on trends. Trends for the two sets of variables should be calculated either through the method of least squares or through the method of moving average. The trend so calculated should be compared units the corresponding variables and the deviations should be ascertained. Then the following common formula should be applied to ascertain the value of correlation co-efficient.

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

Illustration 9.17

The following data related to the index of supply and index of demand. Calculate co-efficient of correlation for short term variations through Karl Pearson's method.

Year	Index of Supply	Index of Demand
1990	230	170
1991	240	176
1992	234	184
1993	250	194
1994	246	196
1995	250	200
1996	270	216
1997	274	214
1998	290	224
1999	296	226
2000	290	230

Year	Index of supply	Five yearly Moving average X	Deviations of Activate values from M.A. x	Square of Deviations X^2	Index of Demand	Five yearly Moving average y	Deviations of Activate values from M.A. y	Square of Deviations Y^2	xy
1990	230				170				
1991	240				176				
1992	234	240	-6	36	184	184	0	0	0
1993	250	244	6	36	194	190	4	16	24
1994	246	250	-4	16	196	198	-2	4	8
1995	250	258	8	64	200	204	-4	16	32
1996	270	266	4	16	216	210	6	36	24
1997	274	284	-10	100	214	216	-2	4	20
1998	290	286	4	16	224	222	2	4	8
1999	296				226	226			
2000	290				230				
$\Sigma x^2 = 284$						$\Sigma y^2 = \Sigma xy = 80 \quad 116$			

Solutions

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

$$\begin{aligned}
 &= \frac{116}{\sqrt{284} \times \sqrt{80}} \\
 &= \frac{116}{16.85 \times 8.94} = \frac{116}{150.64} \\
 r &= 0.77
 \end{aligned}$$

Co-efficient of Correlation and Probable Error Probable Error of correlation co-efficient is a measure of testing the reliability of an observed value in so far as it depends upon the condition of random sampling.

$$\text{Probable Error (P.E.)} = 0.6745 \times \frac{1-r^2}{\sqrt{N}}$$

P.E. stands for Probable Error

r stands for co-efficient of correlation.

N stands for the total number of pairs of observations.

0.6745 is a constant number.

If the value of r is less than 6 times of Probable Error ($r < 6$ P.E.) then the value of r is insignificant. Hence, there should be no correlation between the two variables. If the value of r is more than 6 times of Probable Error ($r > 6$ P.E.) then the value of r is significant. Hence, the correlations between the two variables are very certain. The limits of the co-efficient of correlation of the population can be ascertained by adding and deducting the co-efficient of correlation with the Probable Error. It can be stated as

$$P = r \pm \text{P.E.}$$

Illustration 9.18

Find the significance of correlation co-efficient when $r = 0.46$ for 40 observations.

Solutions

Significance of correlation can be tested through the Probable Error.

$$\begin{aligned}
 \text{P.E.} &= 0.6745 \times \frac{1 - (0.46)^2}{\sqrt{40}} \\
 &= 0.6745 \times \frac{1 - 0.246}{6.32} \\
 &= 0.6745 \times \frac{0.7884}{6.32}
 \end{aligned}$$

$$\text{P.E.} = 0.084$$

If $r > 6$ P.E. then, there exists the significance for the correlation co-efficient.

$$6 \times \text{P.E.} = 6 \times 0.084 = 0.504$$

P.E. is greater than correlation co-efficient, the correlation co-efficient is not significant.

Illustration 9.19

If $r = 0.8$, and $n = 56$, find the Probable Error of correlation co-efficient and determine the limits for population r .

$$\begin{aligned}\text{P.E.} &= 0.6745 \times \frac{1-r^2}{\sqrt{n}} \\ \text{P.E.} &= 0.6745 \times \frac{1-(0.86)^2}{\sqrt{56}} \\ &= 0.6745 \times \frac{1-0.64}{7.48} \\ &= 0.6745 \times \frac{0.36}{7.48} \\ &= 0.032\end{aligned}$$

Limits of population correlation = 0.8 ± 0.032
0.768 to 0.832

Co-efficient of Determination The square of the co-efficient of correlation is called co-efficient of determination, that is, r^2 . It gives the percentage variation in the dependent variable that is accounted for or explained by the independent variable. The formula is

$$\text{Co-efficient of Determination } r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

The value of co-efficient of determination is always with positive value. For example, when $r = -0.8$, the co-efficient of determination is $(-0.8)^2 = +0.64$, when $r = +0.4$, then the co-efficient of determination, $r^2 = (+0.4)^2 = +0.16$.

Co-efficient of non-determination is calculated through the nature of unexplained variance to total variance. It is denoted by K^2 .

The square root of the co-efficient of non-determination is called co-efficient of alienation. It is denoted by k . Hence,

$$\begin{aligned}\text{Co-efficient of non-determination} &= \frac{\text{Unexplained Variance}}{\text{Total Variance}} \\ &= 1 - \frac{\text{Explained Variance}}{\text{Total Variance}} \\ &= 1 - r^2\end{aligned}$$

Illustration 9.20

Calculate co-efficient of determination and co-efficient of non-determination for the following data:

x	y
30	20
36	30
42	40
48	50
54	60
60	76

Solutions

x	x (x - x̄) x - 45	x²	y	y (y - ȳ) y - 46	y²	xy
30	- 15	225	20	- 26	676	390
36	- 9	81	30	- 16	256	144
42	- 3	9	40	- 6	36	18
48	3	9	50	4	16	12
54	9	81	60	14	196	126
60	15	225	76	24	576	360
$\Sigma x = 0$		$\Sigma x^2 = 630$	$\Sigma y = 0$		$\Sigma y^2 = 1756$	$\Sigma xy = 1050$

$$\begin{aligned}
 r &= \frac{\sum xy}{\sqrt{\sum x^2} \cdot \sqrt{\sum y^2}} \\
 &= \frac{1050}{25.10 \times 41.90} \\
 &= \frac{1050}{1051.69}
 \end{aligned}$$

Co-efficient of determination = $r^2 = (0.998)^2 = 0.996$

$$\begin{aligned}
 \text{Co-efficient of Non-determination} &= 1 - r^2 \\
 &= 1 - 0.996 = 0.004.
 \end{aligned}$$

9.8 MISCELLANEOUS ILLUSTRATIONS

Illustration 9.21

Find Karl Pearson's Co-efficient of correlation from the following data.

Firm No.	1	2	3	4	5
Sales (Rs '000)	145	155	165	175	185
Expenses (Rs '000)	90	100	110	120	130

Solutions

Let the sales be denoted by x and the expenses by y .

Computation of co-efficients of correlation

Firm No.	x	$x(x - \bar{x})$	x^2	y	$y(y - \bar{y})$	y^2	xy
1	145	-20	400	90	-20	400	400
2	155	-10	100	100	-10	100	100
3	165	0	0	110	0	0	0
4	175	+10	100	120	+10	100	100
5	185	+20	400	130	+20	400	400
	$\Sigma x = 825$		$\Sigma x^2 = 1000$	$\Sigma y = 550$		$\Sigma y^2 = 1000$	$\Sigma xy = 1000$

$$\bar{x} = \frac{\sum x}{N} = \frac{825}{5} = 165$$

$$\bar{y} = \frac{\sum y}{N} = \frac{550}{5} = 110$$

$$\sum x^2 = 1000$$

$$\sum y^2 = 1000$$

$$\sum xy = 1000$$

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

$$r = \frac{1000}{\sqrt{1000} \times \sqrt{1000}}$$

$$r = \frac{1000}{31.62 \times 31.62}$$

$$r = \frac{1000}{999.82} = +1.00$$

$$r = +1.$$

Therefore, there is perfect positive correlation between the variables of sales and expenses.

Illustration 9.22

Find Karl Pearson's Co-efficient of correlation from the following data.

Sales (Rs '000)	15	18	22	28	32	46	52
Profit (Rs '000)	52	66	78	87	96	125	141

Solutions

Let the sales be denoted by x and the profit by y .

Computation of co-efficients of correlation

x	$(x = x - \bar{x})$	x^2	y	$(y = y - \bar{y})$	y^2	xy
15	-15.43	238.98	52	-40.14	1611.22	619.36
18	-12.43	154.50	66	-26.14	683.30	324.92
22	-8.43	71.06	78	-14.14	199.94	119.20
28	-2.43	5.90	87	-5.14	26.42	12.49
32	+1.57	2.46	96	+3.86	14.90	6.06
46	+15.57	242.42	125	+32.86	1079.78	511.63
52	+21.57	465.26	141	+48.86	2387.30	1053.91
$\Sigma x =$	$\Sigma x =$	$\Sigma x^2 =$	$\Sigma y =$	$\Sigma y =$	$\Sigma y^2 =$	$\Sigma xy =$
213	-0.01	1179.68	645	0.02	6002.86	2647.5

$$\bar{x} = \frac{\sum x}{N} = \frac{213}{7} = 30.43$$

$$\bar{y} = \frac{\sum y}{N} = \frac{645}{7} = 92.14$$

$$\sum x^2 = 1179.68$$

$$\sum y^2 = 6002.86$$

$$\sum xy = 2647.57$$

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

$$= \frac{2647.57}{\sqrt{1179.68} \cdot \sqrt{6002.86}}$$

$$= \frac{2647.57}{34.35 \times 77.48}$$

$$r = \frac{2647.57}{2661.44} = +0.99$$

$$r = +0.99.$$

Therefore, there is a high degree positive correlation between the x and y .

Illustration 9.23

Find out the co-efficient of correlation from the following data.

Output of pen (in'000s)	12	14	15	16	17	19	21
Cost per pen (Rs.)	270	290	320	440	480	520	570

Solutions

Let the output of pen be denoted by x and the cost per pen by y .

Computation of co-efficients of correlation

x	$dx = (x - A)$ $x - 16$	dx^2	y	$dy = (y - A)$ $y - 480$	dy^2	$dxdy$
12	-4	16	270	-210	44100	840
14	-2	4	290	-190	36100	380
15	-1	1	320	-160	25600	160
16	0	0	440	-40	1600	0
17	+1	1	480	0	0	0
19	+3	9	520	+40	1600	120
21	+5	25	570	+90	8100	450
$\Sigma dx =$		$\Sigma dx^2 =$		$\Sigma dy =$	$\Sigma dy^2 =$	$\Sigma dxdy =$
2		56		-470	117100	1950

$$\begin{aligned}
 r &= \frac{N \sum dxdy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \sqrt{N \sum dy^2 - (\sum dy)^2}} \\
 &= \frac{7 \times 1950 - (2 \times -470)}{\sqrt{7 \times 56 - (2)^2} \sqrt{7 \times 117100 - (-470)^2}} \\
 &= \frac{13650 + 940}{\sqrt{392 - 4} \sqrt{819700 - 220900}} \\
 &= \frac{14590}{\sqrt{388} \sqrt{598800}} \\
 &= \frac{14590}{19.70 \times 773.82} \\
 &= \frac{14590}{15244.25} = +0.96
 \end{aligned}$$

$$r = +0.96$$

Illustration 9.24

Find out the co-efficient of correlation from the following data:

Demand	100	120	120	132	142	158	160
Supply	90	110	100	130	138	140	152

Solutions

Computation of co-efficient of correlation

Height of Uncle (in inches)	x Series			y Series			Products of x and y series $dxdy$
	Deviations from assu- med mean	Square of Devia- tions dx^2 (in inches)	Height of Aunty y	Deviations from assu- med mean	Square of Devia- tions dy^2		
	x (120) dx	y	(130) dy				
100	- 20	+ 400	90	- 40	1600	800	
120	- 0	0	110	- 20	400	0	
120	0	0	100	- 30	900	0	
132	+ 12	+ 144	130	0	0	0	
142	+ 22	+ 484	138	+ 8	16	176	
158	+ 38	+ 1444	140	+ 10	100	380	
160	+ 40	+ 1600	152	+ 22	484	880	
$\Sigma x =$ 932	$\Sigma dx =$ 92	$\Sigma dx^2 =$ 4072	$\Sigma y =$ 860	$\Sigma dy =$ -50	$\Sigma dy^2 =$ 3500	$\Sigma dxdy =$ 2236	

$$r = \frac{N \sum dxdy - \sum dx \sum dy}{\sqrt{N \sum dx^2 - (\sum dx)^2} \cdot \sqrt{N \sum dy^2 - (\sum dy)^2}}$$

$$\sum dxdy = 2236, \sum dx = 92, \sum dy = -50, \sum dx^2 = 4072, \sum dy^2 = 3200, \\ N = 7$$

$$r = \frac{2236 \times 7 - (92 \times -50)}{\sqrt{4072 \times 7 - (92)^2} \times \sqrt{3200 \times 7 - (-50)^2}} \\ = \frac{15652 + 4600}{\sqrt{(28504 - 8464)} \times \sqrt{(22400 - 2500)}} \\ = \frac{20252}{\sqrt{20040 \times 19900}}$$

$$r = \frac{20252}{20923.67} = + 0.96$$

or

x	x^2	y	y^2	xy
100	10000	90	8100	9000
120	14400	110	12100	13200
120	14400	100	10000	12000
132	17424	130	16900	17160
142	20164	138	19044	19596
158	24964	140	19600	22120
160	25600	152	23104	24320
$\Sigma x = 932$	$\Sigma x^2 = 126952$	$\Sigma y = 860$	$\Sigma y^2 = 108848$	$\Sigma xy = 117396$

$$\begin{aligned}
 r &= \frac{N \sum xy - (\sum x) \times (\sum y)}{\sqrt{N \sum x^2 - (\sum x)^2} \cdot \sqrt{N \sum y^2 - (\sum y)^2}} \\
 &= \frac{7 \times 117396 - (932 \times 860)}{\sqrt{7 \times 126952 - (932)^2} \times \sqrt{7 \times 108848 - (860)^2}} \\
 &= \frac{821772 - 801520}{\sqrt{888664 - 868624} \sqrt{761936 - 739600}} \\
 &= \frac{20252}{\sqrt{20040} \sqrt{22336}} \\
 &= \frac{20252}{141.56 \times 149.45} \\
 r &= \frac{20252}{21156.14} = 0.96
 \end{aligned}$$

Hence, there is a high degree of positive correlation between demand and supply.

Illustration 9.25

From the following table, calculate the co-efficient of correlation by Karl Pearson's method:

x	26	22	20	14	28
y	19	21	?	18	17

Arithmetic means of x and y series are 22 and 18 respectively.

Solutions

The missing value of y can be calculated as follows:

$$y = \frac{\sum y}{N}$$

$$\therefore Ny = \sum y$$

$$\Sigma y = 5 \times 18 = 90$$

Missing value of the third item = $90 - 75 = 15$.

Computation of correlation co-efficient

x	$x(x - \bar{x})$ $x = 22$	x^2	y	$y(y - \bar{y})$ $y = 18$	y^2	xy
26	4	16	19	1	1	4
22	0	0	21	3	9	0
20	-2	4	15	-3	9	6
14	-8	64	18	0	0	0
28	6	36	17	-1	1	-6
$\Sigma x = 110$		$\Sigma x^2 = 120$	$\Sigma y = 90$		$\Sigma y^2 = 20$	$\Sigma xy = 4$

$$\bar{x} = \frac{110}{5} = 22$$

$$\bar{y} = \frac{90}{5} = 18$$

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

$$= \frac{4}{\sqrt{120} \times \sqrt{20}}$$

$$= \frac{4}{\sqrt{2400}}$$

$$= \frac{4}{48.99}$$

$$r = +0.08$$

Illustration 9.26

Calculate r from the following data.

x	24	27	33	45	56
y	22	20	19	18	14

Solutions

Computation of correlation of co-efficient

x	x^2	y	y^2	xy
24	576	22	484	528
27	729	20	400	540

Contd.

x	x²	y	y²	xy
33	1089	19	361	627
45	2025	18	324	810
56	3136	14	196	784
185	7555	93	1765	3289

$$\begin{aligned}
 r &= \frac{N \sum xy - (\sum x) \times (\sum y)}{\sqrt{N \sum x^2 - (\sum x)^2} \cdot \sqrt{N \sum y^2 - (\sum y)^2}} \\
 &= \frac{(5 \times 3289) - (185 \times 93)}{\sqrt{(5 \times 7555) - (185)^2} \sqrt{5 \times 1765 - (93)^2}} \\
 &= \frac{16445 - 17205}{\sqrt{37775 - 34225} \sqrt{8825 - 8649}} \\
 &= \frac{-760}{\sqrt{3550} \sqrt{176}} \\
 &= \frac{-760}{59.58 \times 13.27} \\
 &= \frac{-760}{790.63} \\
 r &= -0.96
 \end{aligned}$$

Hence, there is a high degree of negative correlation between the variables of x and y .

Illustration 9.27

Calculate Pearson's co-efficient of correlation from the following data. Take 55 and 80 as the assumed average of the variables x and y respectively.

X	35	45	46	48	50	55	58	60	65	79	75
Y	66	60	58	70	72	74	75	80	84	90	86

Solutions

Calculation of Co-efficient of correlation

X	$dx = x - 55$	dx^2	y	$dy = y - 80$	dy^2	$dxdy$
35	-20	400	66	-14	196	280
45	-10	100	60	-20	400	200
46	-9	81	58	-22	484	198
48	-7	49	70	-10	100	70

Contd.

X	$dx = x - 55$	dx^2	y	$dy = y - 80$	dy^2	$dxdy$
50	-5	25	72	-8	64	40
55	0	0	74	-6	36	0
58	3	9	75	-5	25	-15
60	5	25	80	0	0	0
65	10	100	84	4	16	40
79	24	576	90	10	100	240
75	20	400	86	6	36	120
$\Sigma dx = 11$		$\Sigma dx^2 = 1765$	$\Sigma dy = -65$		$\Sigma dy^2 = 1457$	$\Sigma dxdy = 1173$

$$r = \frac{N \sum dxdy - (\sum dx)(\sum dy)}{\sqrt{N \sum dx^2 - (\sum dx)^2} \times \sqrt{N \sum dy^2 - (\sum dy)^2}}$$

$$\sum dxdy = 1173, \sum dx = 11, \sum dy = -65, \sum dx^2 = 1765, \sum dy^2 = 1457, N = 11$$

$$\begin{aligned} r &= \frac{(1173 \times 11) - (1173 \times -65)}{\sqrt{1765 \times 11 - (11)^2} \times \sqrt{1457 \times 11 - (-65)^2}} \\ &= \frac{12903 + 715}{\sqrt{(19415 - 121)} \times \sqrt{(16027 - 4225)}} \\ &= \frac{13618}{\sqrt{19294 \times 11802}} \\ &= \frac{13618}{138.90 \times 108.64} = \frac{13618}{15090.10} \\ r &= +0.90 \end{aligned}$$

Illustration 9.28

The following are the ranks obtained by 10 students in two subjects Costing and Business Maths. To what extent the knowledge of the students in the two subjects is related?

Costing	11	12	13	14	15	16	17	18	19	20
Business Maths	12	14	11	15	13	19	17	20	16	18

Solutions

Let the Costing be denoted by x and the Business Maths by y .

Calculation of Rank correlation co-efficient

Rank of Costing (R_1)	Rank of B. Maths (R_2)	$D (R_1 - R_2)$	D^2
11	12	-1	1
12	14	-2	4

Contd.

Rank of Costing (R_1)	Rank of B. Maths (R_2)	$D (R_1 - R_2)$	D^2
13	11	+ 2	4
14	15	- 1	1
15	13	+ 2	4
16	19	- 3	9
17	17	0	0
18	20	- 2	4
19	16	+ 3	9
20	18	+ 2	4
$\Sigma D^2 = 40$			

$$\begin{aligned}
 R &= 1 - \frac{6 \sum D^2}{N^3 - 1} \\
 &= 1 - \frac{6 \times 40}{10(10^2 - 1)} \\
 &= 1 - \frac{240}{10(100 - 1)} = 1 - \frac{240}{990} \\
 &= 1 - 0.24 \\
 R &= + 0.76
 \end{aligned}$$

There is high degree of positive correlation. Between Costing and Business Maths.

Illustration 9.29

A random sample of 5 school students is selected and their grades in Management Accounting and Tally are found to be:

Mgt. A/cg.	98	92	95	82	72
Tally	61	41	52	68	93

Calculate rank correlation co-efficient.

Solutions

Computation of Rank correlation co-efficient

Marks in Mgt.A/cg (x)	Ranks R_1	Marks in Tally (y)	Ranks R_2	Rank Diff. ($R_1 - R_2$)	D^2
98	1	61	3	- 2	4
92	3	41	5	- 2	4
95	2	52	4	- 2	4
82	4	68	2	+ 2	4
72	5	93	1	+ 4	16
$\Sigma D^2 = 32$					

$$\begin{aligned}
 \text{Pearson's Rank Correlation } (R) &= 1 - \frac{6\sum D^2}{N^3 - N} \\
 N &= 5, \quad D^2 = 32 \\
 &= 1 - \frac{6 \times 32}{5(5^2 - 1)} \\
 &= 1 - \frac{192}{5(25 - 1)} \\
 &= 1 - \frac{192}{5 \times 24} = 1 - \frac{192}{120} \\
 &= 1 - 1.6 = -0.6 \\
 &= -0.6.
 \end{aligned}$$

Illustration 9.30

From the following data calculate the rank correlation co-efficient after making adjustment for tied ranks.

<i>x</i>	58	43	30	9	26	26	75	14	26	67
<i>y</i>	23	23	24	6	25	14	30	19	16	29

Solutions

First we have to assign ranks to the variables.

x	Rank R_1	y	Rank R_2	$D(R_1 - R_2)$	D^2
58	8	23	5.5	2.5	6.25
43	7	23	5.6	1.5	6.25
30	6	24	7	-1	1
9	1	6	1	0	0
26	4	25	8	-4	16
26	4	14	2	2	4
75	10	30	10	0	0
14	2	19	4	-2	4
26	4	16	3	1	1
67	9	29	9	0	0
$\Sigma D^2 = 34.5$					

$$\begin{aligned}
 R &= 1 - \frac{6 \left[\sum D^2 + 1/12(m^3 - 3) + 1/12(m^3 - 3) \right]}{N^3 - N} \\
 &= 1 - \frac{6 \left[34.5 + 1/12(3^3 - 3) + 1/12(2^3 - 2) \right]}{10^3 - 10}
 \end{aligned}$$

$$\begin{aligned}
 &= 1 - \frac{6[34.5 + 1/12(24) + 1/12(6)]}{990} \\
 &= 1 - \frac{6[34.5 + 2 + 0.5]}{990} = 1 - \frac{270}{990} \\
 &= 1 - 0.27 = + 0.73 \\
 R &= + 0.73
 \end{aligned}$$

Illustration 9.31

The competitors in a face cream contest are ranked by three judges in the following order:

I Judge	23	25	24	28	29	26	39	27	23	22
II Judge	24	28	27	26	25	29	30	23	22	23
III Judge	26	27	28	24	25	30	29	22	23	24

Use rank correlation co-efficient to discuss which pair of judges has the nearest approach to common tastes in face cream.

First Judge R_1	Second Judge R_2	Third Judge R_3	$(R_1 - R_2)^2$ D_1	$(R_2 - R_3)^2$ D_2	$(R_1 - R_3)^2$ D_3	
23	24	26	1	4	9	
25	28	27	9	1	4	
24	27	28	9	1	16	
28	26	24	4	4	16	
29	25	25	16	0	16	
26	29	30	9	1	16	
39	30	29	81	1	100	
27	23	22	16	1	25	
23	22	23	1	1	0	
22	23	24	1	1	4	
$\Sigma D_1^2 = 147$			$\Sigma D_2^2 = 15$		$\Sigma D_3^2 = 206$	

Solutions

1st and 2nd Judges:

$$\begin{aligned}
 R &= 1 - \frac{6\sum D_1^2}{N^3 - N} \\
 \text{where, } N &= 10, \quad \sum D_1^2 = 147; \\
 &= 1 - \frac{6 \times 147}{10^3 - 10} \\
 &= 1 - \frac{882}{1000 - 10} = 1 - \frac{882}{990} \\
 &= 1 - 0.89 = + 0.11
 \end{aligned}$$

2nd and 3rd Judges:

$$\begin{aligned}
 R &= 1 - \frac{6\sum D_2^2}{N^3 - N} \\
 N = 10, \quad \sum D^2 &= 15; \\
 &= 1 - \frac{6 \times 15}{10^3 - 10} \\
 &= 1 - \frac{90}{1000 - 10} = 1 - \frac{90}{990} \\
 &= 1 - 0.09 = + 0.91
 \end{aligned}$$

1st and 3rd Judges:

$$\begin{aligned}
 R &= 1 - \frac{6\sum D_3^2}{N^3 - N} \\
 N = 10, \quad \sum D^2 &= 206 \\
 &= 1 - \frac{6 \times 206}{10^3 - 10} \\
 &= 1 - \frac{1236}{990} \\
 &= 1 - 1.25 = - 0.25
 \end{aligned}$$

The second and third judges have the nearest approach in common tastes in face cream because the coefficient of correlation is the highest between them.

Illustration 9.32

The following table gives the frequency according to production and distribution.

Distribution	Production					Total
	30	31	32	33	34	
400 – 450	1	2	6	2	5	16
450 – 500	2	4	8	4	2	20
500 – 550	2	3	7	6	1	19
550 – 600	4	1	5	8	4	22
600 – 650	3	1	4	3	10	21
Total	12	11	30	23	22	98

Is there any relationship between production and distribution?

Solutions

Let production be denoted by x and distribution by y .

Solutions

Calculation of correlation co-efficient

X	M	$\frac{dx}{dy}$	30	31	32	33	34	f	fdy	fdy^2	$fdxdy$
Y		$\frac{dx}{dy}$	-2	-1	0	1	2				
400–450	425	-2	<u>4</u> 1	<u>4</u> 2	<u>0</u> 6	<u>-4</u> 2	<u>-20</u> 5	16	-32	64	-16
450–500	475	-1	<u>4</u> 2	<u>4</u> 4	<u>0</u> 8	<u>-4</u> 4	<u>-4</u> 2	20	-20	20	0
500–550	525	0	<u>0</u> 2	<u>0</u> 3	<u>0</u> 7	<u>0</u> 6	<u>0</u> 1	19	0	0	0
550–600	575	1	<u>-8</u> 4	<u>-1</u> 1	<u>0</u> 5	<u>8</u> 8	<u>8</u> 4	22	22	22	7
600–650	625	2	<u>-12</u> 3	<u>-2</u> 1	<u>0</u> 4	<u>6</u> 3	<u>40</u> 10	21	42	84	32
	f	12	11	30	23	22		N = 98 = 12	Σfdy^2 190 = 23		
	fdx	-24	-11	0	23	44	$\Sigma f dx$ = 32				
	fdx²	48	11	0	23	88	$\Sigma f dx^2$ = 170				
	fdxdy	-12	5	0	6	24	23				

$\Sigma f dx dy$

$\Sigma f dx dy^2$

Co-efficient of Correlations

$$r = \frac{N \sum f dx dy - (\sum f dx)(\sum f dy)}{\sqrt{N \sum f dx^2 - (\sum f dx)^2} \sqrt{N \sum f dy^2 - (\sum f dy)^2}}$$

$\Sigma f dx dy = 23, \Sigma f dx = 32, \Sigma f dy = 12, \Sigma f dx^2 = 170, \Sigma f dy^2 = 190, N = 98$

$$r = \frac{23 \times 98 - (32 \times 12)}{\sqrt{98 \times 170 - 32^2} \times \sqrt{98 \times 190 - 12^2}}$$

$$= \frac{98 - 32 \times 12}{\sqrt{98 \times 170 - 1024} \times \sqrt{98 \times 190 - 144}}$$

$$= \frac{23 - 3.92}{\sqrt{170 - 10.45} \times \sqrt{190 - 1.47}}$$

$$= \frac{19.08}{\sqrt{159.55} \sqrt{188.53}}$$

$$= \frac{19.08}{12.63 \times 13.73}$$

$$= \frac{19.08}{173.41}$$

$$r = + 0.11$$

There is low degree positive correlation between production and distribution.

Illustration 9.33

The co-efficient of correlation between two variables x and y is 0.52. Their Co-variance is 48. The variance of x is 24. Find the standard deviation of y series.

Solutions

$$\text{Co-variance} = r \times \sigma_x \times \sigma_y$$

$$\text{Co-variance} = 48; r = 0.52; \sigma_x = \sqrt{24} = 4.90$$

$$48 = 0.52 \times 4.90 \times \sigma_y$$

$$48 = 2.548 \sigma_y$$

$$\sigma_y = \frac{48}{2.548} = 18.84.$$

Illustration 9.34

The coefficients of correlation of the marks obtained by 17 students in two subjects were found to be 0.8. It was detected that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 7 instead of 5. What should be the correct rank correlation coefficient?

Solutions

$$R = 1 - \frac{6 \sum D^2}{N^3 - N}$$

$$R = 0.8; N = 17$$

Substituting the value,

$$\begin{aligned} 0.8 &= 1 - \frac{6 \sum D^2}{17^3 - 17} \\ &= 1 - \frac{6 \sum D^2}{4913 - 17} \\ 0.8 &= 1 - \frac{6 \sum D^2}{4896} \quad \text{or} \quad \frac{6 \sum D^2}{4896} = 0.2 \end{aligned}$$

$$6 \sum D^2 = 4896 \times 0.2$$

$$6 \sum D^2 = 979.2$$

$$6 \sum D^2 = \frac{979.2}{6} = 163.2$$

But the correct value of $\sum D^2$:

$$\begin{aligned} \text{Correct } \sum D^2 &= 163.2 - [(7)^2 - (5)^2] \\ &= 163.2 - (49 - 25) \\ &= 163.2 - 24 = 139.2 \end{aligned}$$

$$\begin{aligned} \text{Correct } R &= 1 - \frac{6 \times 139.2}{4896} \\ &= 1 - \frac{835.2}{4896} \\ &= 1 - 0.17 = 0.83 \end{aligned}$$

Illustration 9.35

Find the Probable Error if $n = 4$, $r = 0.6$.

Solutions

$$\begin{aligned} \text{Probable Error (P.E.)} &= 0.6745 \times \frac{1 - r^2}{\sqrt{n}} \\ &= 0.6745 \times \frac{1 - 0.6^2}{\sqrt{4}} \end{aligned}$$

$$\begin{aligned}
 &= 0.6745 \times \frac{1 - 0.36}{2} \\
 &= 0.6745 \times \frac{0.64}{2} \\
 &= 0.6745 \times 0.32 \\
 &= 0.2158 \\
 &= 0.22
 \end{aligned}$$

Illustration 9.36

Calculate the co-efficient of correlation between two values of x and y and its P.E.

Variable x	3	4	6	5	8
Variable y	13	15	14	17	16

Solutions

Calculation of co-efficient of correlation

x	$dx = x - 6$	dx^2	y	$dy = y - 14$	dy^2	$dxdy$
3	-3	9	13	-1	1	3
4	-2	4	15	1	1	-2
6	0	0	14	0	0	0
5	-1	1	17	3	9	-3
8	2	4	16	2	4	4
$\Sigma x =$	$\Sigma dx =$	$\Sigma dx^2 =$	$\Sigma y =$	$\Sigma dy =$	$\Sigma dy^2 =$	$\Sigma dxdy =$
26	-4	18	75	5	15	2

$$\begin{aligned}
 r &= \frac{N \cdot \sum dxdy - (\sum dx)(\sum dy)}{\sqrt{N \cdot \sum dx^2 - (\sum dx)^2} \times \sqrt{N \cdot \sum dy^2 - (\sum dy)^2}} \\
 r &= \frac{(5 \times 2) - (-4 \times 5)}{\sqrt{5 \times 18 - (-4)^2} \times \sqrt{5 \times 15 - (5)^2}} \\
 &= \frac{10 + 20}{\sqrt{90 - 16} \times \sqrt{75 - 25}} \\
 &= \frac{30}{\sqrt{74} \times \sqrt{50}} \\
 &= \frac{30}{8.60 \times 7.07} \\
 &= \frac{30}{60.802} = 0.49
 \end{aligned}$$

$$\begin{aligned}
 \text{Probable Error (P.E.)} &= 0.6745 \times \frac{1-r^2}{\sqrt{n}} \\
 &= 0.6745 \times \frac{1-0.49^2}{\sqrt{5}} \\
 &= 0.6745 \times \frac{1-0.2401}{2.236} \\
 &= 0.6745 \times \frac{0.5126}{2.236} = 0.229
 \end{aligned}$$

Illustration 9.37

Find the correlation coefficient between sales and advertisements of a company. The company follows a time lag of two months between advertisement and sales.

Year	Sales (in lakhs)	Advertisement (in thousands)
1995	200	300
1996	230	350
1997	260	340
1998	240	400
1999	270	410
2000	250	390
2001	260	420

Solutions

Calculation of co-efficient of correlation

Year	x	$\frac{dx}{x-26}$	dx^2	y	$\frac{dy}{y-14}$	dy^2	$dxdy$
1995	200	-6	36	340	-1	1	3
1996	230	-3	9	400	1	1	-2
1997	260	0	0	410	0	0	0
1998	240	-2	4	390	3	9	-3
1999	270	1	1	420	2	4	4
$\Sigma x =$	1200	$\Sigma dx = -10$	$\Sigma dx^2 = 50$	$\Sigma y = 1960$	$\Sigma dy = -9$	$\Sigma dy^2 = 55$	$\Sigma dxdy = 50$

$$r = \frac{N \cdot \sum dxdy - (\sum dx)(\sum dy)}{\sqrt{N \cdot \sum dx^2 - (\sum dx)^2} \times \sqrt{N \cdot \sum dy^2 - (\sum dy)^2}}$$

$$\begin{aligned}
 r &= \frac{(5 \times 50) - (-10 \times -9)}{\sqrt{5 \times 50 - (-10)^2} \times \sqrt{5 \times 55 - (-9)^2}} \\
 &= \frac{250 - 90}{\sqrt{250 - 100} \times \sqrt{275 - 81}} \\
 &= \frac{160}{\sqrt{150} \times \sqrt{194}} \\
 &= \frac{160}{1.07 \times 13.92} \\
 &= \frac{160}{12.25 \times 13.92} \\
 &= \frac{160}{170.52} = 0.94
 \end{aligned}$$

Illustration 9.38

Calculate the coefficient of correlation between the deposit and advances of a private sector banking in India over a period of time.

Deposit (Rs in crores)	1575	2137	2537	3090	3615	4180	5121	5911	6672
Advances (Rs in crores)	956	1155	1047	1807	2254	2460	3344	4023	4620

Solutions

Calculation of co-efficient of correlation

Deposit	$dx = x - A$ (3615)	dx^2	y	$dy = x - A$ (2254)	dy^2	$dxdy$
1575	-20404	4161600	956	-1298	1684804	2647920
2137	-1478	2184484	1155	-1099	1207801	1624322
2537	-1078	1162084	1047	-1207	1456849	1301146
3090	-525	275625	1807	-447	199809	234675
3615	0	0	2254	0	0	0
4180	565	319225	2460	206	42436	116390
5121	1506	2268036	3344	1090	1188100	1641540
5911	2296	5271616	4023	1769	3129361	4061624
6672	3057	9345249	4620	2366	5597956	7232862
	$\Sigma dx = 2303$	$\Sigma dx^2 = 24987917$	$\Sigma y = 1960$	$\Sigma dy = 1380$	$\Sigma dy^2 = 14507116$	$\Sigma dxdy = 18860479$

$$r = \frac{N \cdot \sum dxdy - (\sum dx)(\sum dy)}{\sqrt{N \cdot \sum dx^2 - (\sum dx)^2} \times \sqrt{N \cdot \sum dy^2 - (\sum dy)^2}}$$

$$\begin{aligned}
 r &= \frac{(9 \times 18860479) - (-2303 \times 1380)}{\sqrt{(9 \times 24987919)} - (-2303)^2 \times \sqrt{(9 \times 14507116) - (-1380)^2}} \\
 &= \frac{169744311 - 3178140}{\sqrt{(224891271 - 5303809)} \sqrt{(130564044 - 1904400)}} \\
 &= \frac{166566171}{\sqrt{219587462} \times 128659644} \\
 &= \frac{166566171}{14818.48 \times 11342.82} \\
 &= \frac{166566171}{168083351} \\
 r &= 0.99
 \end{aligned}$$

There is a high degree positive correlation between the deposit and Advances.

Illustration 9.39

Calculate Karl Pearson's co-efficient of correlation taking deviation from actual mean. (x series = 26) (y series = 22)

x	22	23	23	24	26	27	?	28	30	30
y	18	20	21	20	?	22	23	24	25	26

Solutions

With the help of mean, we can calculate the missing value of x and y .

$$\begin{aligned}
 \bar{X} &= \frac{\sum x}{N} = 26 & \sum x &= \frac{\sum x}{10} \\
 260 &= \sum x \\
 \sum x &= 260 \\
 \text{Total of } 9 &= \frac{233}{27} \\
 y &= \frac{\sum y}{N}, \quad 22 = \frac{\sum y}{10} \\
 220 &= \sum y \\
 \sum y &= 220 \\
 \text{Total of } 9 &= \frac{199}{21}
 \end{aligned}$$

Solutions

x	x (x - x̄)	x²	y	y(y - ȳ)	y²	xy
22	-4	16	18	-4	16	16
23	-3	9	20	-2	4	6
23	-3	9	21	-1	1	3
24	-2	4	20	-2	4	4
26	0	0	21	-1	1	0
27	1	1	22	0	0	0
27	1	1	23	1	1	1
28	2	4	24	2	4	4
30	4	16	25	3	9	12
30	4	16	26	4	16	16
$\Sigma x^2 = 76$			$\Sigma y^2 = 56$			$\Sigma xy = 62$

$$\begin{aligned}
 r &= \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \\
 &= \frac{62}{\sqrt{76} \times \sqrt{56}} \\
 &= \frac{62}{8.72 \times 7.48} \\
 &= \frac{62}{65.24} = +0.95 \\
 r &= +0.95
 \end{aligned}$$

There is a high degree positive correlation between x and y .

SUMMARY

Correlation

Analysis of the relationship of two or more variables.

Types of Correlation

- Positive or Negative correlation
- Simple, Partial and multiple correlation
- Linear and non-linear correlation.

Positive Correlation

If the increase in one variable influence the increase in the corresponding variables.

Negative Correlation

If the increase in one variable decreases the value of other variables.

Simple Correlation

Means association of only two variables.

Partial Correlation

Means the correlation of two variables keeping third variables as constant.

Multiple Correlation

Means study of correlation among all variables.

Linear Correlation

If the ratio of change between two set of variables is same.

Non-linear Correlation

If the rate of change in one variable does not bear the constant ratio to the rate of change in the other variables.

Methods of Correlation Analysis

- **Graphic method**
 - ◆ Scatter Diagram method
 - ◆ Graphic method.
- **Statistical method**
 - ◆ Karl Pearson's coefficient of correlation
 - ◆ Rank method
 - ◆ Concurrent Deviation method
 - ◆ Method of least squares

Probable Error

The technique of probable error should be applied to know the reliability of the coefficient of correlation.

Co-efficient of Determination (r^2)

The square of the coefficient of correlation is called co-efficient of determination. It is calculated through the ratio of explained variance to total variance.

FORMULAE

Karl Pearson's co-efficient of correlations

1. When deviations are taken from actual mean

$$r = \frac{\sum xy}{N \sigma x \sigma y}$$

or

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

r = Karl Pearson's co-efficient of correlation

$$x = (x - \bar{x})$$

$$y = (y - \bar{y})$$

N = Number of pairs of observations

σ_x = Standard deviation of X services

σ_y = Standard deviation of Y services

2. When deviations are taken from assumed mean

$$r = \frac{N \cdot \sum dxdy - (\sum dx)(\sum dy)}{\sqrt{N \cdot \sum dx^2 - (\sum dx)^2} \times \sqrt{N \cdot \sum dy^2 - (\sum dy)^2}}$$

$$dx = (X - A)$$

$$dy = (Y - A)$$

3. Bivariate frequency distribution

$$r = \frac{N \sum f dxdy - (\sum f dx)(\sum f dy)}{\sqrt{N \sum f dx^2 - (\sum f dx)^2} \times \sqrt{N \sum f dy^2 - (\sum f dy)^2}}$$

4. When we take actual values of X and Y

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

5. Spearman's rank correlation co-efficient

$$r = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

$$\text{or } r = 1 - \frac{6 \sum D^2}{N^3 - N}$$

r = Spearman's rank correlation

D = Difference of Rank

N = Number of pairs of observations

6. When rank are repeated

$$R = 1 - \frac{\left[6 \sum D^2 + 1/12(m^3 - m) + \frac{1}{12}(m^3 - m) \right]}{N^3 - N}$$

M = The number of times, the values are repeated or the ranks are common

7. Concurrent Deviation

$$rc = \pm \sqrt{\pm \frac{2C - N}{N}}$$

rc = co-efficient of correlation by the concurrent deviation method

c = Number of concurrent deviations or the number of positive signs

N = Number of pairs of deviations compared, i.e., $(N - 1)$

8. Probable Error

$$P.E. = 0.6745 \frac{1-r^2}{\sqrt{N}}$$

$$S.E. = \frac{1-r^2}{\sqrt{N}}$$

P.E. = Probable Error

r = correlation

n = number of pairs at observation.

S.E. = Standard Error

9. Co-efficient of determination = r^2

10. Co-efficient of non-Determination = $1 - r^2$

EXERCISES

(a) Choose the best option

1. An analysis of the relationship of two or more variables is
 - (a) Correlation
 - (b) Regression
 - (c) Kurtosis
2. Decrease in one variable influences the decrease in other variable is
 - (a) Positive correlation
 - (b) Negative correlation
 - (c) Simple correlation
3. When decrease in one variable increases the other variables then it is
 - (a) Positive correlation
 - (b) Negative correlation
 - (c) Linear correlation
4. If the ratio of change between two sets of variables is same, then it is called
 - (a) Linear correlation
 - (b) Non-Linear correlation
 - (c) Negative correlation

- 5.** Curvilinear correlation is
 (a) Linear correlation
 (b) Non-Linear correlation
 (c) Simple correlation
- 6.** Perfect Negative correlation is
 (a) $r = -1$ (b) $r = +1$ (c) $r = 0$
- 7.** Perfect Positive correlation is
 (a) $r = -1$ (b) $r = +1$ (c) $r = 0$
- 8.** Absence of correlation is
 (a) $r = -1$ (b) $r = +1$ (c) $r = 0$
- 9.** Rank co-efficient correlation is
 (a) $r_R = 1 - \frac{6 \sum D^2}{N^3 - N}$ (b) $r_R = 1 + \frac{6 \sum D^2}{N^3 - N}$ (c) $r_R = \frac{6 \sum D^2}{N^3 - N}$
- 10.** Co-efficient of concurrent correlation is
 (a) $r_c = \pm \sqrt{\frac{2c - N}{N}}$ (b) $r_c = \sqrt{\frac{2c + N}{N}}$ (c) $r_c = \frac{2c - N}{N}$
- 11.** Co-efficient of determination is
 (a) r^2 (b) r^3 (c) r
- 12.** Co-efficient of non-determination is
 (a) $1 - r^2$ (b) $1 + r^2$ (c) $1 > r^2$
- 13.** Standard error is
 (a) $\frac{1 - r^2}{\sqrt{n}}$ (b) $\sqrt{\frac{1 + r^2}{n}}$ (c) $\frac{1 + r^2}{n}$
- 14.** Probable error is
 (a) $0.6745 \times S.E.$ (b) $0.6745 + S.E.$ (c) $0.6745 - S.E.$

Answers

- | | | | |
|----------------|----------------|----------------|----------------|
| 1. (a) | 2. (a) | 3. (b) | 4. (a) |
| 5. (b) | 6. (a) | 7. (b) | 8. (c) |
| 9. (a) | 10. (a) | 11. (a) | 12. (a) |
| 13. (a) | 14. (a) | | |

(b) Theoretical Questions

- Define correlation. What are the properties of the co-efficient of correlation?
- Explain the meaning and significance of the concept of correlation.
- What are the different types of correlation?
- Distinguish between co-efficient of correlation from co-efficient of variation.

5. What are the different measures of correlation?
6. What is a scatter diagram?
7. Define Karl Pearson's co-efficient of correlation. Explain the special characteristics of it.
8. What are the merits and demerits of Karl Pearson's correlation method?
9. Define Rank correlation. Explain the method of calculating Spearman's rank correlation co-efficient.
10. What is Probable Error?
11. What is Co-efficient of determination? Why is it computed?
12. What is concurrent deviation method?
13. Distinguish between
 1. Positive and Negative Correlation
 2. Linear and Non-Linear Correlation
 3. Simple, Partial and Multiple correlation
14. What is spurious or non-sense correlation? Explain with an example.
15. Interpret the r in the following cases.
 $r = 0.92, r = -0.82, r = +1, r = -1, r = 0.$

(c) Practical Problems

16. From the following data, calculate the co-efficient of correlation between age of students and their playing habit:

Age (in Years)	14	15	16	17	18
No. of students	300	250	200	180	150
Regular Players	200	130	125	110	100

Answer $r = +0.23$

17. Draw a scatter diagram for the following distribution:

x	5	10	15	20	25	30
y	15	17	19	16	14	28

18. The height and weight of 10 students in I M.Com. class of a college is as follows:

Height (cm)	161	162	163	164	165	166	167	168	169	170
Weight (kg)	48	50	52	47	53	56	46	54	55	58

Answer $r = +0.61$

Draw a scatter diagram and find out whether the correlation is positive or negative.

19. The average income and expenditure of staff in an office is given below:

Months	Jan.	Feb.	Mar.	Apr.	May.	June	July
Income (Rs)	5000	6500	7200	8000	9500	9800	10000
Expenses (Rs)	4500	5900	7000	6900	7500	7700	7900

Calculate the co-efficient of correlation

Answer $r = 0.953$

20. From the following data of the marks obtained by 8 students in the Accountancy and statistics papers, compute the rank co-efficient of correlation:

Marks in Accountancy	15	20	28	12	40	60	20	80
Marks in Statistics	40	30	50	30	20	10	30	60

(B.Com., MKU)

21. From the following data, find the rank correlation:

A	115	109	112	87	98	98	120	100	98	118
B	75	73	85	70	76	65	82	73	68	80

(B.Com., MKU, BDU)

22. From the following data, calculate the co-efficient of correlation between age and playing habit of the following students:

Age	15	16	17	18	19	20
No. of students	250	200	150	120	100	80
Regular players	200	150	90	48	30	12

Answer $R = -0.991$

23. Calculate Karl Pearson's co-efficient of correlation and its probable error from the following data of imports and exports:

Year	Value of Imports	Value of Exports
1994	903	620
1995	1036	625
1996	904	573
1997	961	640
1998	1122	642
1999	1092	661
2000	1131	685
2001	1190	793
2002	1314	816
2003	1350	805

Answer $r = +0.9173$, P.E. = 0.0338

(B.Com., CHU)

24. Calculate r from the following data:

Demand	100	60	40	30	24	11	6	3
Supply	55	40	40	36	22	18	15	40

Answer $r = -0.0898$

25. Calculate the co-efficient of correlation from the following data by the method of rank differences:

Rank of x	5	8	6	9	5	10	4	2
Rank of y	5	8	5	9	4	10	6	2

Answer $r = 0.827$ (B.Com., BDU)

26. From the following data calculate the rank correlation co-efficient after making adjustment for tied ranks:

x	48	33	40	9	16	16	65	24	16	57
y	13	13	24	6	15	4	20	9	6	19

Answer $r = 0.733$ (B.Com., BU)

27. Calculate from the data reproduced below pertaining to 66 selected villages in a district, the value of r between total cultivable area and the area under paddy:

Area under paddy (in acres)	Total Cultivable area (in acres)					Total
	0 – 500	500 – 1000	1000 – 1500	1500 – 2000	2000 – 2500	
0–200	12	6	–	–	–	18
200–400	2	18	4	2	1	27
400–600	–	4	7	3	–	14
600–800	–	1	–	2	1	4
800–1000	–	–	–	1	2	3
Total	14	29	11	8	4	66

Answer $r = 0.746$

28. The following table gives the distribution of the ages in years of husbands and wives in 100 couples:

Age of wives	Age of husbands				
	20–25	25–30	30–35	35–40	
15–20	20	10	3	2	
20–25	4	28	6	4	
25–30	–	5	11	–	
30–35	–	–	2	–	

Age of wives		20–25	25–30	30–35	35–40				
30–35	—	—	2	—	—				
35–40	—	—	—	5	—				
Calculate the co-efficient of correlation between the age of the husbands and the age of the wives.									
Answer $r = 0.6133$		(B.Com., MKU, BDU, BU)							
29. If the value of r is = 0.9 and its probable error is 0.0128, what would be the value of n ?									
Answer $n = 100$									
30. Compute Karl Pearson's co-efficient of correlation in the following series relating to cost of living and wages:									
Wages	100	101	103	102	100				
Cost of Living	98	99	99	97	95				
Answer. $r = + 0.85$									
31. Calculate the Pearson's co-efficient of correlation from the following data using 44 and 26 respectively as the origin of x and y :									
X	43	44	46	40	44				
y	29	31	19	18	19				
Answer $r = - 0.777$ (B.Com., MKU, CHU)									
32. Find out the co-efficient of correlation between output and cost of the automobile factory from the following data:									
Output 10 of cars ('000)	3.5	4.2	5.6	6.5	7.0				
Cost per car (Rs.'000)	9.8	9.4	8.8	8.4	8.3				
Answer $r = - 0.938$									
33. Calculate Karl Pearson's co-efficient of correlation between expenditure on advertising and sales from the data given below:									
Advertising Expenses ('000s)	39	65	62	90	82				
Sales (Lakhs) (Rs)	47	53	58	86	62				
Answer $r = + 0.7804$ (B.Com., MKU, BDU)									
34. Calculate the co-efficient of correlation between x and y series from the following data:									

	x series	y series									
No. of pairs of observation	15	15									
Arithmetic Mean	25	18									
Standard Deviation	3.01	3.03									
Sum of Squares of deviation	136	138									
Summation of product deviations of x and y series from their respective Arithmetic Mean = 122.											
Answer $r = + 0.89$		(B.Com., MKU, BU)									
35. The co-efficient of correlation between two variables x and y is 0.64. Their covariance is 16. The variance of x is 9. What is the standard deviation of y series?											
36. Two ladies were asked to rank 7 different types of lipsticks. The ranks given by them are given below:											
Lipsticks	A	B	C	D	E	F	G				
Surya	2	1	4	3	5	7	6				
Rama	1	3	2	4	5	6	7				
Calculate Spearman's rank correlation co-efficient.											
Answer $r = + 0.786$		(B.Com., MKU, BU, BDU)									
37. Calculate the co-efficient of correlation by concurrent deviation method:											
Price	1	4	3	5	5	8	10	10	11	15	
Demand	100	80	80	60	58	50	40	40	35	30	
Answer $r = + 0.89$											
38. Calculate the co-efficient of concurrent deviation from the following data.											
x	80	45	50	56	58	60	65	68	70	75	85
y	82	56	50	48	60	62	64	64	70	74	90
Answer $r = - 0.577$		(B.A., MKU)									
39. The following statistics relate to income and expenditure of 11 workers of a factory. Find the correlation co-efficient by concurrent deviation method.											
Income	54	43	40	65	55	73	58	49	62	60	53
Expenditure	50	48	45	56	52	60	57	54	58	55	58
Answer $r = + 0.89$		(B.Com., CHU, BDU)									
40. Find the r from the following data:											
x	2		4		6		8		10		
y		1		2		3		4		5	
Answer $r = + 1$											

41. Find the r from the following data:

Mark in Science	Mark in Maths			Total
20 – 40	–	–	–	0
40 – 60	1	4	2	7
60 – 80	1	4	10	6
80 – 100	–	–	3	9
	2	8	15	15
				40

42. Calculate r from the following data:

Mark in Science	Mark in Maths			Total
20 – 40	–	–	–	0
40 – 60	–	4	3	7
60 – 80	–	3	11	4
80 – 100	–	–	7	8
	0	7	21	12
				40

10

CHAPTER

REGRESSION ANALYSIS

10.1 MEANING

Regression means stepping back or going back. It was first used by Francis Galton in 1877. He studied the relationship between the height of father and their sons. The study revealed that

- (i) Tall fathers have tall sons and short fathers have short sons.
- (ii) The mean height of the sons of tall father is less than mean height of their fathers.
- (iii) The mean height of sons of short fathers is more than the mean height of their fathers.

The tendency to going back was called by Galton as ‘Line of Regression’. This line describing the average relationship between two variables is known as the Line of Regression.

Regression is a statistical technique through which the estimation of unknown variable from the known variable can be done. The known variable which is used to estimate an unknown variable is called an independent variable or explanatory variable. The unknown variable or the variable for which the value is to be predicted is called ‘Dependent Variable’ or ‘Explained Variable’.

10.2 DEFINITIONS

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data —**Blair**

One of the most frequently used techniques in Economics and Business research, to find a relations between two or more variables that are related casually, is regression analysis —**Taro Yamane**

Regression Analysis attempts to establish the nature of relationship between variables that is to study the function relationship between the variables and thereby provide a mechanism for prediction or forecasting.

10.3 USES OF REGRESSION ANALYSIS

1. Regression analysis studies the relationship of two or more variables.
2. It helps to find out the value of dependent variables from the values of independent variables.
3. It is useful for calculating co-efficients of correlation (r) and co-efficients of determination (r^2).
4. It is widely useful for quality control in corporate sector.
5. It is useful for estimation of statistical curve for demand, supply, price consumption and cost.

Differences between Correlation and Regression

Correlation	Regression
1. Correlation is the relationship between two or more variables.	1. Regression is the mathematical measure showing the average relationship between two variables.
2. It is the measure of degree of relationship between two variables.	2. It is the measure of nature of relationship.
3. It does not study the cause and effect of the variables.	3. It indicates the cause and effect relationship between the variables and establish a functional relationship.
4. The co-efficients of correlation is a relative measure. The range lies between ± 1 .	4. Regression co-efficient is an absolute figure. If we know the value of independent variable, we can find out the value of the dependent variable.
5. It is not very much useful for further algebraic treatment.	5. It is very much useful for further algebraic treatment.

10.4 REGRESSION LINES

A regression line is a straight line fitted to the data by the method of least squares. There are two regression lines for two set of variables. The two regression lines are the Regression line of X on Y and the Regression line of Y on X . The line of regression gives the best average value of one variable for any given value of the other variable. If there exists perfect positive correlation between the two variables then the two regression lines will coincide with each other. In that case, there will be only one line. If the degree of correlation is very high, the two lines are nearer to each other. If the degree is very low, the two regression lines are far away from each other.

10.5 REGRESSION EQUATIONS

Regression equations are algebraic expressions of regression lines. As there are two regression lines, there are two regression equations. For the two variables X and Y , there are two regression equations. They are:

- (i) Regression equation of X on Y
- (ii) Regression equation of Y on X .

10.5.1 Regression Equation of X on Y

The straight line equation is

$$X = a + by$$

Here a and b are unknown constants, which determines the position. The constant a is the intercept on the other value; the constant b is the slope.

The following two normal equations are derived:

$$\begin{aligned}\sum x &= na + b \sum y \\ \sum xy &= a \sum x + b \sum y^2\end{aligned}$$

The Regression equation X on Y is used to find out the values of X for given value of Y .

10.5.2 Regression Equation of Y on X

The straight line equation is

$$Y = a + bx$$

The following two normal equations are derived:

$$\begin{aligned}\sum y &= na + b \sum x \\ \sum xy &= a \sum x + b \sum x^2\end{aligned}$$

The Regression equation Y on X is used to ascertain the value of y for a given value of x .

Illustration 10.1

Find out the regression equation, x on y and y on x from the following data:

x	15	20	25	30	35	40	45
y	8	14	20	26	32	38	44

Solutions

x	y	x^2	y^2	xy
15	8	225	64	120
20	14	400	196	280
25	20	625	400	500
30	26	900	676	780
35	32	1225	1024	1120
40	38	1600	1444	1520
45	44	2025	1936	1980
$\sum x = 210$		$\sum y = 182$	$\sum x^2 = 7000$	$\sum y^2 = 5740$
				$\sum xy = 6300$

$$\sum x = 210; \sum y = 182; \sum x^2 = 7000; \sum y^2 = 5740; \sum xy = 6300$$

Regression equation x on y is $y = -a + bx$

Hence,

$$\begin{aligned}\sum y &= Na + b \sum y \\ \sum xy &= a \sum y + b \sum y^2\end{aligned}$$

$$210 = 7a + 182b \quad (1)$$

$$6300 = 182a + 5740b \quad (2)$$

Multiplying Eq. (1) by 26

$$5460 = 182a + 4732b \quad (3)$$

$$6300 = 182a + 5740b \quad (4)$$

Deducting (3) from Eq. (2)

$$6300 = 182a + 5740b \quad (4)$$

$$5460 = 182a + 4732b \quad (3)$$

$$\begin{array}{r} (-) \quad (-) \quad (-) \\ \hline 840 = 0 \quad + \quad 1008b \end{array}$$

Therefore,

$$b = \frac{840}{1008} = 0.83$$

Substituting the value of b in Eq. (1)

$$210 = 7a + (182 \times 0.83)$$

$$210 = 7a + 151.06$$

$$7a + 151.06 = 210$$

$$7a = 210 - 151.06$$

$$7a = 58.94$$

$$a = 8.42$$

Hence,

$$x = a + by$$

$$x = 8.42 + 0.83y$$

Regression Eq. of y on x :

Hence,

$$y = a + bx$$

$$\sum y = Na + b \sum y^2 \quad (1)$$

$$182 = 7a + 210b \quad (1)$$

$$6300 = 210a + 7000b \quad (2)$$

Multiplying Eq. (1) by 30

$$5460 = 210a + 6300b \quad (3)$$

$$6300 = 210a + 7000b \quad (4)$$

Deducting Eq. (4) from Eq. (3)

$$6300 = 210a + 7000b \quad (4)$$

$$\begin{array}{r} 5460 = 210a + 6300b \\ \hline 840 = 0 \quad + 700b \end{array} \quad (3)$$

$$700b = 840$$

$$b = \frac{840}{700} = 1.2$$

Substituting the value of b in Eq. (1)

$$182 = 7a + (210 \times 1.2)$$

$$\begin{aligned}
 182 &= 7a + 252 \\
 7a + 252 &= 182 \\
 7a &= 182 - 252 \\
 7a &= -70 \\
 a &= -10 \\
 \text{Therefore, } y &= -10 + 1.2x
 \end{aligned}$$

10.5.3 Calculation of Regression Equations through Arithmetic Mean Method

The calculation of regression equation through the simple linear equation is not easy when the variables are large and complicated. Hence, the regression equations can be calculated through arithmetic mean method for such kind of variables. The following are the two regression equations under this method.

1. Regression Equation x on y

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

where \bar{x} indicates arithmetic mean of x series, \bar{y} indicates Arithmetic mean of y series.

$\frac{\sigma_x}{\sigma_y}$ indicates regression co-efficient of x on y . $\frac{\sigma_x}{\sigma_y}$ may be denoted as b_{xy} .

$$\text{Hence, } b_{xy} = r \frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2}$$

2. Regression Equation y on x

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

where $\frac{\sigma_y}{\sigma_x}$ indicates regression co-efficient of y on x . \bar{x} indicates arithmetic mean of x series and \bar{y} indicates arithmetic mean of y series. $\frac{\sigma_y}{\sigma_x}$ may be called as b_{yx} .

$$\text{Hence, } b_{yx} = r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2}$$

Illustration 10.2

Find out the regression equations, x on y and y on x from the following data.

x	10	20	30	40	50	60	70
y	11	25	35	43	60	67	74

Solutions

Calculation of Regression Equations

x	y	X(x - x) x - 40 = x	y(y - y) y - 45 = y	x²	y²	xy
10	11	-30	-34	900	1156	110
20	25	-20	-20	400	400	500
30	35	-10	-10	100	100	1050
40	43	0	-2	0	4	1720
50	60	10	15	100	225	3000
60	67	20	22	400	484	4020
70	74	30	29	900	841	5180
$\sum x = 280$		$\sum y = 315$		$\sum x^2 = 2800$		$\sum y^2 = 3210$
						$\sum xy = 15580$

$$\bar{x} = \frac{\sum y}{N} = \frac{280}{7} = 40$$

$$\bar{y} = \frac{\sum y}{N} = \frac{315}{7} = 45$$

Regression Equation of x on y:

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2} = \frac{15580}{3210} = 4.85$$

Hence,

$$y - 40 = 4.85 (y - 45)$$

$$x - 40 = 4.85y - 218.25$$

$$x = 4.85y - 218.25$$

$$x = -178.25 + 4.85y$$

Regression Equation of y on x:

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (\bar{x} - x)$$

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum y^2} = \frac{15580}{2800} = 5.56$$

Hence,

$$y - 45 = 5.56 (x - 40)$$

$$y - 45 = 222.4 + 45$$

$$y = -177.4 + 5.56x$$

10.5.4 Calculation of Regression Equation on the Basis of Assumed Mean

The calculation of regression equation is further simplified, if the deviations are taken from the assumed mean. The formula for the calculation of regression equation under this method is as follows:

(i) **Regression equation of x on y**

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}) = \frac{\sigma_x}{\sigma_y}$$

can be calculated through the following formula:

(a) **For Individual Observations:**

$$\frac{\sigma_x}{\sigma_y} = \frac{\sum dx dy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

where, dx = variables, x – Assumed mean, $(x - A)$; dy = variables, y – Assumed mean, $(y - A)$; N = Total number of pairs of variables x and variables y .

(b) **For Discrete Series:**

$$\frac{\sigma_x}{\sigma_y} = \frac{\sum f dx dy - \frac{\sum f dx \times \sum f dy}{N}}{\sum f dy^2 - \frac{(\sum f dy)^2}{N}}$$

Here, f indicates frequency.

(c) **For Continuous Series:**

$$\frac{\sigma_x}{\sigma_y} = \frac{\sum f dx dy - \frac{\sum f dx \times \sum f dy}{N}}{\sum f dy^2 - \frac{(\sum f dy)^2}{N}} \times \frac{ix}{iy}$$

where, ix indicates differences of the class intervals of x series; iy indicates differences of the class intervals of y series;

(ii) **Regression equation of y on x**

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) = \frac{\sigma_y}{\sigma_x}$$

(a) For Individual Observations:

$$\frac{\sigma_y}{\sigma_x} = \frac{\sum dx dy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

(d) For Discrete Series:

$$\frac{\sigma_y}{\sigma_x} = \frac{\sum f dx dy - \frac{\sum f dx \times \sum f dy}{N}}{\sum f dx^2 - \frac{(\sum f dy)^2}{N}}$$

where, f indicates frequency.

(c) For Continuous Series:

$$\frac{\sigma_y}{\sigma_x} = \frac{\sum f dx dy - \frac{\sum f dx \times \sum f dy}{N}}{\sum f dx^2 - \frac{(\sum f dx)^2}{N}} \times \frac{ix}{iy}$$

where, ix indicates differences of the class intervals of x series; iy indicates differences of the class intervals of y series.

Illustration 10.3

Find out the regression equations from the following data:

x	17	22	29	31	37	42	51
y	8	25	14	27	13	17	37

Solutions

Calculation of Regression Equations

x	y	$dx(x-x)$ $x-30$	$dy(y-y)$ $y-20$	dx^2	dy^2	$dxdy$
17	11	-13	-12	169	144	156
22	25	-8	5	64	25	-40
29	35	-1	-6	1	36	6
31	43	1	7	1	49	7
37	60	7	-7	49	49	-49
42	67	12	-3	144	9	-36
51	74	21	17	441	289	357
$\sum dx = 19$		$\sum dy = 1$	$\sum dx^2 = 869$	$\sum dy^2 = 601$	$\sum dxdy = 401$	

$$\bar{x} = A \pm \frac{\sum dx}{N} = 30 + \frac{19}{7} = 32.71$$

$$\bar{y} = A \pm \frac{\sum dy}{N} = 20 + \frac{1}{7} = 20.14$$

$$\frac{\sigma_x}{\sigma_y} = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

$$= \frac{401 - \frac{19 \times 1}{7}}{601 - \frac{(1)^2}{7}}$$

$$= \frac{401 - 2.71}{601 - 0.14} = \frac{398.29}{600.86} = 0.66$$

$$\frac{\sigma_y}{\sigma_x} = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

$$= \frac{401 - \frac{19 \times 1}{7}}{869 - \frac{(19)^2}{7}}$$

$$= \frac{401 - 2.71}{869 - 51.57} = \frac{398.29}{817.43} = 0.49$$

Regression Equation of x on y :

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

i.e.,

$$y - 32.71 = 0.66 (y - 20.14)$$

Regression Equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

Hence,

$$y - 20.14 = 0.49 (x - 32.71)$$

Regression Equation of y on x :

$$y - 20.14 = 0.49 (x - 32.71)$$

$$y - 20.14 = 0.49 x - 16.03$$

$$y = 0.49 x - 16.03 + 20.14$$

$$y = 4.11 + 0.49x$$

10.6 STANDARD ERROR OF REGRESSION

Standard error of regression is otherwise called as standard error of estimate. It is a statistical technique to know the reliability of the regression analysis. It is purely an estimate. It may not give the accurate date for the estimate. The prediction made by the regression analysis can be tested through the standard error of estimate. It can be ascertained through the following formula:

10.6.1 Standard Error of Regression of x Value from y

$$yx = \frac{\sqrt{\sum(y - yx)^2}}{N}$$

or

$$\sum yx = y\sqrt{1 - r^2}$$

It may be otherwise called as = $\frac{\sqrt{\text{Unexplained Variation}}}{N}$

10.6.2 Standard Error or Estimate may Otherwise be Calculated as

$$\sum xy = \frac{\sqrt{\sum y^2 - a \sum y - b \sum xy}}{N}$$

10.6.3 Standard Error of Regression of x Value from x_c

$$\sum xy = \frac{\sqrt{\sum(x - xc)^2}}{N}$$

or

$$\sum yx = y\sqrt{1 - r^2}$$

It may be otherwise called as = $\frac{\sqrt{\text{Unexplained Variation}}}{N}$

It may otherwise be calculated through the following formula

$$\sum xy = \frac{\sqrt{\sum x^2 - a \sum x - b \sum xy}}{N}$$

10.6.4 Standard Error of Regression of y Value from y_c

$$\sum yx = \frac{\sqrt{\sum(y - yc)^2}}{N}$$

or

$$\sum yx = y\sqrt{1-r^2}$$

It may otherwise be calculated through the following formula.

$$\sum yx = \frac{\sqrt{\sum y^2 - a \sum y - b \sum xy}}{N}$$

Illustration 10.4

From the following data, find the two regression equations and calculate the standard error of the estimate:

x	15	17	22	21	20
y	14	15	18	13	15

Solutions

Calculation of Regression Equations

x	(x - x̄)	x²	y	y(y - ȳ)	y²	xy
15	-4	16	14	-1	1	4
17	-2	4	15	0	0	0
22	3	9	18	3	9	9
21	2	4	13	-2	4	-4
20	1	1	15	0	0	0
$\sum x = 45$		$\sum x^2 = 34$	$\sum y = 25$	$y = 0$	$\sum y^2 = 14$	$\sum xy = 9$

Regression Equation of x on y:

$$x - \bar{x} = \frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2} = \frac{9}{14} = 0.26$$

Regression Equation of y on x:

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2} = \frac{9}{34} = 0.26$$

Regression Equation of x on y:

$$(x - 9) = 0.64(y - 5)$$

$$x - 9 = 0.64y - 3.2$$

$$x = 0.64y - 3.2 + 9$$

$$x = 5.8 + 0.64y$$

Regression Equation of y on x :

$$\begin{aligned}
 (y - 5) &= 0.26(x - 9) \\
 y - 5 &= 0.26x - 2.34 \\
 y &= 0.26x - 2.34 + 5 \\
 y &= 2.66 + 0.26x
 \end{aligned}$$

Calculation of standard error of estimate

$$\begin{aligned}
 y &= 2.66 + 0.26x \\
 x &= 5.8 + 0.64y
 \end{aligned}$$

x	y	yc	xc	$(y - yc)^2$	$(x - xc)^2$
15	14	6.56	14.76	55.3536	0.0576
17	15	7.08	15.40	62.7264	2.56
22	18	8.38	17.32	92.5444	21.9024
21	13	8.12	14.12	23.8144	47.3344
20	15	7.86	15.40	50.9796	21.16
				$(y - yc)^2$	$(x - xc)^2$
				= 285.4184	= 93.0144

Calculations:

$$\begin{aligned}
 y &= 2.66 + 0.26x \\
 (x - 15)y &= 2.66 + 0.26 \times 15 \\
 y &= 2.66 + 3.9 \\
 yc &= 6.56 \\
 (x - 17)y &= 2.66 + 0.26 \times 17 \\
 &= 2.66 + 4.42 \\
 &= 7.08 \\
 (x - 22)y &= 2.66 + 0.26 \times 22 \\
 &= 2.66 + 5.72 \\
 &= 8.38 \\
 (x - 21)y &= 2.66 + 0.26 \times 21 \\
 &= 2.66 + 5.46 \\
 &= 8.12 \\
 (x - 20)y &= 2.66 + 0.26 \times 20 \\
 &= 2.66 + 5.2 \\
 &= 7.86 \\
 x &= 5.8 + 0.64y \\
 (y - 14)x &= 5.8 + 0.64 \times 14 \\
 &= 5.8 + 8.96 \\
 xc &= 14.76 \\
 (y - 15)x &= 5.8 + 0.64 \times 15 \\
 &= 5.8 + 9.6 \\
 &= 15.4 \\
 (y - 18)x &= 5.8 + 0.64 \times 18 \\
 &= 5.8 + 11.52 \\
 &= 17.32
 \end{aligned}$$

$$\begin{aligned}
 (y - 13)x &= 5.8 + 0.64 \times 13 \\
 &= 5.8 + 8.32 \\
 &= 14.12 \\
 (y - 15)x &= 5.8 + 0.64 \times 15 \\
 &= 5.8 + 9.6 \\
 &= 15.4
 \end{aligned}$$

Standard Error of regression

$$\begin{aligned}
 \sum xy &= \frac{\sqrt{\sum(x - xc)^2}}{N} \\
 &= \frac{\sqrt{93.0144}}{5} \\
 &= \sqrt{18.60288} \\
 \sum xy &= 4.31
 \end{aligned}$$

Standard Error of regression of y value from y :

$$\begin{aligned}
 \sum yx &= \frac{\sqrt{\sum(y - yc)^2}}{N} \\
 &= \frac{\sqrt{285.4184}}{5} \\
 &= \sqrt{57.08368} \\
 \sum yx &= 7.56 \\
 \sum yx &= \sigma_x \sqrt{1 - r^2} \\
 \sum x^2 &= 34; \sum y^2 = 14; N = 5; \\
 r &= \sqrt{bxy \times byx}, bxy \text{ or } r \frac{x}{y} = 0.64 \\
 &= \sqrt{0.64 \times 0.26}, byx \text{ or } r \frac{y}{x} = 0.26 \\
 &= \sqrt{0.1664} = 0.41 \\
 \sigma_x &= \frac{\sqrt{\sum x^2}}{n} = \sqrt{\frac{34}{5}} = \sqrt{6.8} = 2.61 \\
 \sum xy &= \sigma_x \sqrt{1 - r^2} \\
 \sum xy &= 2.61 \times \sqrt{1 - (0.41)^2} \\
 &= 2.61 \times \sqrt{1 - 0.1681} \\
 &= 2.61 \times \sqrt{0.8319}
 \end{aligned}$$

$$\begin{aligned}
 &= 2.61 \times 0.91 \\
 \sum_{yx} &= \sigma_y \sqrt{1 - r^2} \\
 \sigma_y &= \sqrt{\frac{\sum y^2}{n}} = \sqrt{14/5} = \sqrt{2.8} = 1.67 \\
 \sum_{yx} &= 1.67 \times 0.91 = 1.52
 \end{aligned}$$

10.6.5 Regression Equations in Bi-variate Grouped Distribution

Regression Equations for Bi-variate Grouped Frequency Distribution can be calculated by preparing the Bi-variate Correlation table. This correlation table would be used to find out the sum of deviations taken from assumed mean and the sum of the squares of the deviation etc. which are required for calculating the regression equation. The following formula can be applied to find out the regression equations:

Regression co-efficient of x on y

$$b_{xy} = \frac{N \sum f dx dy - \sum f dx \times \sum f dy}{N \sum f dy^2 - (\sum f dy)^2} \times \frac{ix}{iy}$$

Regression co-efficient of y on x

$$b_{yx} = \frac{N \sum f dx dy - \sum f dx \times \sum f dy}{N \sum f dx^2 - (\sum f dx)^2} \times \frac{ix}{iy}$$

where, ix – width of the class intervals of x variable.

iy – width of the class intervals of y variable.

Illustration 10.5

Find out the two regression equations from the following distribution:

Sales (Rs '000)	10 – 20	Profit (in Rs '000) 20 – 30	30 – 40	40 – 50
60 – 70	2	5	4	6
70 – 80	4	6	5	4
80 – 90	6	7	6	8
90 – 100	8	8	5	9

Calculate (i) The co-efficient of correlation (ii) The profit corresponding to sales of Rs 2,50,000 and (iii) The sales corresponding to the profit of Rs 1,70,000.

Solutions

Let sales be denoted by x and profit by y .

Calculation of Regression Lines

x	y	M.P. $\frac{dy}{dx}$	10– 20	20–30	30–40	40–50	f	$\sum f dx$	$\sum f dx^2$	$\sum f dx dy$
60–	65	-1	4	5	0	-6	17	-17	17	3
70–	75	0	2	5	4	6	19	0	0	0
80–	85	1	-12	-7	0	5	4	27	27	-11
90–	95	2	-32	-16	0	18	30	60	120	-30
100			8	8	5	9				
		f	20	26	20	27	$N=93$	$\sum f dx =$ 70	$\sum f dx^2 =$ 164	$\sum f dx dy =$ $= -38$
		$\sum f dy$	-40	-26	0	27	$\sum f dy =$ -39			
		$\sum f dy^2$	80	26	0	27	$\sum f dy^2 =$ 133			
		$\sum f dx dy$	-40	-18	0	20	$\sum f dx dy =$ -38			

$$N = 93; \sum fdy = -39; \sum fdy^2 = 133; \sum fdx = 70; \sum fdx^2 = 164; \sum fdx dy = -38$$

Regression Co-efficient of x on y

$$\begin{aligned} b_{xy} &= \frac{\sum fdx dy - \sum fdx \times \sum fdy}{\sum fdy^2 - (\sum fdy)^2} \times \frac{ix}{iy} \\ &= \frac{-38 - \frac{70 \times (-39)}{93}}{133 - \frac{(-39)^2}{93}} \times \frac{10}{10} \\ &= \frac{-38 + \frac{2730}{93}}{133 - \frac{1521}{93}} \times 1 \\ &= \frac{-38 + 29.35}{33 - 16.35} \times 1 \\ &= \frac{-8.65}{116.65} \times 1 = -0.07 \end{aligned}$$

Regression Co-efficient of y on x

$$\begin{aligned} b_{yx} &= \frac{\sum fdx dy - \sum fdx \times \sum fdy}{\sum fdx^2 - (\sum fdx)^2} \times \frac{iy}{ix} \\ &= \frac{-38 - \frac{70 \times (-39)}{93}}{164 - \frac{(-70)^2}{93}} \times \frac{10}{10} \\ &= \frac{-38 + \frac{2730}{93}}{164 - \frac{4900}{93}} \times 1 \\ &= \frac{-38 + 29.35}{164 - 52.69} \times 1 \\ &= \frac{-8.65}{111.31} \times 1 \\ &= -0.08 \\ \bar{X} &= A \pm \frac{\sum fdx}{N} \times c \end{aligned}$$

$$= 75 + \frac{70}{93} \times 10 \\ = 75 + 7.53$$

$$\bar{X} = 82.53$$

$$\bar{Y} = A \pm \frac{\sum f dy}{N} \times C$$

$$= 35 + \frac{-39}{93} \times 10 \\ = 35 - 4.19$$

$$\bar{Y} = 30.81$$

Regression Equation of x on y :

$$x - \bar{x} = \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$x - 82.53 = -0.07(y - 30.81) \\ x = -0.07y + 2.16 + 82.53$$

$$x = 84.69 - 0.07$$

Regression Equation of y on x :

$$y - \bar{y} = \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 30.81 = -0.08(x - 82.53) \\ y = -0.08x + 6.60 + 30.81$$

$$y = 37.41 - 0.08x$$

(i) **Co-efficient of Correlation**

$$r = \sqrt{b_{xy} \times b_{yx}} \\ = \sqrt{(-0.07) \times (-0.08)} \\ = \sqrt{0.0056} = 0.07$$

(ii) The profit corresponding to sales Rs 2,50,000.

$$y_{250} = 37.41 - 0.08 \times 250 \\ = 37.41 - 20 \\ = 17.41 \\ \text{Rs } 17,410$$

(iii) The sales corresponding to profit Rs 1,70,000.

$$x_{170} = 84.69 - 0.07 \times 170 \\ = 84.69 - 11.9 \\ = 72.79$$

$$\text{i.e., } 72.79 \times 10,000 = \text{Rs } 7,27,900$$

Second Degree Parabola The equation for the second degree parabola for calculating non-linear trend is stated as,

$$Y_c = a + bx + cx^2$$

The trend value for any year may be compiled with the help of this equation by substituting the value of a , b , c and x . The values of a , b and c can be calculated by solving the following three normal equations.

$$\Sigma y = Na + b \Sigma x + c \Sigma x^2$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2 + c \Sigma x^3$$

$$\Sigma x^2 y = a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4$$

If the origin is taken between two middle years Σx and Σx^3 would be zero. Hence, the above equations can be reduced as follows.

$$\Sigma y = Na + c \Sigma x^2$$

$$\Sigma xy = b \Sigma x^2$$

$$\Sigma x^2 y = a \Sigma x^2 + c \Sigma x^4$$

Hence,

$$a = \frac{\Sigma y - c \Sigma x^2}{N}$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$c = \frac{N \Sigma x^2 y - \Sigma x^2 \Sigma y}{N \Sigma x^4 - (\Sigma x^2)^2}$$

Illustration 10.6

Fit a parabola $y = a + bx + cx^2$ from the following data. Estimate the price of the commodity for the year 1999.

Year	1992	1993	1994	1995	1996
Price (Rs.)	60	75	97	108	130

Solutions

Computation of Second Degree Parabola

Year	Price y	X	x^2	x^3	x^4	xy	x^2y	Trend yc
1992	60	-2	4	-8	16	-120	240	59.82
1993	75	-1	1	-1	1	-75	75	76.49
1994	97	0	0	0	0	0	0	93.58
1995	108	1	1	1	1	108	108	111.09
1996	130	2	4	8	16	260	520	129.02

$$y = 470 \sum x = 0 \sum x^2 = 10 \sum x^3 = 0 \quad \frac{3.5}{4.6} x^4 = 34 \quad \sum xy = 943 \quad \sum x^2y = 943 \quad \sum yc = 470$$

Trend values can be calculated as follows:

$$\begin{aligned}
 Y_{1992} &= 93.58 + 17.3(-2) + 0.21(-2)^2 \\
 &= 93.58 - 34.6 + 0.84 \\
 &= 59.82 \\
 Y_{1993} &= 93.58 + 17.3(-1) + 0.21(-1)^2 \\
 &= 93.58 - 17.3 + 0.21 \\
 &= 76.49 \\
 Y_{1994} &= 93.58 + 17.3(0) + 0.21(0)^2 \\
 &= 93.58 \\
 Y_{1995} &= 93.58 + 17.3(1) + 0.21(1)^2 \\
 &= 93.58 + 17.3 + 0.21 \\
 &= 111.09 \\
 Y_{1996} &= 93.58 + 17.3(2) + 0.21(2)^2 \\
 &= 93.58 + 34.6 + 0.84 \\
 &= 129.02
 \end{aligned}$$

To calculate the value of a , b and c , the following normal equations should be obtained:

$$\begin{aligned}
 \Sigma y &= Na + b \sum x + c \sum x^2 & (i) \\
 \Sigma xy &= a \sum x^2 + b \sum x^3 + c \sum x^4 & (ii) \\
 \Sigma x^2 y &= a \sum x^2 + b \sum x^3 + c \sum x^4 & (iii) \\
 470 &= 5a + 0 + 10c & (i) \\
 173 &= 0 + 10b + 0 & (ii) \\
 943 &= 10a + 0 + 34c & (iii) \\
 5a + 10c &= 470 & (i) \\
 10b &= 173 & (ii) \\
 10a + 34c &= 943 & (iii) \\
 10a + 20c &= 940 & (iv) \\
 14c &= 3 & \\
 (i) x^2 & & \\
 (iv) - (iii) & & \\
 & c = \frac{3}{14} = 0.21 & \\
 (i) & 5a + (10 \times 0.21) = 470 & \\
 & 5a = 470 - 2.1 & \\
 & a = \frac{467.90}{5} = 93.58 & \\
 (ii) & 10b = 173 & \\
 & b = \frac{173}{10} = 17.3 & \\
 \text{Hence, } & yc = a + bx + cx^2 & \\
 & y = 93.58 + 17.3x + 0.21x^2 &
 \end{aligned}$$

The price for 1999 can be calculated as follows:

$$\begin{aligned}
 \text{For 1999} & & x = 4 \\
 & & y = 93.58 + 17.3(4) + 0.21(4)^2 \\
 & & = 93.58 + 69.2 + 3.36 \\
 & & y = 166.14
 \end{aligned}$$

Illustration 10.7

From the following data, calculate

1. Correlation co-efficient
 2. Standard Deviation of y (σ_y)
- $$b_{xy} = 0.95y$$
- $$b_{yx} = 0.99x$$

Solutions

1. Co-efficient of correlation

$$\begin{aligned} r &= \sqrt{b_{xy} \times b_{yx}} \\ &= \sqrt{(0.95 \times 0.99)} \\ &= \sqrt{0.9405} = 0.97 \end{aligned}$$

2. Standard Deviation of y

$$\begin{aligned} r \frac{\sigma_x}{\sigma_y} &= 0.95 \\ 0.97 \times 3 / \sigma_y &= 0.85 \\ 0.95 \sigma_y &= 0.97 \times 3 \\ 0.95 \sigma_y &= 2.91 \\ \sigma_y &= \frac{2.91}{0.95} \\ \sigma_y &= 3.06 \end{aligned}$$

Illustration 10.8

From the following data, calculate the expected value of y when $x = 22$

	x	y
Average	8.6	15.5
Standard Deviation	4.6	3.5
$r = 0.98$		

Solutions

We have to calculate the expected value of y when x is 22. So, we have to find out regression equation of y on x .

Mean of x series $\bar{x} = 8.6$

Mean of x series $\bar{y} = 15.8$

σ of x series = 4.6

σ of y series = 3.5

Co-efficient of correlation $r = 0.98$

Regression of y on x

$$\begin{aligned}
 y - \bar{y} &= r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \\
 (y - 15.8) &= 0.98 \times \frac{3.5}{4.6} (x - 8.6) \\
 y - 15.8 &= 0.75(x - 8.6) \\
 y - 15.8 &= 0.75x - 6.45 \\
 y &= 0.75x - 6.45 + 15.8 \\
 y &= 0.75x + 9.35
 \end{aligned}$$

When x is 22

$$\begin{aligned}
 y &= 0.75(22) + 9.35 \\
 &= 16.5 + 9.35 = 25.85
 \end{aligned}$$

Hence, the expected value of y is 25.

10.4 MISCELLANEOUS ILLUSTRATIONS**Illustration 10.9**

Obtain the lines of two regressions for the following data and calculate the co-efficient of correlation.

X	10	20	30	40	50	60	70
Y	8	6	10	14	7	12	9

Obtain the value of y when $x = 6.5$.

(B.Com., MKU)

Solutions

Calculation of Regression Equations and correlation co-efficient

X	X(x - x) $x - 40 = x$	x^2	y	y(y - y) $y - 4 = y$	y^2	xy
10	-30	900	8	-1.4	1.96	42
20	-20	400	6	-3.4	11.56	68
30	-10	100	10	0.6	0.36	-6
40	0	0	14	4.6	21.16	0
50	10	100	7	-2.4	5.76	-24
60	20	400	12	2.6	6.76	52
70	30	900	9	-0.4	0.16	-12
$\sum x = 280$		$\sum x = 0$	$\sum x^2 = 2800$	$\sum y = 66$	$\sum y = 0.2$	$\sum y^2 = 47.72$
$\sum xy = 120$						

$$x = \frac{\sum x}{N} = \frac{280}{7} = 40$$

$$y = \frac{\sum y}{N} = \frac{66}{7} = 9.4$$

Regression Equation of x on y :

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2} = \frac{120}{47.72} = 2.5$$

$$x - 40 = 2.5(y - 9.4)$$

$$x - 40 = 2.5y - 23.5$$

$$x = 2.5y - 23.5 + 40$$

$$\boxed{x = 2.5y + 16.5}$$

Regression Equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum y^2} = \frac{120}{2800} = 0.043$$

$$y - 9.4 = 0.043(x - 40)$$

$$x - 40 = 0.043x - 1.72$$

$$y = 0.043x - 1.72 + 9.4$$

$$\boxed{y = 7.68 + 0.043x}$$

The value of y when

$$x = 6.5$$

$$y = 0.043x + 7.68$$

$$= 0.043 \times 6.5 + 7.68 = 0.2795 + 7.68$$

$$\boxed{y = 7.9595}$$

Correlation co-efficient i.e.,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$b_{xy} = 2.5$; $b_{yx} = 0.043$

$$r = \sqrt{2.5 \times 0.043}$$

$$r = \sqrt{0.1075}$$

$$\boxed{r = +0.328}$$

Illustration 10.10

The following table gives the various values of two variables:

X	21	22	29	28	43	49	33
Y	28	24	29	26	33	37	28

Determine the regression equations which may be associated with these values and calculate Karl Pearson's co-efficient of correlation. (M.Com., BDU)

Solutions

Calculation of Regression Equations and correlation co-efficient

X	$x = x - \bar{x}$	x^2	y	$y = y - \bar{y}$	y^2	xy	
21	$x = 32.1$	-11.1	123.21	28	-1.3	1.69	14.43

Contd.

X	x = x - \bar{x}	x^2	y	y = y - \bar{y}	y^2	xy
	x - 32.1			y - 29.3		
22	-10.1	102.01	24	-5.3	28.09	53.53
29	-3.1	9.61	29	-0.3	0.09	0.93
28	-4.1	16.81	26	-3.3	10.89	13.53
43	10.9	118.81	33	3.7	13.69	40.33
49	16.9	285.61	37	7.7	59.29	130.13
33	0.9	0.81	28	-1.3	1.69	-1.17
$\sum x = 225 \sum x = 0.3 \sum x^2 = 656.87 \sum y = 205 \sum y = -0.1 \sum y^2 = 115.43 \sum xy = 251.71$						

$$\bar{x} = \frac{\sum x}{N} = \frac{225}{7} = 32.1$$

$$\bar{y} = \frac{\sum y}{N} = \frac{205}{7} = 29.3$$

Regression Equation of x on y:

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{251.71}{115.43}$$

$$b_{xy} = 2.18$$

$$x - 32.1 = 2.18(y - 29.3)$$

$$x - 32.1 = 2.18y - 63.874$$

$$x = 2.18y - 63.874 + 32.1$$

$$x = 2.18y - 31.774$$

Regression Equation of y on x:

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{251.71}{656.87} = 0.38$$

$$y - 29.3 = 0.38(x - 32.1)$$

$$y - 29.3 = 0.38x - 12.198$$

$$y = 0.38x - 12.198 + 29.3$$

$$y = 0.38x + 17.102$$

Karl Pearson's co-efficient of correlation, i.e.,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$b_{xy} = 2.18 ; b_{yx} = 0.38$$

$$r = \sqrt{2.18 \times 0.38}$$

$$r = \sqrt{0.8284}$$

$$r = +0.91$$

Illustration 10.11

Find out from the following:

- The two regression equations
- Co-efficient of correlation
- Most likely value of x when $y = 12$
- Most likely value of y when $x = 22$

X	2	4	6	8	10
Y	8	6	10	14	7

(B.Com., BDU, MKU)

Solutions

Calculation of Regression Equations and correlation co-efficient

x	$dx = (x - A)$ $x - 6$	dx^2	y	$dy = (y - A)$ $y - 9$	dy^2	$dxdy$
2	-4	16	3	-6	36	24
4	-2	4	6	-3	9	6
6	0	0	9	0	0	0
8	2	4	12	3	9	6
10	4	16	15	6	36	24
30	0	40	45	0	90	60

$$\bar{x} = \frac{\Sigma x}{N} = \frac{30}{5} = 6$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{45}{5} = 9$$

Regression Equation of x on y:

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$b_{xy} = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

$$b_{xy} = \frac{60 - \frac{0 \times 0}{7}}{90 - \frac{0}{7}} = \frac{60}{90} = 0.67$$

$$b_{xy} = 0.67$$

$$x - 6 = 0.67(y - 9)$$

$$x - 6 = 0.67y - 6.03$$

$$x = 0.67y - 6.03 + 6$$

$$x = 0.03 - 0.67y$$

Regression Equation of y on x :

$$y - \bar{y} = byx(x - \bar{x})$$

$$byx = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

$$byx = \frac{60 - 0}{40 - 0} = \frac{60}{40} = 1.5$$

$$byx = 1.5$$

$$y - 9 = 1.5(x - 6)$$

$$y - 9 = 1.5x - 9$$

$$y = 1.5y - 9 + 9$$

$$\mathbf{y = 1.5x}$$

Co-efficient of Correlation i.e., $r = \sqrt{bxy \times byx}$

$$r = \sqrt{0.67 \times 1.5}$$

$$r = \sqrt{1.005} = +1.002 \text{ (or) } +1$$

The value of x when $y = 12$

$$x = 0.67y - 6.03 + 6$$

$$x = 0.67 \times 12 - 0.03 = 8.04 - 0.03$$

$$\mathbf{x = 8.01}$$

The value of y when $x = 22$

$$y = 1.5x$$

$$y = 1.5 \times 22$$

$$\mathbf{y = 33}$$

Illustration 10.12

In a correlation analysis between production and price of a commodity, the following constants were obtained.

	Production Index	Price Index
\bar{x}	110	98
S.D.	12	5
r	0.4	

Write down the two regression equations. Find the price index when the production index is 116.

Solutions

Let,

Production = x

$$\text{Price} = y$$

Regression equation of x on y :

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$\bar{x} = 110; r = 0.4; 5x = 12; \sigma_x = 12; \sigma_y = 5; \bar{y} = 98$$

$$x - 110 = 0.4 \times 12/5(y - 98)$$

$$x - 110 = 0.96(y - 98)$$

$$x - 110 = 0.96y - 94.08$$

$$x = 0.96y - 94.08 + 110$$

$$x = 15.92 + 0.96y$$

Regression equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 98 = 0.4 \times 5/12(x - 110)$$

$$y - 98 = 0.167(x - 110)$$

$$y - 98 = 0.167x - 18.37$$

$$y = 0.167x - 18.37 + 98$$

$$y = 79.63 + 0.167x$$

Price index when the production index = 116

$$y = 0.167x + 79.63$$

$$y = 0.167 \times 116 + 79.63$$

$$y = 19.37 + 79.63$$

$$y = 99$$

Illustration 10.12

The following results were worked out from the scores in costing and statistics in a certain examinations:

	Scores in Costing	Scores in Statistics
Mean	39.5	47.5
S.D.	10.8	17.8
Correlation co-efficient	- 0.4	

Find both the regression lines. Find the value of y for $x = 50$ and also find the value of x for $y = 30$. **(B.Com., CHU)**

Solutions

Regression equation of x on y :

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

400 Business Statistics

$$\bar{x} = 39.5; r = -0.4; \sigma_x = 10.8; \sigma_y = 17.8; \bar{y} = 47.5$$

$$x - 39.5 = -0.4 \cdot 10.8/17.8(y - 47.5)$$

$$x - 39.5 = 0.24(y - 47.5)$$

$$x - 39.5 = -0.24y + 11.4$$

$$x = -0.24y + 11.4 + 39.5$$

$$x = \mathbf{50.9 - 0.24y}$$

Regression equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 47.5 = -0.4 \times \frac{17.8}{10.8} (x - 39.5)$$

$$y - 47.5 = 0.66(x - 39.5)$$

$$y - 47.5 = -0.66x + 26.07$$

$$y = -0.66x + 26.07 + 47.5$$

$$y = \mathbf{73.57 - 0.66x}$$

The value of y for $x = 50$

$$y = -0.66 \times 50 + 73.57$$

$$y = -33 + 73.57$$

$$y = \mathbf{-40.57}$$

The value of x for $y = 30$

$$x = -0.24y + 50.9$$

$$= -0.24 \times 30 + 50.9$$

$$= -7.2 + 50.9$$

$$x = \mathbf{43.7}$$

Illustration 10.13

From 10 observations on Price (x) and Supply (y) of a commodity, the following summary of figures were obtained.

$$\sum x = 130 \quad \sum y = 220 \quad \sum x^2 = 2288 \quad \sum xy = 3467$$

Compute the line of regression of y on x and interpret the result. Estimate the supply when price is 16 units.

Solutions

Regression Equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$b_{yx} = \frac{\Sigma xy - N \bar{x} \bar{y}}{\Sigma x^2 - N (\Sigma x)^2}$$

$$N = 10; \Sigma x = 130; \Sigma y = 220; \Sigma x^2 = 2288; \Sigma xy = 3467$$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{130}{10} = 22$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{220}{10} = 22$$

$$= \frac{3467 - 10(13 \times 22)}{2288 - 10(13)^2}$$

$$= \frac{3467 - 2860}{2288 - 1690}$$

$$= \frac{607}{578} = 1.015$$

$$b_{yx} = 1.015$$

$$y - 22 = 1.015(x - 13)$$

$$y - 22 = 1.015x - 13.195$$

$$y = 1.015x - 13.195 + 22$$

$$y = \mathbf{8.805 + 1.015x}$$

Calculation of supply when price = 16 units

$$y = 1.015x + 8.805$$

$$y = 1.015(16) + 8.805$$

$$= 16.24 + 8.805 = 25.045$$

$$y = \mathbf{25.045}$$

$$2y - x - 50, \text{ } x \text{ on } y$$

$$-x = 50 + 2y$$

$$x = -50 - 2y$$

$$x = 2y - 50$$

$$x = y - 2$$

$$b_{xy} = \mathbf{1}$$

$$3y - 2x = 10, \text{ } y \text{ on } x$$

$$3y = 10 + 2x$$

$$y = 10/3 + 20/3x$$

$$y = 3.33 + 0.67x$$

$$y = 0.67x + 3.33$$

$$\begin{aligned}
 b_{yx} &= 0.67 \\
 r &= \sqrt{b_{xy} \times b_{yx}} \\
 &= \sqrt{1 \times 0.67} \\
 &= 0.818
 \end{aligned}$$

Illustration 10.14

Find the mean value of variables x and y and the correlation co-efficient from the following regression equations:

$$\begin{aligned}
 2y - x &= 50 \\
 3y - 2x &= 10
 \end{aligned}$$

Solutions

$$2y - x = 50 \quad (1)$$

$$3y - 2x = 10 \quad (2)$$

Multiply Eq. (1) by Eq. (2) and subtract equation from it,

$$4y - 2x = 100 \quad (3)$$

$$3y - 2x = 10 \quad (4)$$

$$\begin{array}{rcl}
 (-) & + & (-) \\
 & & y = 90 \\
 \hline
 & & y = 90
 \end{array}$$

Substitute $y = 90$ in Eq. 2

$$\begin{aligned}
 3y - 2x &= 10 \\
 3 \times 90 - 2x &= 10 \\
 270 - 2x &= 10 \\
 -2x &= 10 - 270 \\
 -2x &= -260 \\
 x &= -260 / -2 = 130 \\
 x &= 130
 \end{aligned}$$

Regression equation x on y :

$$\begin{aligned}
 3y - 2x &= 10 \\
 -2x &= 10 - 3y \\
 2x &= 3y - 10 \\
 x &= \frac{3y - 10}{2} \\
 x &= 1.5y - 5 \\
 b_{xy} &= 1.5
 \end{aligned}$$

Regression equation y on x :

$$\begin{aligned}
 2y - x &= 50 \\
 2y &= 50 + x
 \end{aligned}$$

$$\begin{aligned}
 y &= \frac{50+x}{2} \\
 y &= 25 + 0.5x \\
 byx &= 0.5 \\
 \text{Correlation co-efficient} &= \sqrt{bxy \times byx} \\
 &= \sqrt{1.5 \times 0.5} \\
 &= \sqrt{0.75} \\
 r &= 0.87
 \end{aligned}$$

Illustration 10.15

Calculate

- (i) The regression equation of x on y and y on x from the following data.
(ii) Estimate x when $y = 40$

x	20	24	26	34	36
y	10	12	14	18	26

(B.Com., M.Com., MKU)

Solutions

Calculation of Regression Equation

x	$dx = (x - A)$	dx^2	y	$dy = (y - A)$	dy^2	$dxdy$
	$x - 26$			$y - 14$		
20	-6	36	10	-4	16	24
24	-2	4	12	-2	4	4
26	0	0	14	0	0	0
34	8	64	18	4	16	32
36	10	100	26	12	144	120
140	10	204	80	10	180	180

Regression equation of x on y :

$$\begin{aligned}
 x - \bar{x} &= b_{xy}(y - \bar{y}) \\
 b_{xy} &= \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}} \\
 b_{xy} &= \frac{180 - \frac{10 \times 10}{5}}{180 - \frac{(10)^2}{5}} = \frac{180 - \frac{100}{5}}{180 - \frac{100}{5}}
 \end{aligned}$$

$$= 160/160 = 1$$

$$bxy = 1$$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{140}{5} = 28$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{80}{5} = 16$$

$$x - \bar{x} = bxy(y - \bar{y})$$

$$x - 28 = 1(y - 16)$$

$$x - 28 = y - 16$$

$$x = y - 16 + 28$$

$$x = y + 12$$

x when

$$y = 40$$

$$x = y + 12$$

$$x = 40 + 12$$

$$x = 52$$

Illustration 10.16

From the following information, calculate line of regression of *y* on *x*:

	x	y
Mean	40	60
Standard Deviation	10	15
Correlation co-efficient	0.7	

(B.Com., MKU, CHU)

Solutions

Regression equation of *y* on *x*:

$$y - \bar{y} = r \frac{\sigma y}{\sigma x} (x - \bar{x})$$

$$\bar{y} = 60; \bar{x} = 40; \sigma x = 10; \sigma y = 15; r = 0.7$$

$$y - 60 = 0.7 \times 15/10(x - 40)$$

$$y - 60 = 1.05(x - 40)$$

$$y - 60 = 1.05x - 42$$

$$y = 1.05x - 42 + 60$$

$$y = 18 + 1.05x$$

Illustration 10.17

Find the two regression equations from the following data:

Height of mothers (cm)	146	147	148	149	150
Height of sons (cm)	151	155	147	160	158

Find the height of sons when mothers height is 160 cms and find also height of mothers when sons height is 150 cm.

Solutions

Let height of mothers = x

Let height of sons = y

Calculation of Regression Equations

X	$dx = (x - A)$	dx^2	Y	$dy = (y - A)$	dy^2	$dxdy$
	$x - 148$			$y - 155$		
146	-2	4	151	-4	16	8
147	-1	1	155	0	0	0
148	0	0	147	-8	64	0
149	1	1	160	5	25	5
150	2	4	158	3	9	6
740	0	10	771	-4	114	19

$$\bar{x} = \frac{\sum x}{N} = \frac{740}{5} = 148$$

$$\bar{y} = \frac{\sum y}{N} = \frac{771}{5} = 154.2$$

Regression equation of x on y :

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$b_{xy} = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

$$b_{xy} = \frac{19 - \frac{(0 \times -4)}{5}}{114 - \frac{(-4)^2}{5}} = \frac{19}{114 - \left(\frac{16}{5}\right)}$$

$$= \frac{19}{114 - 32} = \frac{19}{110.8} = 0.17$$

$$b_{xy} = 0.17$$

$$\begin{aligned}
 x - 148 &= 0.17(y - 154.2) \\
 x - 148 &= 0.17y - 26.214 \\
 x &= 0.17y - 26.214 + 148.6 \\
 x &= \mathbf{0.17y + 121.79}
 \end{aligned}$$

Regression equation of y on x :

$$\begin{aligned}
 y - \bar{y} &= b_{yx}(x - \bar{x}) \\
 b_{yx} &= \frac{\sum dx \times \sum dy}{\sum dx^2 - \frac{(\sum dx)^2}{N}} \\
 b_{yx} &= \frac{19 - \frac{(0 \times -4)}{5}}{10 - \left(\frac{0}{5}\right)^2} \\
 &= \frac{19}{10} \\
 b_{yx} &= 1.9 \\
 y - 154.2 &= 1.9(x - 148) \\
 y - 154.2 &= 1.9x - 281.2 \\
 y &= 1.9x - 281.2 + 154.2 \\
 y &= 1.9x - 127 \\
 y &= \mathbf{-127 + 1.9x}
 \end{aligned}$$

Height of sons when mother's height is 150 cm

$$\begin{aligned}
 y &= 1.9 \times 12 - 281.2 \\
 y &= -127 + 1.9(160) \\
 y &= 127 + 304 \\
 y &= \mathbf{177}
 \end{aligned}$$

Height of mothers when son's height is 150 cm

$$\begin{aligned}
 x &= 0.17y + 121.79 \\
 x &= 0.17y + 150 + 121.79 \\
 x &= 25.5 + 121.79 \\
 x &= \mathbf{147.29}
 \end{aligned}$$

Illustration 10.18

Obtain the Regression equation from the following:

x	10	20	30	40	50
y	50	40	30	20	10

Solutions

X	$dx = (x - A)$ $x - 30$	dx^2	Y	$dy = (y - A)$ $y - 155$	dy^2	$dxdy$
10	-20	400	50	20	400	-400
20	-10	100	40	10	100	-100
30	0	0	30	0	0	0
40	10	100	20	-10	100	-100
50	20	400	10	-20	400	-400
$\sum x = 150$		$\sum dx = 0$	$\sum dx^2 = 1000$	$\sum y = 150$	$\sum dy = 0$	$\sum dy^2 = 1000$
					$\sum dxdy = -1000$	

Regression equation of x on y :

$$x - \bar{x} = bxy(y - \bar{y})$$

$$\bar{x} = \frac{\sum x}{N} = \frac{150}{5} = 30$$

$$\bar{y} = \frac{\sum y}{N} = \frac{150}{5} = 30$$

$$bxy = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

$$= \frac{-1000 - 0}{1000 - 0} = -1$$

$$bxy = -1$$

$$x - 30 = -1(y - 30)$$

$$x - 30 = -y + 30$$

$$x = -y + 30 + 30$$

$$x = 60 - y$$

Regression equation of y on x :

$$y - \bar{y} = byx(x - \bar{x})$$

$$byx = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

$$\begin{aligned}
 &= \frac{-1000 - 0}{1000 - 0} = -1 \\
 b_{yx} &= -1 \\
 y - 30 &= -1(x - 30) \\
 y - 30 &= -x + 30 \\
 y &= -y + 30 + 30 \\
 y &= 60 - x
 \end{aligned}$$

Illustration 10.19

Find the Regression equation from the following particulars:

x	2	4	6	8	10
y	1	2	3	4	5

Solutions

X	$dx = (x - 6)$	dx^2	Y	$dy = (y - 3)$	dy^2	$dxdy$
2	-4	16	1	-2	4	8
4	-2	4	2	-1	1	2
6	0	0	3	0	0	0
8	2	4	4	1	1	2
10	4	16	5	2	4	8
$\sum x = 30$		$\sum dx = 0$	$\sum dx^2 = 40$	$\sum y = 15$	$\sum dy = 0$	$\sum dy^2 = 10$
$\sum dxdy = 20$						

$$\bar{x} = \frac{\sum x}{N} = \frac{30}{5} = 6$$

$$\bar{y} = \frac{\sum y}{N} = \frac{15}{5} = 3$$

Regression equation of x on y :

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$\begin{aligned}
 b_{xy} &= \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}} \\
 &= \frac{-20 - 0}{10 - 0} = 2
 \end{aligned}$$

$$b_{xy} = 2$$

$$y - \bar{y} = b_{xy}(x - \bar{x})$$

$$\begin{aligned}
 x - 6 &= 2(y - 3) \\
 x - 6 &= 2y - 6 \\
 x &= 2y - 6 \\
 x &= 2y
 \end{aligned}$$

Regression equation of y on x :

$$\begin{aligned}
 y - \bar{y} &= b_{yx}(x - \bar{x}) \\
 b_{yx} &= \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}} \\
 &= \frac{20 - 0}{40 - 0} = 0.5 \\
 b_{yx} &= 0.5 \\
 (y - \bar{y}) &= b_{yx}(x - \bar{x}) \\
 y - 3 &= 0.5(x - 60) \\
 y - 3 &= 0.5x - 30 \\
 y &= 0.5x - 3 + 3 \\
 y &= 0.5x
 \end{aligned}$$

Illustration 10.20

A panel of two judges M and N graded seven dramatic performances by independently awarding marks as follows:

Performance	1	2	3	4	5	6	7
Marks by M	44	30	46	38	42	44	48
Marks by N	30	28	26	35	29	27	31

The eight performances, which judge ' N ' could not attend, was awarded 37 marks by judge ' M '. If judge ' N ' had also been present, how many marks would be expected to have been awarded by him to the eighth performance.

Solutions

Let marks by M taken us be x
marks by N taken us be y

Performance	X	$dx = (x - A)$ $x - 38$	dx^2	Y	$dy = (y - A)$ $y - 26$	dy^2	$dxdy$
1	44	6	36	30	4	16	24
2	30	-8	64	28	2	4	-16

Contd.

Performance	X	$dx = (x - A)$	dx^2	Y	$dy = (y - A)$	dy^2	$dxdy$
		$x - 38$			$y - 26$		
3	46	8	64	26	0	0	0
4	38	0	0	35	9	81	0
5	42	4	16	29	3	9	12
6	44	6	36	27	1	1	6
7	48	10	100	31	5	25	50
$\sum x = 292$		$\sum dx = 26$	$\sum dx^2 = 316$	$\sum y = 206$	$\sum dy = 24$	$\sum dy^2 = 136$	$\sum dxdy = 76$

Regression equation of x on y :

$$b_{xy} = \frac{\sum dx \times \sum dy}{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}$$

$$= \frac{76 - \frac{26 \times 24}{7}}{136 - \frac{(24)^2}{7}}$$

$$= \frac{76 - \frac{624}{7}}{136 - \frac{576}{7}}$$

$$= \frac{76 - 89.14}{136 - 82.28}$$

$$= \frac{-13.14}{53.72}$$

$$b_{xy} = -0.24$$

$$\bar{x} = \frac{\sum x}{N} = \frac{292}{7} = 41.7$$

$$\bar{y} = \frac{\sum y}{N} = \frac{206}{7} = 29.43$$

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 41.7 = -0.24(y - 29.43)$$

$$x - 41.7 = -0.24y + 7.06$$

$$x = -0.24y + 7.06 + 41.7$$

$$x = \mathbf{48.76 - 0.24 y}$$

Regression equation of y on x :

$$y - \bar{y} = byx(x - \bar{x})$$

$$byx = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

$$= \frac{76 - \frac{24 \times 26}{7}}{316 - \frac{(26)^2}{7}}$$

$$= \frac{76 - 89.14}{316 - 96.6}$$

$$= \frac{-13.14}{219.4}$$

$$byx = -0.06$$

$$y - \bar{y} = byx(x - \bar{x})$$

$$y - 29.43 = -0.06(x - 41.7)$$

$$y - 29.43 = -0.06x + 2.502$$

$$y = -0.06x + 2.502 + 29.43$$

$$\mathbf{y = 31.932 - 0.06 x}$$

when

$$x = 37$$

$$y = -0.06x + 31.932$$

$$y = -0.06(37) + 31.932$$

$$y = 2.22 + 31.932$$

$$\mathbf{y = 34.152}$$

$$y = 34$$

Illustration 10.21

x	1	2	3	4	5
y	2	5	7	6	8

Calculate regression equation and find out the value of y , when $x = 12$.

Solutions

x	x^2	y	y^2	xy
1	1	2	4	2
2	4	5	25	10
3	9	7	49	21
4	16	6	36	24
5	25	8	64	40
$\Sigma x = 15$		$\Sigma dx^2 = 55$	$\Sigma y = 28$	$\Sigma dy^2 = 178$
				$\Sigma dxdy = 97$

Regression equation of x on y

$$x = a + by$$

The simultaneous equations are:

$$\Sigma x = Na + b \Sigma y$$

$$\Sigma xy = a \Sigma y + b \Sigma y^2$$

By substituting the values

$$15 = 5a + b28 \quad (1)$$

$$97 = 28a + 178b \quad (2)$$

Equation (1) multiplying by 28 and Eq. (2) multiplying by 2

$$420 = 140a + 784b \quad (3)$$

$$485 = 140a + 890b \quad (4)$$

Subtract Eq. (3) from Eq. (4)

$$485 = 140a + 890b$$

$$420 = 140a + 784b$$

$$\begin{array}{r} (-) \quad (-) \quad (-) \\ \hline 65 = 106b \end{array}$$

$$65/106 = b$$

$$\mathbf{b = 0.61}$$

Substitute the b value in Eq. (7)

$$15 = 5a + 38b$$

$$15 = 5a + 28 \times 0.61$$

$$15 = 5a + 17.08$$

$$15 - 17.08 = 5a$$

$$-2.08 = 5a$$

$$-2.08/5 = a$$

$$a = 0.416$$

Equation

$$y = a + bx \quad x = a + by$$

$$y = -0.416 + 0.61 \quad x = -0.416 + 0.61x$$

$$y = 0.61x - 0.416 \quad x = 0.61 - 0.416$$

Regression equation of y on x :

$$y = a + bx$$

The simultaneous equations are

$$\Sigma y = Na + b \Sigma x$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2$$

By substituting the values

$$28 = 5a + 15b \quad (1)$$

$$97 = 15a + 55b \quad (2)$$

Equation (1) multiplying by 3 and subtract it from Eq. (2)

$$84 = 15a + 45b \quad (3)$$

$$(-) \quad (-) \quad (-)$$

$$\begin{array}{r} 97 = 15a + 55b \\ \hline 13 = 10b \end{array} \quad (4)$$

$$b = 13/10 = 1.3$$

$$\mathbf{b = 1.3}$$

Substitute the b value in Eq. (1)

$$28 = 5a + 15b$$

$$28 = 5a + 15 \times 1.3$$

$$28 = 5a + 19.5$$

$$8.5 = 5a$$

$$a = 1.7$$

Equation y on x

$$y = a + bx$$

$$\mathbf{y = 1.7 + 1.3x}$$

When

$$x = 12$$

$$y = 1.7 + 1.3x$$

$$y = 1.7 + (1.3 \times 12)$$

$$= 1.7 + 15.6$$

$$= 17.3$$

$$\mathbf{y = 17.3}$$

Illustration 10.22

You are supplied with the following data only:

$$\text{Variance of } x = 32$$

$$8 \times 10y - 90 = 0$$

$$50x - 25y = 325$$

Calculate:

- (a) The average value of x and y
- (b) Co-efficient of correlation between x and y

Solutions

$$(a) \quad \begin{array}{l} 8x - 10y - 90 = 0 \\ \quad 50x - 25y = 325 \end{array} \quad (1)$$

Multiplying Eq. (1) by 25 and multiplying Eq. (2) by 10

$$\begin{array}{r} 200x - 250y = 3250 \\ 500x - 250y = 2250 \end{array} \quad (3)$$

$$\begin{array}{r} (-) \quad (+) \quad (-) \\ \hline 300x = 1000 \end{array} \quad (4)$$

$$x = 1000/300$$

$$x = 3.33 \text{ or } x = 3.33$$

Substitute x value in Eq. (1)

$$\begin{aligned} 8x - 10y &= 90 \\ 8 \times 3.33 - 10y &= 90 \\ 26.04 - 10y &= 90 \\ -10y &= 90 - 26.04 \\ -10y &= 63.96 \\ y &= -63.96/10 \\ y &= -6.396 \end{aligned}$$

$$\begin{aligned} (b) \quad 8x - 10y &= 90 \\ -10y &= 90 - 8x \\ 10y &= 8x - 90 \\ y &= 8/10x - 90/10 \\ y &= 0.8x - 9 \\ b_{yx} &= 0.8 \\ 50x &= 325 + 25y \\ 50x - 25y &= 325 \\ x &= 325/50 + 25/50 y \\ x &= 6.5 + 0.5y \\ b_{xy} &= 0.5 \end{aligned}$$

Co-efficient of Correlation i.e., $r = \sqrt{b_{xy} \times b_{yx}}$

$$r = \sqrt{0.5 \times 0.8}$$

$$r = \sqrt{0.4}$$

$$r = +0.63$$

Illustration 10.23

	X Series	Y Series
Mean	18	100
Standard Deviation	12	30

- (a) Co-efficient of correlation between x and y series is + 0.6
 (b) Find out the most probable value of y if x is 60 and most probable value of x if y is 80.
 (c) If two regression co-efficients are 0.6 and 0.7, what would be the value of the co-efficient of correlation? (B.Com, MDU, BDU, BU)

Solutions

Regression equation of x on y :

$$x - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$\bar{x} = 18; \bar{y} = 100; \sigma_x = 12; \sigma_y = 30; r = 0.6$

$$x - 18 = 0.6 \times 12/30(y - 100)$$

$$x - 18 = 0.24(y - 100)$$

$$x - 18 = 0.24y - 24$$

$$\mathbf{x = -6 + 0.24y}$$

When

$$y = 80,$$

$$x = 0.24y - 6$$

$$x = 0.24(80) - 6$$

$$x = 19.2 - 6 = 13.2$$

$$x = 13.2$$

Regression equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 100 = 0.6(30/12)(x - 18)$$

$$y - 100 = 1.5(x - 18)$$

$$y - 100 = 1.5x - 27$$

$$y = 1.5x - 27 + 100$$

$$\mathbf{y = 73 + 1.5x}$$

when $x = 60$

$$y = 1.5x + 73$$

$$y = (1.5 \times 60) + 73$$

$$y = 90 + 73$$

$$\mathbf{y = 163}$$

Co-efficient of Correlation

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$r = \sqrt{0.6 \times 0.7}$$

$$r = \sqrt{0.42}$$

$$\mathbf{r = 0.65}$$

Illustration 10.24

The mean value of $x = 52$

The mean value of $y = 28$

The regression co-efficient of y on $x = -1.5$

The regression co-efficient of x on $y = +0.2$

Find

- The most probable value of y when $x = 60$
- r , the co-efficient of correlation.

Solutions

In order to find out the most probable value of y when $x = 60$, we must know the regression equation of y on x :

Regression equation of y on x :

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \text{ (or) } y - \bar{y} = b_{yx} (x - \bar{x})$$

$$\bar{y} = 28; \bar{x} = 52; b_{yx} = -1.5$$

$$y - 28 = -1.5(x - 52)$$

$$y - 28 = -1.5x + 78$$

$$y = -1.5x + 106$$

$$\mathbf{y = 106 - 1.5x}$$

$$y \text{ when } x = 60$$

$$y = 106 - 1.5(60)$$

$$y = 106 - 90$$

$$\mathbf{y = 16}$$

Co-efficient of Correlation, i.e., $r = \sqrt{b_{xy} \times b_{yx}}$

$$r = \sqrt{-1.5 \times 0.2}$$

$$r = \sqrt{-0.3}$$

$$\mathbf{r = -0.55}$$

Illustration 10.25

The lines of regression of a bivariable population are:

$$8x - 10y + 62 = 0$$

$$20x - 8y = 200$$

The variance of x is 16. Find:

- The mean values of x and y
- Correlation co-efficient between x and y
- Standard deviation of y .

Solutions

$$8x - 10y = -62 \quad (1)$$

$$20x - 8y = 200 \quad (2)$$

Multiplying Eq. (1) by 8 and multiplying Eq. (2) by 10

$$64x - 80y = -496 \quad (3)$$

$$200x - 80y = -2000 \quad (4)$$

Subtracting (3) from (4)

$$200x - 80y = -2000 \quad (4)$$

$$64x - 80y = -496 \quad (3)$$

$$\begin{array}{r} (-) \quad (+) \quad (+) \\ \hline 136x = 2496 \\ x = 2496/136 \\ x = 18.35 \end{array}$$

Substitute x value in Eq. (1)

$$\begin{aligned} 8x - 10y &= -62 \\ (8 \times 18.35) - 10y &= -62 \\ 146.8 - 10y &= -62 \\ -10y &= -62 - 146.08 \\ y &= -208.8/10 \\ y &= \mathbf{20.88} \end{aligned}$$

The mean values

$$\bar{x} = 18.35 \quad \bar{y} = 20.88$$

(ii) Rewriting the Eq. (2) x on y

$$\begin{aligned} 20x - 8y &= 200 \\ 20x &= 8y + 200 \\ x &= 8/20y + 200/20 \\ x &= 0.4y + 10 \\ b_{xy} &= 0.4 \end{aligned}$$

Rewriting the Eq. (1)

$$\begin{aligned} 8x - 10y &= -62 \\ -10y &= -62 - 8x \\ 10y &= 62 + 8x \\ y &= 62/10 + 8/10x \\ y &= 6.2 + 0.8x \\ y &= 0.8x + 6.2 \\ b_{yx} &= 0.8 \end{aligned}$$

Co-efficient of Correlation

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$r = \sqrt{0.4 \times 0.8}$$

$$r = \sqrt{0.32}$$

$$r = 0.56$$

The variance of x is 16.

$$\sigma_x^2 = 16$$

$$\sigma_x = 4$$

$$byx = r \frac{\sigma_y}{\sigma_x} = 0.8$$

$$= r \frac{\sigma_y}{\sigma_x} = 0.8$$

$$0.56 \frac{\sigma_y}{4} = 0.8$$

$$0.56 \sigma_y = 3.2$$

$$\sigma_y = 3.2 / 0.56 = 5.7$$

$$\sigma_y = 5.7$$

Illustration 10.26

From the following data, calculate

(i) Correlation co-efficient

(ii) Standard deviation of y (σ_y)

$$x = 0.79y$$

$$y = 0.75x$$

$$\sigma_x = 3$$

Solutions

(i) Co-efficient of Correlation, i.e., $r = \sqrt{b_{xy} \times b_{yx}}$
 $b_{xy} = 0.79$; $b_{yx} = 0.75$

$$r = \sqrt{0.79 \times 0.75}$$

$$r = \sqrt{0.5925}$$

$$r = 0.77$$

(ii) Standard deviation of y :

$$r \frac{\sigma_x}{\sigma_y} = 0.79$$

$$0.77 \frac{3}{\sigma_y} = 0.79$$

$$\frac{0.77 \times 3}{0.79} = \sigma_y$$

$$\frac{2.31}{0.79} = \sigma_y$$

$$\sigma_y = 2.92$$

Illustration 10.27

For 100 students of a class, the regression equation of marks in Statistics (x) on the marks in Accountancy (y) is $6y - 10x + 260 = 0$. The mean marks in Accountancy is 88 and variance of marks in Statistics is $9/16$ th of the variance of marks in Accountancy. Find the mean marks in Statistics and co-efficient of correlation between marks in the two subjects.

Solutions

We are given $6y - 10x + 260 = 0$ (or) $6y + 260 = 10x$
Let,

Marks in Statistics denoted by x .

Marks in Accountancy denoted by y

$$\begin{aligned} x \text{ when } & y = 88 \\ & 10x = 6y + 260 \\ & 10x = 6(88) + 260 \\ & 10x = 528 + 260 \\ & 10x = 788 \\ & x = \frac{788}{10} = 78.8 \\ & \boxed{x = 78.8} \end{aligned}$$

Hence, the mean marks in statistics are 78.8

Calculation of correlation co-efficient:

The regression equation of x on y from the given equation is

$$10x = 6y + 260$$

$$x = \frac{6y}{10} + \frac{260}{10}$$

$$x = 0.6y + 26$$

$$b_{xy} = 0.6; \sigma_x^2 = 9; \sigma_y^2 = 16; r = ?$$

$$b_{xy} = r \frac{\sigma_x^2}{\sigma_y^2}$$

$$0.6 = r = \frac{\sqrt{9}}{\sqrt{16}}$$

$$0.6 = r \frac{3}{4}$$

$$3r = 4 \times 0.6$$

$$3r = 2.4$$

$$r = \frac{2.4}{3} = 0.8$$

$$\boxed{r = 0.8}$$

Illustration 10.28

The following calculation have been made for prices of twelve stocks (x) on the Calcutta Stock Exchange on a certain day along with the volume of sales in thousands of shares (y). From these calculations, find the regression equation of prices on stocks, on the volume of sales of shares.

$$\sum x = 600; \quad \sum y = 400; \quad \sum xy = 12300; \quad \sum x^2 = 41658; \quad \sum y^2 = 17300.$$

(B.Com, MSU, MKU, CHU)

Solutions

Equation of prices on stocks, on the volume of sales of shares (a) x on y .

$$\begin{aligned}x - \bar{x} &= b_{xy} (y - \bar{y}) \\ \bar{x} &= \frac{\sum x}{N} = \frac{600}{12} = 50 \\ \bar{y} &= \frac{\sum y}{N} = \frac{400}{12} = 33.33 \\ b_{xy} &= \frac{\sum xy - N\bar{x}\bar{y}}{\sum y^2 - N(\bar{y})^2} = \frac{12300 - 12(50 \times 33.33)}{17300 - 12(33.33)^2} \\ &= \frac{12300 - 12(1666.5)}{17300 - 12(1110.89)} \\ &= \frac{12300 - 19998}{17300 - 13330.68} \\ &= \frac{-7698}{3969.32} = -1.94\end{aligned}$$

$$b_{xy} = -1.94$$

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

$$x - 50 = -1.94(y - 33.33)$$

$$x - 50 = -1.94y + 64.66$$

$$x = -1.94y + 64.66 + 50$$

$$x = -1.94y + 114.66$$

$$\boxed{x = 114.66 - 1.94y}$$

Illustration 10.29

The following data are given for marks in Company Law and Statistics in a certain year:

Mean marks in Statistics	40
Mean marks in Company Law	47.6
Standard deviation of Marks in Statistics	10.8
Standard deviation of Marks in Company Law	15.5

$$r = 0.42$$

Find regression equation.

Solutions

Let,

Marks in Statistics taken as x

Marks in Company Law taken as y

Regression equation of x on y :

$$(x - \bar{x}) = r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$\bar{x} = 40; \bar{y} = 47.6; \sigma_x = 10.8; \sigma_y = 15.5; r = 0.42$$

$$x - 40 = 0.42 \times 10.8/15.5 (y - 47.6)$$

$$x - 40 = 0.29(y - 47.6)$$

$$x - 40 = 0.29y - 13.80$$

$$x = 0.29y + 26.2$$

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$y - 47.6 = 0.42 \times 15.5/10.8 (x - 40)$$

$$y - 47.6 = 0.6 (x - 40)$$

$$y - 47.6 = 0.6x - 24$$

$$y = 0.6x + 23.6$$

Illustration 10.30

From the following data, calculate b_{xy} , b_{yx} and value of r

y	1	2	3	4
10	—	—	3	2
20	3	1	0	0
30	1	2	1	1
40	3	—	2	—

Solutions

Calculation of correlation co-efficient

X	y	$\sum f$	$\sum fdx$	$\sum fdy$	$\sum f dx dy$	$\sum f dy^2$	$\sum f dx^2$
10	-1	7	0	1	2	5	-5
20	0	7	-	3	<u>-3</u> 2	5	5
30	1	7	<u>0</u> 2	<u>0</u> 1	<u>1</u> 1	4	0
40	2	7	<u>-6</u> 3	-	<u>4</u> 2	5	5
		f	7	3	6	3	N = 19
		$\sum f$	$\sum fdx$	$\sum fdy$	$\sum f dx dy$	$\sum f dy^2$	$\sum f dx^2$
		7	-7	10	30	-7	25

$$b_{xy} = \frac{\sum f d x d y - \frac{(\sum f d x) \times (\sum f d y)}{N}}{\sum f d y^2 - \frac{(\sum f d y)^2}{N}}$$

$$r = \frac{-7 - \frac{(5)(10)}{19}}{30 - \frac{(10)^2}{19}} = \frac{-7 - \frac{50}{19}}{30 - \frac{100}{19}}$$

$$r = \frac{-7 - 2.63}{30 - 5.26} = \frac{-9.63}{24.74} = -0.39$$

$$b_{yx} = \frac{\sum f d x d y - \frac{(\sum f d x) \times (\sum f d y)}{N}}{\sum f d x^2 - \frac{(\sum f d x)^2}{N}}$$

$$= \frac{-7 - \frac{(5)(10)}{19}}{25 - \frac{(5)^2}{19}} = \frac{-7 - \frac{50}{19}}{25 - \frac{25}{19}}$$

$$= \frac{-7 - 2.63}{25 - 1.32} = \frac{-9.63}{23.68} = -0.41$$

$$= b_{yx} = -0.41$$

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$= \sqrt{-0.39 \times -0.41}$$

$$= \sqrt{0.16} = 0.4$$

$$r = 0.4$$

Illustration 10.31

The following are the marks obtained by 97 students in test x and

$y \backslash x$	30–40	40–50	50–60	60–70	Total
20–30	2	5	3	—	10
30–40	—	—	2	6	8
40–50	5	22	14	1	42
50–60	2	16	2	9	29
60–70	—	2	2	4	8
Total	9	45	23	20	97

Calculate (a) The correlation co-efficient
 (b) The regression equations.

Solutions

Calculation of regression equations and Correlation co-efficient

Y	X	30–40	40–50	50–60	60–70	f	fdy	fdy²	fdx dy
M	35	45	55	65					
	dx dy	-2	-1	0	1				
20–30	25	-2	8	10	3	-	10	-20	40
30–40	35	-1	-	-	2	0	6	-6	8
40–50	45	0	0	0	14	0	1	42	0
50–60	55	1	-4	-16	2	0	9	29	29
60–70	65	2	-	-4	2	0	4	8	16
f	9	45	23	20	N = 97	Σfdy = 17	Σfdy² = 109	5	
fdx	-18	-45	0	20	Σfdx = 43				
fdx²	36	45	0	20	Σfdx² = 101				
fdxdy	4	-10	0	11	5				

(a) Co-efficient of correlation:

$$b_{xy} = \frac{\Sigma f dx dy - \frac{(\Sigma f dx) \times (\Sigma f dy)}{N}}{\Sigma f dy^2 - \frac{(\Sigma f dy)^2}{N}}$$

$$r = \frac{5 - \frac{(-43)(17)}{97}}{109 - \frac{(17)^2}{97}}$$

$$r = \frac{5 - (-731)/97}{109 - \frac{289}{97}}$$

$$= \frac{5 - (-7.54)}{109 - 2.98}$$

$$= \frac{12.54}{106.02}$$

$$b_{xy} = 0.12$$

$$b_{yx} = \frac{\Sigma f dx dy - \frac{(\Sigma f dx) (\Sigma f dy)}{N}}{\Sigma f dx^2 - \frac{(\Sigma f dx)^2}{N}}$$

$$r = \frac{5 - \frac{(-43 \times 17)}{97}}{101 - \frac{(-43)^2}{97}}$$

$$= \frac{5 - \frac{(-7.54)}{97}}{101 - \frac{1849}{97}}$$

$$= \frac{5 + 7.54}{101 - 19.06}$$

$$= \frac{12.54}{81.94}$$

$$b_{xy} = 0.15$$

426 Business Statistics

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$r = \sqrt{0.12 \times 0.15}$$

$$r = \sqrt{0.018}$$

Mean \Rightarrow

$$r = 0.134 \text{ (or) } 0.13$$

$$\bar{x} = A + \frac{\sum f dx}{N} \times C$$

$$= 55 + \frac{-43}{97} \times 10$$

$$= 55 - 4.43$$

$$\bar{X} = 50.57$$

$$\bar{y} = A + \frac{\sum f dy}{N} \times C$$

$$= 45 + \frac{17}{97} \times 10$$

$$= 45 + 1.75$$

$$\bar{y} = 46.75$$

(b) Regression equation of x on y :

$$x - \bar{x} = b_{xy} (y - \bar{y})$$

$$x - 50.57 = 0.09(y - 46.75)$$

$$x - 50.57 = 0.09y - 4.2075$$

$$x = 0.09y - 4.2075 + 50.57$$

$$x = 0.09y + 46.3625$$

$$\mathbf{x = 0.09y + 46.36}$$

(c) Regression equation of y on x :

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y - 46.75 = 0.153(x - 50.57)$$

$$y - 46.75 = 0.153x - 7.74$$

$$\mathbf{y = 39.01 + 0.153x}$$

SUMMARY

Regression

It is a statistical technique through which the estimation of unknown variable from the known variable.

Independent Variable

The known variable which is used to estimate an unkown variable.

Dependent Variable

The unknown variable for which the value is to be predicted.

Simple linear Regression Analysis

The analysis which is used to find out an unknown variable from the known variable.

Importance of Regression

- To find out an unkown variable from the group of known variables.
- It can be applied in companies for quality control.
- To find out coefficient of correlation and co-efficient of detemination.

Regression Equation

Regression line which can be expressed in algebra terms.

Standard Error of Regression or Standard Error of Estimate

Statistical tchnique to know the reliabilty of the regression analysis.

FORMULAE

Regression Equations

X on *Y*

$$Y = a + by$$

$$x - \bar{x} = bxy(Y - \bar{Y})$$

Y on *X*

$$Y = a + bx$$

$$= Y - \bar{Y} = byx(x - \bar{X})$$

Normal Equations

X on *Y*

$$\sum x = Na + b\sum y$$

$$\sum xy = a\sum y + b\sum y^2$$

Y on *X*

$$\sum y = Na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Regression co-efficient

X on *Y*

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

X on Y

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

When Deviations are taken from mean

X on Y

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum xy}{\sum y^2}$$

$$X = (x - \bar{x}); Y = (y - \bar{y})$$

Y on X

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy}{\sum x^2}$$

When Deviations are taken from Assumed mean

X on Y

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

$$dx = (x - A); dy = (y - A)$$

Y on X

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

When the original values of X and Y series are taken

X on Y

$$r \frac{\sigma_x}{\sigma_y} = \frac{\sum xy - N \bar{x} \bar{y}}{\sum y^2 - N (\bar{y})^2}$$

Y on X

$$r \frac{\sigma_y}{\sigma_x} = \frac{\sum xy - \bar{x} \bar{y}}{\sum x^2 - N (\bar{x})^2}$$

$$\text{Co-efficient of correlation} = r = \sqrt{b_{xy} \times b_{yx}}$$

EXERCISES**(a) Choose the best option.**

1. Analysis which is used to find out an unknown variable from the known variable is
(a) Regression (b) Correlation (c) Sewness
2. The known variable which is used to estimate an unknown variable is called
(a) Independent variable
(b) Dependent variable
(c) Regression
3. The unknown variable which is used to be predicted is called
(a) Explained variable
(b) Explanatory variable
(c) Independent variables
4. When the regression lines are expressed in an algebra terms it is known as
(a) Regression equation
(b) Regression Analysis
(c) Correlation
5. For calculating non-linear trend, the equation for the second degree parabola is
(a) $Y_c = a + bx + cx^2$ (b) $Y_c = a + bx$ (c) $Y_c = a + bx^2$
6. Standard error of regression is also called as
(a) Standard error of estimate
(b) Standard error
(c) Standard error of determination.
7. Co-efficient of regression of Y on x is
(a) b_{yx} (b) b_{xy} (c) $bx+y$
8. Co-efficient of regression of x on y is
(a) b_{yx} (b) b_{xy} (c) $bx-y$
9. The two lines are very close to each other, if the degree of correlation is
(a) high (b) low (c) medium
10. The two regression lines will keep distance with each other, when the degree of correlation is
(a) Low (b) High (c) Medium

Answers

- | | | | | |
|--------|--------|--------|--------|---------|
| 1. (a) | 2. (b) | 3. (c) | 4. (a) | 5. (b) |
| 6. (a) | 7. (a) | 8. (b) | 9. (a) | 10. (a) |

(b) Theoretical Questions

1. What do you mean by Regression?
2. What is meant by Regression lines?
3. What is meant by Regression equations?
4. What are regression co-efficients? How are they computed?
5. Explain the concept of regression and ratio of variation.
6. What do you understand by regression? What are the properties of regression co-efficients?
7. Distinguish between regression and correlation.
8. What is regression line? What are its uses? Why are there generally two regression lines?
9. Briefly explain the concept of regression and write down the equations of the regression lines. When do the lines coincide?
10. How do you fit a regression equation to a set of bivariate data? Explain.
11. What are the assumptions of least square regression method?
12. What is regression analysis? Discuss its utility in predicting future events.
13. What would be the lines of regression if (i) $r = +1$, (ii) $r = -1$, (iii) $r = 0.7$. Give your interpretation in each case.
14. Explain the concept of regression and ratio of variation and state their utilities in the field of economics.
15. "If all the points in a scatter diagram lie exactly on the regression line, two variables are perfectly correlate." Comment.

(c) Practical Problems

16. Fit a straight line of y on x from the following data.

X	0	1	2	3	4	5	6
Y	2	1	3	2	4	3	5

Answer $y = 1.357 + 0.5x$

17. Calculate the two regression equations of x on y and y on x from the data given below taking deviations from actual mean of x and y .

Price (Rs)	10	12	13	12	16	15
Demand	40	38	43	45	37	43

Estimate the likely demand when the price is Rs. 20.

Answer $x = 17.92 - 0.12y$; $y = 44.25 + 0.25x$, when $x = 20$,
then $y = 49.25$

18. Fit a straight line to the following data by the method of least squares.

x	0	1	2	3	4
y	1	18	33	4.5	6.3

Answer $y = -4.29x + 11.96$ (B.Com., MKU, BDU, CHU)

19. Find both the regression equations from the following data:
 $\Sigma x = 60$; $\Sigma y = 40$; $\Sigma x^2 = 4160$; $\Sigma y^2 = 1720$ $\Sigma xy = 1150$ $N = 10$;
Answer $y = 0.24x + 2.56$, $x = 0.58y + 3.68$
20. Calculate the regression co-efficient for the following information:
 $\Sigma x = 50$; $\Sigma y = 30$; $\Sigma x^2 = 3000$; $\Sigma y^2 = 180$; $\Sigma xy = 1000$; $N = 10$;
Answer $b_{yx} = 0.309$ $b_{xy} = 0.497$

21. In a partially destroyed record, the following data are available:

Variance of $x = 25$

Regression equation of x on y : $5x - y = 22$;

Regression equation of y on x : $64x - 45y = 24$;

Find (a) mean values of x and y (b) Co-efficient of correlation between x and y (c) Standard deviation of y .

Answer (a) $x = 6$, $y = 8$, (b) $b_{yx} = 64/65$, $b_{xy} = 1/5$ $r = 8/15$.
(c) $\sigma_y = 13.33$ (B.Com., MKU, BDU)

22. From the following results, obtain the two regression equations and estimate the yield, when the rainfall is 29 cms, and the rainfall, when the yield is 600 kg.

	Yield in kg	Rainfall in cm
Mean	508.4	26.7
S.D.	36.8	4.6

(B.Com., MKU, BDU, CHU)

23. Co-efficient of correlation between yield and rainfall is +0.52
Answer $y = 4.16x + 397.328$ when $x = 29$, $y = 517.968$ kg. $x = 0.065$
 $y - 6$. when $y = 600$, $x = 32.654$ cms

24. The following data are given regarding expenditure on advertising and sales of particular firm.

	Advertisement Expenditure (x)	Sales (Rs. Lakhs) (y)
Mean	10	90
Standard deviation	3	12

(i) Calculate the regression equation of y on x .

(ii) Estimate the advertisement expenditure required to attain a sales target of Rs. 120 lakhs.

Answer (i) $y = 3.2x + 58$, (ii) $x = 0.2y - 8$ when $y = 120$ $x = 16$

(B.Com., MKU, BDU, CHU, MSU)

- 25.** Given that the variance of $x = 9$ and the regression equations are $8x - 10y + 66 = 0$, $40x - 18y = 214$. Find (a) mean values of x and y ; (b) co-efficient of correlation between x and y . (c) Standard deviation of y .

Answer (a) $x = 13$; $y = 17$; (b) $b_{yx} = 4/5$, $b_{xy} = 9/20$, $r = 6/10$, (c) $\sigma_y = 4$

- 26.** For a bivariate data, the mean value of x is 20 and the mean value of y is 45. The regression co-efficient of y on x is 4 and that of x on y is $(1/9)$. Find

- The co-efficient of correlation.
- The standard deviation of x if the standard deviation of y is 12.
- Also write down the equations of regression lines.

Answer (i) $r = 0.667$; (ii) $\sigma_x = 2$. (iii) $x = 1/9y + 15$, $y = 45 - 35$

- 27.** For 10 observations on price (x) and supply (y), the following data were obtained (in appropriate units);

$\Sigma x = 130$; $\Sigma y = 220$; $\Sigma x^2 = 2288$; $\Sigma y^2 = 5506$; $\Sigma xy = 3467$. Obtain the line of regression of y on x and estimate the supply when the price is 16 units.

Answer $y = 8.8 + 1.015x$ put $x = 16$ to get $y = 25.04$

(B.Com., MKU, BDU, CHU)

- 28.** Compute the two regression co-efficients from the data given below and find the value of r (correlation co-efficient) using the same.

x	7	4	8	6	5
y	6	5	9	8	2

Answer $b_{yx} = 6/5$, $b_{xy} = 2/5$, $r = 0.693$ **(B.Com, BDU, CHU)**

- 29.** Compute the two regression co-efficients using the values of actual means of x and y from the data given below and then work out the value of the correlation co-efficient.

x	7	4	8	6	5
y	6	5	9	8	2

Answer $b_{yx} = 1.2$, $b_{xy} = 0.4$, $r = 0.693$

- 30.** Find out the regression co-efficient of y on x from the following data.

x	1	2	3	4	5
y	160	180	140	180	200

Answer $b_{yx} = 8$

- 31.** A panel of judges, P and Q graded seven debators and independently awarded the following marks:

Debators	1	2	3	4	5	6	7
Marks by judge P	40	34	28	30	44	38	31
Marks by judge Q	32	39	26	30	38	34	28

An eighth debator was awarded 36 marks by judge P while judge Q was not present. If judge Q were also present, how many marks would you expect him to award to the eighth debator assuming the same degree of relationship exists between their judgement?

Answer $b_{yx} = 0.587, y = 0.587x + 11.885$, when $x = 36, y = 33$
(B.Com., MKU, BDU, BU)

32. The following table gives the age of cars of a certain make and actual maintenance costs. Obtain the regression equation for costs related to age. Also estimate the maintenance cost for a 10-year old car.

Age of Car (years)	2	4	6	8
Maintenance cost (Rs. 100)	10	20	25	30

Answer $b_{yx} = 3.25, y = 5 + 3.25x$. putting $x = 10, y = 37.5$ or Rs 37.50

33. The heights of a sample of 10 fathers and their eldest sons are given below (to the nearest cm):

Height of father (x)	170	167	162	163	167	166	169	171	164	165
Height of son (y)	168	167	166	166	168	165	168	170	165	168

- (i) Compute the correlation co-efficient y .
- (ii) Find the regression of y on x .
- (iii) Compute the co-efficient of determination and give your comments.

Answer $r = 0.76, y = 0.405x + 99.708, r^2 = 0.58$, the regression line accounts for 58% of total variation

(B.Com., BU, BDU, CHU)

34. From the following data, obtain the two regression equations:

Sales	91	97	108	121	67	124	51	73	111	57
Purchases	71	75	69	97	70	91	39	61	80	47

Answer $y = 0.6132x + 14.812; x = 1.36y - 5.2$

35. Find the two lines of regression from the following data:

Age of husband (in years)	25	22	28	26	35	20	22	40	20	18
Age of wife (in years)	18	15	20	17	22	14	16	21	15	14

Hence estimate:

- (i) the age of husband when the age of wife is 19 and
- (ii) the age of wife when the age of husband is 30.

Answer $b_{xy} = 2.23, b_{yx} = 0.385, x = 2.23y - 12.76, y = 0.385x + 7.34$,

- (i) when $y = 19, x = 29.61$ (or) 30 nearly, (ii) when $x = 30, y = 18.89$ (or) 19 nearly

Hence

- (i) Required estimated age of the husband $\rightarrow 30$ years

(ii) Required estimated age of the wife 19 years.

(B.Com., MKU, BDU, CHU)

36. The following data represents rainfall (x) and yield of paddy per hectare (y) in a particular area. Find the linear regression of x on y .

x	113	102	95	120	140	130	125
y	1.8	1.5	1.3	1.9	1.1	2.0	1.7

Answer $x = 19.965y + 85.637$ (B.Com., BDU, BU)

37. From the following data find the regression equation and estimate the likely value of y when $x = 100$

x	72	90	76	81	56	76	92	88	49
y	124	131	117	132	96	120	136	97	85

Answer $b_{xy} = 0.64; b_{yx} = 0.83; x = 0.64y + 2.63, y = 0.83 + 51.88$

38. Using the following data, obtain the two regression equations

x	14	19	24	21	26	22	15	20	19
y	31	36	48	37	50	45	33	41	3

Answer $x = 0.557y - 2.28, y = 1.608x + 7.84$

39. Find the two line of regression from the following data.

X	1	2	3	4	5
Y	3	6	9	12	15

(B.Com., MSU, CHU)

40. Using the following data, fit a two regression lines

X	3	6	9	12	15
Y	2	4	6	8	10

(i) estimate the value of x when $y = 9$

(ii) estimate the value of y when $x = 8$.

41. Find the r form the following date

Mark in Science	Mark in Maths				Total
	20–40	40–60	60–80	80–100	
20–40	—	—	—	—	0
40–60	—	4	2	—	7
60–80	—	3	10	6	21
80–100	—	—	3	9	12
Total	8	8	15	15	40

Answer $r_{12} = 0.62; X = 13.4 + 0.8y; y = 37.92 + 0.48x$

42. Find Karl Pearson's co-efficient of correlation

Mark in Science	Mark in Socal Science				Total
	20–40	40–60	60–80	80–100	
20–40	—	—	—	—	0
40–60	—	4	3	—	7

	Mark in science		Mark in Socal Science			Total
	20– 40	40–60	60–80	80–100		
60–80	—	3	11	4	18	
80–100	—	—	7	8	15	
Total	0	7	21	12	40	

Answer $r^{2/3} = 0.58$; $Y = 31.68 + 0.562 z$; $Z = 28.8 + 0.6y$

43. Find Karl Pearson's co-efficient of correlation

	Mark in Science		Mark in Social Science			Total
	20–40	40–60	60–80	80–100		
20–40	—	—	—	—	—	2
40–60	—	5	2	—	—	8
60–80	—	2	3	3	—	15
80–100	—	—	10	12	—	15
Total	0	7	18	12	—	40

Answer $r_{31} = 0.65$; $Z = 33.66 + 0.54x$; $x = 13.4 + 0.8z$

11

CHAPTER

INDEX NUMBERS

11.1 INTRODUCTION

An Index number is a statistical device for comparing the general level of magnitude of a group of related variables in two or more situations. It deals with the average of changes in a group of related variables over a period of time or between places. An index number is a number which is used to measure a certain phenomenon as compared to the level of same phenomenon at some standard period.

11.2 DEFINITIONS

Index numbers are devices for measuring differences in the magnitude of a group of related variables. —**Croxton and Cowden**

An index number is a percentage relative that compares economic measures in a given period with those same measures at a fixed period in the past. —**Clark and Schkade**

An index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographical location or other characteristics. —**Spiegel**

In its simplest form, an index number is nothing more than a relative number; or a relative which expresses the relationship between two figures, where one of the figures is used as a base. —**Morris Hamburg**

11.2.1 Characteristics

The following are the characteristics of Index Numbers

- 1. Index numbers are specialised averages:** Normally, averages can be used to compare only those variables which are expressed in the same units. But

index numbers help in comparing the changes in variables which are in different units.

2. **Index numbers are expressed in percentages:** Index numbers are expressed in terms of percentages so as to show the extent of change. However, percentage sign (%) is never used.
3. **Index numbers are for comparison:** The index numbers by their nature are comparative. They compare changes taking place over time or between places and like categories.
4. **Index numbers measure changes not capable of direct measurement:** Where it is difficult to measure the variation in the effects of a group of variables directly or where the variations are entirely incapable of direct quantitative study, relative variations are measured with the help of index numbers.

11.2.2 Uses

The following are the uses of Index Numbers:

1. **They measure the relative change.** Index numbers are particulars useful in measuring relative changes. They give better idea of changes in levels of prices, production, business activity, employment.
2. **They are of better comparison.** The index numbers reduce the changes to price level into more useful and understandable form. The numbers of the changes are further reduced to percentages which are easily comparable.
3. **They are good guides.** Index numbers are not restricted to the price phenomenon alone. Any phenomenon, which is spread over a period to time, is capable of being expressed numerically through index numbers. Thus, various kinds of index numbers serve different uses.
4. **They are economic barometers.** Various index numbers computed for different purposes, say employment, trade, transport, agriculture, industry etc., are of immense value in dealing with different economic problems. Thus, index numbers are the economic barometers.
5. **They are the pulse of the economy.** The stability of prices or their inflating or deflating conditions can well be observed with the help of indices. Index numbers of general price level will measure the purchasing power of money.
6. **They are a special type of averages.** All the basic ideas of average are employed for the construction of index numbers. In average, the data are homogeneous (in the same units): but in index number averages the variables have different units of measurement. Hence, it is a special type.
7. **They compare the standard of living.** Different places have different index numbers. Index numbers may measure the cost of living of different classes of people. Thus, comparison becomes easy in respect of general price index numbers.

8. **They help in formulating policies.** Formulation of good polities for the future depends upon past trends. Behaviour of the index numbers are studies carefully before making any policies. For instance, increase or decrease in wages required to study the cost of living index numbers.

11.2.3 Types of Index Number

There are three types of Index numbers. They are: Price Index, Quantity Index and Value Index.

Price Index It is an index number which compares the prices for a group of commodities at a certain time or at a place with prices of a base period. There are wholesale price index numbers and retail price index numbers. The wholesale price index reveals the changes in the general price level of a country. Retail price index reveals the changes in the retail prices of commodities, such as consumption goods, bank deposits, bonds.

Quantity Index Quantity index numbers study the changes in the volume of goods produced or consumed; for instance, industrial production, agricultural production, import, export.

Value Index These index numbers compare the total value of a certain period with the total value of the base period. Here, the total value is equal to the price of each, multiplied by the quantity; for instance, indices of profits, sales, inventories.

11.2.4 Problems in the Construction of Index Numbers

The following are some guidelines to be remembered when we construct index numbers.

Purpose or Object The statistician must clearly determine the purpose for which the index numbers are to be constructed, because there is no all purpose index numbers. Every index number has got its own uses and limitations. Cost of living index numbers of workers in an industrial area and those of the workers of an agricultural area are different in respect of requirements. Therefore, it is very essential to define clearly the purposes of the index numbers and that too beforehand.

Selection of Base The base period of an index number is very important as it is used for the construction of index numbers. Every index number must have a base. One cannot say whether the price level has increased or decreased, unless one compares the price level of the current year with the price level of the previous year. The year to be selected as base year must be normal year or a typical year and a recent year. The base may be of the following type.

Fixed Base The name reveals that the base year is a fixed one. The prices of a particular year, selected as a base period are treated as equal to 100. The changes in the prices of subsequent years are shown as the percentages of the base year.

Average Base Sometimes, it is difficult to select a year as base through normality. Under such a critical position, the average of several years is considered better, as abnormalities can be reduced to great extent.

Chain Base In fixed base method, the base year once selected, remains fixed and all index numbers are based on the same base year. In this method, there is no fixed base year. It changes from year to year. When a comparison is desired from year to year, a system of chain base is used. It is the previous year that is taken as the base for the current year; and the change is calculated as a percentage of that year.

Selection of Commodities If we study the price changes of one commodity, we have to include only one item. For instance, if we study the changes in production of cloth, then we may include the production of mill cloth, powerloom cloth, handloom cloth, silk, khadi, etc., and there is no problem. Another example, say index of retail price; we cannot include all commodities sold in retail. We include only the important commodities which are representative of tastes, habits and customs of the people. For the purpose of finding the cost of living index number of low income group, we have to select only those items or commodities, which are mostly consumed by that group.

Source of Data The price relating to the thing to be measured must be collected. If we want to study the changes in industrial production, we must collect the prices relating to the production of various good of factories. The price may be collected from reliable sources. The prices of commodities are the raw materials for the construction of index numbers. The prices may be collected from the public sources or from standard commercial magazines.

Selection of Averages One can use any average. But in practice, the arithmetic average is used, because it is easy for computation; geometric mean and harmonic mean are difficult to calculate. But geometric mean is preferred because of the following characteristics (a) Geometric mean is the best measure and (b) It gives less weight to bigger items and more weight to smaller items.

Weighting All commodities are not equally important. The main purpose of an index number of prices is to ascertain the changes in the price level. In case of simple average, all commodities will have equal importance. But in actual practice, different groups of people will have different preferences on different commodities.

The relative importance is the basis of weightage. Generally, the quantities of the commodities produced or the values of quantities of the goods sold,

demanded or purchased are taken as suitable weights. These weights relate to the base year or the current years, depending on their availability or suitability.

There are two methods of weighting:

- (a) Implicit weighting
- (b) Explicit weighting.

Implicit Weighting Weights are not expressly laid down. The weights are implied by the nature of commodities selected. It means the weights relate to the selection of commodities themselves. Emphasis is to be given to the commodities according to the number of times the given commodities are included in the selection. For example, a particular commodity is included diet, 4 times, the weight given to the commodity is four.

Explicit Weighting Weights are expressly laid down on the basis of importance of the items. In case only one variety is included in the construction of index numbers, than the price is multiplied by the quantity.

NOTATIONS

Base year : The year selected for which comparison, that is, the year with reference to which comparisons are made. It is denoted by '0'.

Current year : The year for which comparisons are sought or required.

P_0 = Price of a commodity in the base year.

P_1 = Price of a commodity in the current year.

q_0 = Quantity of a commodity consumed or purchased during the base year.

q_1 = Quantity of a commodity consumed or purchase in the current year.

w = Weight assigned to a commodity according to its relative importance in the group.

P_{01} = Price Index Number for the base year with reference to the base year.

P_{10} = Price Index Number for the current year with reference to the current year.

Q_{01} = Quantity Index Number for the current year with reference to the base year.

Q_{10} = Quantity Index Number for the base year with reference to the current year.

11.3 METHODS OF INDEX NUMBERS

The following are various methods of index numbers.

11.3.1 Unweighted Index Numbers

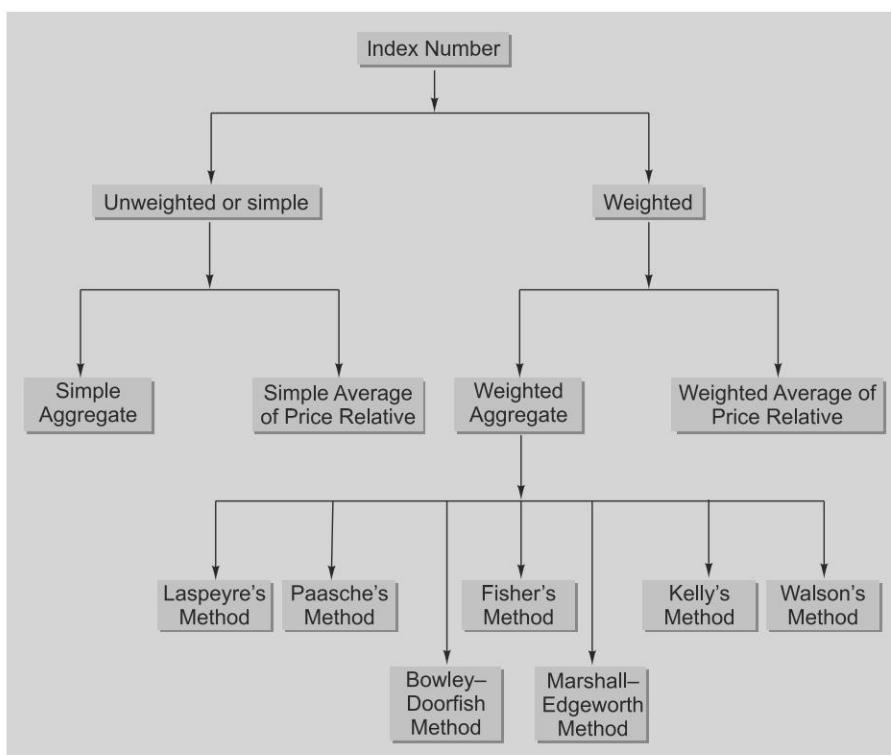
1. Simple Aggregate Method
2. Simple Average of Price Relative Method

11.3.2 Weighted Index Numbers

While calculating Index Numbers, weight is to be considered. The purpose is to give more representation. They are of two types—(i) Weighted aggregate method which has further been classified into:

- 1. Laspeyre's Method
- 2. Paasche's Method
- 3. Bowley–Doorfish Method
- 4. Fisher's Ideal Method
- 5. Marshall–Edgeworth Method
- 6. Kelly's Method
- 7. Walson's Method

Weighted average of price relative method



11.3.3 UNWEIGHTED INDEX NUMBERS

Simple aggregate method The price of the different commodities of the current year are added and the total is divided by the sum of the prices of the base year commodities and it is multiplied by 100;

The formula is $P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$

where

P_{01} = Price index number for the current year with reference to the base year

ΣP_1 = Aggregate of Prices for the current year

ΣP_0 = Aggregate of Prices for the base year

Steps

1. Prices of all commodities of current years are added, i.e., ΣP_1
2. Prices of all commodities of base year are added, i.e., ΣP_0
3. ΣP_1 is divided by ΣP_0 and it is P_{01}
4. $\frac{\Sigma P_1}{\Sigma P_0}$ multiplied with 100.

Illustration 11.1

Commodities	Price in 2006	Price in 2007
A	80	90
B	100	115
C	90	95
D	70	80

Construct an index number for 2007 taking 2006 as a base.

Solutions

Calculation of Index numbers

Commodities	Price in 2006 (P_0)	Price in 2007 (P_1)
A	80	90
B	100	115
C	90	95
D	70	80
	$\Sigma P_0 = 340$	$\Sigma P_1 = 380$

$$\begin{aligned}
 P_{01} &= \frac{\sum P_1}{\sum P_0} \times 100 \\
 &= \frac{380}{340} \times 100 \\
 &= 126.666 = 126.67
 \end{aligned}$$

Limitation of Simple Aggregate Method The relative importance of various commodities is not taken into account in the index as it is unweighted.

Simple Average of Price Relative Method In this method, the price relative of each item is calculated separately and then averaged. A price relative is the price of the current year expressed as a percentage of the price of the base year.

$$\text{i.e., } P_{01} = \frac{\sum \left[\frac{P_1}{P_0} \times 100 \right]}{N} = \frac{\sum P}{N}$$

N = Number of observations

When the Geometric mean is used the formula is

$$P_{01} = \text{antilog of } \frac{\sum \log \left[\frac{P_1}{P_0} \times 100 \right]}{N}$$

$$\text{or } = \text{antilog of } \frac{\sum \log P}{N}$$

where

$$P = \frac{P_1}{P_0} \times 100$$

Illustration 11.2

Construct a price index from the following by (a) simple aggregate method (b) average of price relative method by using both arithmetic mean method and geometric mean method.

Commodities	101	102	103	104	105
Price in 2006 (Rs)	30	36	42	48	60
Price in 2007 (Rs)	35	39	43	59	75

Solutions

Calculation of Price Index

Commodity	Price in 2006 (P_0)	Price in 2007 (P_1)	Price Relative $P = P_1/P_0 \times 100$	$\log P$
101	30	35	116.67	2.0671
102	36	39	108.33	2.0346
103	42	43	102.38	2.0103
104	48	59	122.92	2.0895
105	60	75	125.00	2.0969
$\Sigma P_0 = 216$		$\Sigma P_1 = 251$	$\Sigma P = 575.30$	$\Sigma P = 10.2984$

$$\begin{aligned}
 \text{(a) Simple Aggregative} &= \frac{\sum P_1}{\sum P_0} \times 100 \\
 &= \frac{251}{216} \times 100 = 116.2
 \end{aligned}$$

$$\begin{aligned}
 \text{(b) Arithmetic mean of price relatives} &= \frac{P_{01}}{N} = \frac{\Sigma P}{N} \\
 N &= 5
 \end{aligned}$$

$$= \frac{575.30}{5} = 115.06$$

(c) Geometric mean of price

$$\begin{aligned}
 \text{relative index} &= \text{Antilog} \left(\frac{\sum \log P}{N} \right) \\
 &= \text{Antilog} \frac{10.2984}{5} = \text{Antilog} (2.0598) \\
 &= \text{Antilog} (2.0597) = 114.8
 \end{aligned}$$

Merits

1. Extreme items do not affect the index numbers
2. It gives equal importance to all items
3. The index number will not be affected by the absolute level of individual prices.

Demerits

1. The arithmetic mean should not be used in all situation like changes in ratios.
2. The use of geometric mean involves difficulties of computation
3. The relative importance of each item has not been taken into account in this method.

11.3.4 Weighted Index Numbers

In this method, the quantity consumed is also taken into account, in addition to the prices of commodities. This is the most accurate method for calculating index number.

Weighted index numbers are of two types.

They are:

1. Weighted aggregate method and
2. Weighted average of price relatives.

Weighted Aggregate Method This method is based on the weight of the prices of the selected commodities. According to this method, prices themselves are weighted by quantities, that is, $p \times q$. Thus, physical quantities are used as weights. It can be calculated through various methods.

Following are the most common method, for calculating index under this method:

- | | |
|------------------------------|--------------------------|
| 1. Laspeyre's Method | 2. Paasche's Method |
| 3. Bowley–Doorfish Method | 4. Fisher's Ideal Method |
| 5. Marshall–Edgeworth Method | 6. Kelly's Method |
| 7. Walsch's Method | |

Laspeyre's Method In this method, the weight for the price is calculated by multiplying both the current year prices and base year prices with base year quantities. Price Index under this method is

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

Paasche's Method In this method, current year quantities are taken as weight for the base year prices as well as current year prices. Price index under this method is

$$P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$$

3. Bowley–Doorfish Method Under this method, price index can be calculated by taking the average of the price indices calculated by Laspeyre's and Paasche's method.

Hence, price index is $P_{01} = \frac{L + P}{2}$

where L indicates Laspeyre's method and P indicates Paasche's method.

i.e.,
$$P_{01} = \frac{\sum P_1 q_0 + \sum P_1 q_1}{\sum P_0 q_0 + \sum P_0 q_1} \times 100$$

Fisher's Ideal Method Under this method, price index number can be calculated by the Geometric mean of the index number under Laspeyre's method and Paasche's method.

Hence, the index number is

$$\begin{aligned} P_{01} &= \sqrt{L \times P} \\ \text{i.e.,} \quad &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \end{aligned}$$

Marshall–Edgeworth Method Under the method, the index number is calculated as

$$\begin{aligned} P_{01} &= \frac{\sum P_1(q_0 + q_1)}{\sum P_0(q_0 + q_1)} \times 100 \\ \text{i.e.,} \quad &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \end{aligned}$$

In this method, the arithmetic mean of base year and current year quantities is taken as weights.

Kelly's Method Index number under this method is

$$P_{01} = \frac{\sum P_1 q}{\sum P_0 q} \times 100 \text{ where } q = \frac{q_0 + q_1}{2}$$

This method uses quantities of some period as weights.

Illustration 11.4

Following are the data related with the prices and quantity consumed for 2006 and 2007.

Commodity	2006		2007	
	Price (Rs)	Quantity (kg)	Price (Rs)	Quantity (kg)
Rice	25	10	27	15
Wheat	20	5	22	7
Sugar	22	4	24	6
Tea	15	2	17	5

Construct price index numbers by

- (a) Laspeyre's Method
- (b) Paasche's Method
- (c) Bowley–Doorfish Method
- (d) Fisher's Method

Solutions

Construction of Index Numbers

Commodity	2006		2007		$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
	Price P_0	Quantity q_0	Price P_1	Quantity q_1				
Rice	25	10	27	15	270	250	405	375
Wheat	20	5	22	7	110	100	154	140

Commodity	2006		2007		$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$	
	Price P_0	Quantity q_0	Price P_1	Quantity q_1					
Sugar	22	4	24	6	96	88	144	132	
Tea	15	2	17	5	34	30	85	75	
						$\sum P_1 q_0$ = 510	$\sum P_0 q_0$ = 468	$\sum P_1 q_1$ = 788	$\sum P_0 q_1$ = 722

(a) Index number by Laspeyre's Method

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 = \frac{510}{468} \times 100 = 108.97$$

(b) Index number by Paasche's Method

$$\begin{aligned} P_{01} &= \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100 \\ &= \frac{788}{722} \times 100 = 109.14 \end{aligned}$$

(c) Index number by Bowley–Doorfish Method

$$\begin{aligned} P_{01} &= \frac{L + P}{2} = \frac{108.97 + 109.14}{2} \\ &= 109.055 \end{aligned}$$

(d) Index number by Fisher's Ideal Method

$$\begin{aligned} P_{01} &= \sqrt{L \times P} = \sqrt{108.97 \times 109.14} \\ &= 109.055 \end{aligned}$$

Illustration 11.5

Calculate Index numbers by (a) Marshal–Edgeworth Method (b) Kelly's Method and (c) Walsch's Method from the following:

Commodity	2005		2006	
	Price	Quantity	Price	Quantity
A	20	10	15	20
B	25	20	21	16
C	22	15	18	13
D	35	5	23	15
E	40	3	31	9
F	30	10	18	12

Solutions

Calculation of Index Number

Commodity	2005 Price	2005 Quantity	2006 Price	2006 Quantity	$q = \frac{q_0 + q_1}{2}$	$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$	$P_1 q$	$P_0 q$
	P_0	q_0	P_1	q_1							
A	20	10	15	20	15	150	200	300	400	225	300
B	25	20	21	16	18	420	500	336	400	378	450
C	22	15	18	13	14	270	330	234	286	252	308
D	35	5	23	15	10	115	175	345	525	230	350
E	40	3	31	9	6	93	120	279	360	186	240
F	30	10	18	12	11	180	300	216	360	198	330
						$\Sigma P_1 q_0$	$\Sigma P_0 q_0$	$\Sigma P_1 q_1$	$\Sigma P_0 q_1$	$\Sigma P_1 q$	$\Sigma P_0 q$
						= 1228	= 1625	= 1710	= 2331	= 1469	= 1978

(a) Index number under Marshal–Edgeworth Method

$$\begin{aligned}
 P_{01} &= \frac{\Sigma P_1 q_0 + \Sigma P_1 q_1}{\Sigma P_0 q_0 + \Sigma P_0 q_1} \times 100 \\
 &= \frac{1228 + 1710}{1625 + 2331} \times 100 \\
 &= \frac{2938}{3956} \times 100 = 74.266 \\
 &= 74.27
 \end{aligned}$$

(b) Index number under Kelly's Method

$$\begin{aligned}
 P_{01} &= \frac{\Sigma P_1 q}{\Sigma P_0 q} \times 100 \\
 &= \frac{1469}{1978} \times 100 \\
 &= 74.266 \\
 &= 74.27
 \end{aligned}$$

Illustration 11.6

Calculate price index numbers through (a) Bowley–Doorfish method and (b) Fisher's Ideal Method from the following data.

Commodity	2002		2003	
	Price	Quantity	Price	Quantity
A	36	12	38	18
B	80	15	85	13
C	70	10	72	15
D	42	14	40	10
E	46	10	48	5

Solutions

Calculation of Price Index Number

Commo- dity	2002		2003		P_1q_0	P_0q_0	P_1q_1	P_0q_1
	Price	Quantity	Price	Quantity				
	P_0	q_0	P_1	q_1				
A	36	12	38	18	456	432	684	648
B	80	15	85	13	1275	1200	1105	1040
C	70	10	72	15	720	700	1080	1050
D	42	14	40	10	560	588	400	420
E	46	10	48	5	480	460	240	230
					ΣP_1q_0 =3491	ΣP_0q_0 =3380	ΣP_1q_1 =3509	ΣP_0q_1 =3388

(a) Index number by Bowley–Doorfish Method

$$P_{01} = \frac{\sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}}}{2} \times 10 = \frac{\sqrt{\frac{3491}{3380} + \frac{3509}{3388}}}{2} \times 100 \\ = \frac{\sqrt{1.0328 + 1.0357}}{2} \times 100 = 103.425$$

(b) Index number by Fisher's Ideal method

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 = \sqrt{\frac{3491}{3380} \times \frac{3509}{3388}} \times 100 \\ = \sqrt{1.0328 \times 1.0357} \times 100 = 103.425$$

Weighted Average of Price Relative Method For calculating price Index under this method, price relatives of each commodity should be taken into account. It can be calculated through arithmetic mean and geometric mean methods.

450 *Business Statistics*

Arithmetic Mean Method Under this method, index number can be calculated as

$$P_{01} = \frac{\Sigma PV}{\Sigma V}$$

where P means price relatives, i.e., equal to $\frac{P_1}{P_0} \times 100$

and V means value weights i.e., equal to P_0q_0 .

Steps

1. Calculate the value weights of each commodity by multiplying the price for the base year with quantity for the base year (i.e., $V = P_0q_0$) and ascertain ΣV .
2. Find out the price relatives for each commodity i.e., $P = \frac{P_1}{P_0} \times 100$
3. Multiply the price relative with the value weights for each commodity (i.e., PV) and then compute the value ΣPV .
4. Divide ΣPV with ΣV to find out the price index.

Geometric Mean Method Under this method, the index number can be calculated as $P_{01} = \text{antilog } \frac{\Sigma V \log P}{\Sigma V}$

Steps

- (1) Calculate value weight for each commodity $V = P_0q_0$ and find out ΣV .
- (2) Find out the price relatives for each commodity $P = \frac{P_1}{P_0} \times 100$
- (3) Find out logarithm for the price relatives ($\log P$) for each commodities.
- (4) Multiply the log price relatives with the value weights (i.e., $V \log P$) from this calculate $\Sigma \log P$.
- (5) Then divide $\Sigma V \log P$ with ΣV and find out the antilog value for this, to get the price index number.

Illustration 11.7

Calculate price index from the following data by weighted average of price relatives through (a) Arithmetic mean method and (b) Geometric mean method.

Commodity	2003		2004	
	Price	Quantity	Price	Quantity
A	15	20	20	18

Commodity	2003		2004	
	Price	Quantity	Price	Quantity
B	13	15	22	15
C	9	10	11	8
D	14	12	15	10
E	19	8	28	10
F	7	7	18	9

Solutions

Calculation of price index number

Commo-	Price in 2003 (P_0)	Qty 2003 (q_0)	Price in 2004 (P_1)	$V = P_0 q_0$	$P = P_1 / P_0 \times 100$	PV	$\log P$	$V \log P$
A	15	20	20	300	133.33	39999	2.1249	637.47
B	13	15	22	195	169.23	32999.85	2.2284	434.54
C	9	10	11	90	122.22	10999.8	2.0869	187.82
D	14	12	15	168	107.14	17999.52	2.0298	341.01
E	19	8	28	152	147.37	22400.24	2.1687	329.64
F	7	7	18	49	257.14	12599.86	2.4101	118.09
$\Sigma V = 954$				$\Sigma PV = 136998.27$			$\Sigma V \log P = 2048.57$	

(a) Arithmetic Mean Method

$$P_{01} = \frac{\Sigma PV}{\Sigma V} = \frac{136998.27}{954} = 143.6$$

(b) Geometric Mean Method

$$\begin{aligned} P_{01} &= \text{antilog} \frac{\Sigma V \log P}{\Sigma V} = \text{Antilog} \frac{2048.57}{954} \\ &= \text{Antilog} 2.147 \\ &= 140.4 \end{aligned}$$

Quantity Index Numbers The quantity index numbers permit comparison of the physical quantity of goods produced, consumed or distributed. To construct quantity index numbers, the prices of the commodity either in the base year or in the current year should be taken into consideration.

The quantity index number can be calculated through the following methods.

$$(i) \text{ Laspeyre's Method} \quad Q_{01} = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

$$(ii) \text{ Paasche's Method} \quad Q_{01} = \frac{\sum q_1 P_1}{\sum q_0 P_1}$$

$$(iii) \text{ Fisher's Method} \quad Q_{01} = \sqrt{\frac{\sum q_1 P_0}{\sum q_0 P_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_1}} \times 100$$

or
 $= \sqrt{L \times P}$

where L indicates Laspeyre's method and P indicates Paasche's method.

Illustration 11.8

Calculate quantity index from the following data through (i) Laspeyre's Method (ii) Paasche's Method and (iii) Fisher's Method.

Commodity	2005		2006	
	Price	Quantity	Price	Quantity
A	35	10	35	12
B	31	15	42	16
C	23	12	32	10
D	20	14	25	15
E	40	8	35	10
F	32	6	45	6

Solutions

Calculation of Quantity Index Number

Commodity	2005		2006		$q_1 P_0$	$q_0 P_0$	$q_1 P_1$	$q_0 P_1$
	Price (P_0)	Qty (q_0)	Price (P_1)	Qty (q_1)				
A	35	10	35	12	420	350	420	350
B	31	15	42	16	496	465	672	630
C	23	12	32	10	230	276	320	384
D	20	14	25	15	300	280	375	350
E	40	8	35	10	400	320	350	280
F	32	6	45	6	192	192	270	270
					$\Sigma q_1 P_0$ = 2038	$\Sigma q_0 P_0$ = 1883	$\Sigma q_1 P_1$ = 2407	$\Sigma q_0 P_1$ = 2264

(a) Laspeyres's Method:

$$Q_{0_1} = \frac{\sum q_1 P_0}{\sum q_0 P_0} \times 100 = \frac{2038}{1883} \times 100 = 108.23$$

(b) Paasche's Method:

$$Q_{0_1} =$$

(c) Fisher's Method:

$$Q_{0_1} = \sqrt{L \times P \times 100} = \sqrt{108.23 \times 106.31} = \sqrt{115.593} = 107.27$$

Value Index Numbers Value index numbers are easy to calculate. Here, the value is the product of price and quantity. The value Index or V is the sum of the values of a given year divided by the sum of the values of the base year. The formula under this method is

$$\text{Value Index}, \quad V = \frac{\sum P_1 q_1}{\sum P_0 q_0} \times 100 \quad \text{or} \quad V = \frac{\sum V_1}{\sum V_0}$$

Where P_1 = Price current year

P_0 = Price for base year

q_1 = Quantity for current year

q_0 = Quantity for base year

V_1 = Total values of all commodities in the given period

V_0 = Total values of all commodities in the base period.

11.3.5 Test of Consistency of Index Numbers

Index numbers have been constructed through various formulae. Hence the consistency of Index numbers so calculated should be found out. There are various tests available to find out the consistency of Index Numbers. The important among them are

- | | |
|-------------------------|-----------------------|
| 1. Unit Test | 2. Time Reversal Test |
| 3. Factor Reversal Test | 4. Circular Test |

Unit Test This test requires the index number formula should be independent of the units in which the prices or quantities of various commodities are quoted.

Time Reversal Test This test is advocated by Prof. Irvin Fisher to test the consistency of the Index Number. According to him, the formula for calculating the Index Number should be such that it will give the same ratio for one point of comparison and the other. It can be represented as $P_{01} \times P_{10} = 1$.

where P_{01} means the Index number for time '1' on time '0' as base and P_{10} means the index for time '0' on time '1' as base.

When this time reversal test is applied to all the method for calculating index numbers like Laspeyre's method, Paasche's method, Fisher's method etc. Only Fisher's Ideal method satisfies a time reversal test.

Index number under Fisher's method is

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

Hence, $P_{10} = \sqrt{\frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}}$

$$P_{01} \times P_{10} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}} = \sqrt{1} = 1$$

Factor Reversal Test This test is also recommended by Fisher to test the consistency of Index Number. According to this test, the product of price Index and the quantity Index should be equal to the value Index. It can be represented as

$$V = P_{01} \times Q_{01}$$

$$V = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}}$$

$$Q_{01} = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

Hence,

$$= \sqrt{\frac{(\sum P_1 q_1)^2}{(\sum P_0 q_0)^2}} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

It is clear from the calculations, that only Fisher's Ideal method satisfies the factor reversal test (i.e., $V = P_{01} \times Q_{01}$).

Circular Test Circular Test is applied for testing the consistency of Index numbers especially for the Index number of the changes in price over a period of years.

According to the test,

$$P_{01} \times P_{12} \times P_{23} \dots P_{(n-1)n} \times P_{n0} = 1$$

It is proved that Laspeyre's method, Fisher's method and Paasche's method won't satisfy the circular test. Only simple aggregative method and the fixed weight aggregative method satisfy the circular test.

The circular test based on the simple aggregative method is calculated as follows.

$$\frac{\sum P_1}{\sum P_0} \times \frac{\sum P_2}{\sum P_1} \times \frac{\sum P_0}{\sum P_2} = 1$$

The circular test based on the fixed weight aggregative method is

$$\frac{\sum P_1 q}{\sum P_0 q} \times \frac{\sum P_2 q}{\sum P_1 q} \times \frac{\sum P_0 q}{\sum P_2 q} = 1$$

Illustration 11.9

From the following data, calculate the price index number through Fisher's Ideal method and find out the consistency of the Index Number by time reversal test and factor reversal test.

Commodity	2006		2007	
	Price	Quantity	Price	Quantity
A	9	20	11	12
B	10	25	12	13
C	7	41	9	20
D	13	40	14	35
E	11	30	13	21

Solutions

Calculation of Index Numbers.

Commodity	2006		2007		$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
	Price P_0	Qty q_0	Price P_1	Qty q_1				
A	9	20	11	12	220	180	132	108
B	10	25	12	13	300	250	156	130
C	7	41	9	20	369	287	180	140
D	13	40	14	35	560	520	490	455
E	11	30	13	21	390	330	273	231
					1839	1567	1231	1064

Index number by Fisher's Ideal Number

$$\begin{aligned}
 P_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0}} \times \sqrt{\frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \\
 &= \sqrt{\frac{1839}{1567}} \times \sqrt{\frac{1231}{1064}} \times 100 \\
 &= \sqrt{1.174} \times \sqrt{1.157} \times 100 \\
 &= \sqrt{1.358} \times 100 = 116.53
 \end{aligned}$$

Test of consistency of Index number

(a) Time Reversal test

$$\text{i.e., } P_{01} \times P_{10} = 1$$

$$\begin{aligned}
 P_{10} \times P_{01} &= \sqrt{\frac{\sum P_0 q_0}{\sum P_1 q_0}} \times \sqrt{\frac{\sum P_0 q_1}{\sum P_1 q_1}} \times \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0}} \times \sqrt{\frac{\sum P_1 q_1}{\sum P_0 q_1}} \\
 &= \sqrt{\frac{1567}{1839}} \times \sqrt{\frac{1064}{1231}} \times \sqrt{\frac{1839}{1567}} \times \sqrt{\frac{1231}{1064}} \\
 &= \sqrt{1} = 1
 \end{aligned}$$

(b) Factor Reversal test :

$$\begin{aligned}
 P_{01} \times Q_{01} &= \frac{\sum P_1 q_1}{\sum P_0 q_0} \\
 P_{01} \times Q_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0}} \times \sqrt{\frac{\sum P_1 q_1}{\sum P_0 q_1}} \times \sqrt{\frac{\sum q_1 p_0}{\sum P_0 q_0}} \times \sqrt{\frac{\sum q_1 p_1}{\sum q_0 p_1}} \\
 &= \sqrt{\frac{1839}{1567}} \times \sqrt{\frac{1231}{1064}} \times \sqrt{\frac{1064}{1567}} \times \sqrt{\frac{1231}{1839}} \\
 &= \sqrt{\frac{1231 \times 1231}{1567 \times 1567}} = \frac{1231}{1567} \\
 \frac{\sum P_1 q_1}{\sum P_0 q_0} &= \frac{1231}{1567} = 0.711 \\
 P_{01} \times Q_{01} &= \frac{\sum P_1 q_1}{\sum P_0 q_0}
 \end{aligned}$$

it is also equal to $\frac{1231}{1567}$. Hence, factor reversal test is not satisfied by the given data.

Illustration 11.10

Calculate Index number through Fisher's Ideal Index and test the consistency of it by (a) Time reversal test and (b) Factor reversal test.

Commodity	1995		2003	
	Price	Quantity	Price	Quantity
Beef	25	20	100	25
Mutton	60	35	180	30
Chicken	50	38	100	40
Pork	20	12	50	10
Fish	40	12	120	15

Solutions

Calculation of Index number

Commodity	1995		2003		$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
	Price $P_0 q_0$	Qty P_1	Price P_1	Qty q_1				
Beef	25	20	100	25	2000	500	2500	625
Mutton	60	35	180	30	6300	2100	5400	1800
Chicken	50	38	100	40	3800	1900	4000	2000
Pork	20	12	50	10	600	240	500	200
Fish	40	12	120	15	1440	480	1800	600
					$\sum P_1 q_0$ = 14140	$\sum P_0 q_0$ = 5220	$\sum P_1 q_1$ = 14200	$\sum P_0 q_1$ = 5225

Index number by Fisher's Ideal Index number:

$$\begin{aligned}
 P_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \\
 &= \sqrt{\frac{14140}{5220} \times \frac{14200}{5225}} \times 100 \\
 &= \sqrt{2.7088 \times 2.7177} \times 100 \\
 &= \sqrt{7.3617} \times 100 = 271.33
 \end{aligned}$$

458 Business Statistics

Test of consistency of Index number

(a) Time Reversal Test

$$P_{10} \times P_{01} = 1$$

$$\begin{aligned} P_{10} \times P_{01} &= \sqrt{\frac{\sum P_0 q_0}{\sum P_1 q_0} \times \frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \\ &= \sqrt{\frac{5220}{14140} \times \frac{5225}{14200} \times \frac{14140}{5220} \times \frac{14200}{5225}} \end{aligned}$$

$$P_{10} \times P_{01} = 1$$

(b) Factor Reversal test

$$\begin{aligned} P_{01} \times Q_{01} &= \frac{\sum P_1 q_1}{\sum P_0 q_0} \\ P_{01} \times Q_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_1 q_1} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} \\ &= \sqrt{\frac{14140}{5220} \times \frac{14200}{5225} \times \frac{5225}{5220} \times \frac{14200}{14140}} \\ &= \sqrt{\frac{14200 \times 14200}{5220 \times 5225}} = \frac{14200}{5220} \\ \frac{\sum P_1 q_1}{\sum P_0 q_0} &= \frac{14200}{5220} \end{aligned}$$

i.e., the index number so calculated is consistent one.

Chain Base Method In the fixed base index number, the base remains the same for all the series of the index. In practice, the base once selected may not be suitable for another period due to various reasons such as changes in economic conditions, social behaviour of people etc.

The chain base method is the most suitable method for calculating the index numbers. The index number should be constructed by treating the previous year as the base year. Hence, the base year will be changing year by year.

Steps for the Construction of Chain Indices

1. Calculate the link relatives which is equal to

$$\frac{\text{Current year price}}{\text{Previous year price}} \times 100$$

2. Chain index can be calculated through the following formula

$$\text{Chain index} = \frac{\text{Current year link relative} \times \text{Previous year chain index}}{100}$$

Merits

1. The chain base method helps economists and businessmen to know the extent of change that has come in the current year as compared to the previous year.
2. It helps to take decisions for the management of a company.
3. Weights can be adjusted as frequently as possible; it won't be affected by seasonal fluctuations.

Demerits

1. It is a very difficult task to select the items to be included or omitted.
2. Under this method, the long-term changes in prices can't be calculated.

Illustration 11.11

Convert the following fixed base index numbers into chain base index numbers.

Year	2001	2002	2003	2004	2005	2006	2007
F.B.I	376	392	408	380	392	400	410

Solutions

Calculation of chain base index numbers.

Year	F.B.I	Link Relatives	Chain Index
2001	376		376
2002	392	$\frac{392}{376} \times 100 = 104.3$	$\frac{104.3 \times 376}{100} = 392.2$
2003	408	$\frac{408}{392} \times 100 = 104.1$	$\frac{104.1 \times 392.2}{100} = 408.3$
2004	380	$\frac{380}{408} \times 100 = 93.1$	$\frac{93.1 \times 408.3}{100} = 380.1$
2005	392	$\frac{392}{380} \times 100 = 103.2$	$\frac{103.2 \times 380.1}{100} = 392.3$
2006	400	$\frac{400}{392} \times 100 = 102.0$	$\frac{102 \times 392.3}{100} = 400.1$
2007	410	$\frac{410}{400} \times 100 = 102.5$	$\frac{102.5 \times 400.1}{100} = 410.1$

Conversion of Chain Base Index into Fixed Base Index Sometimes, it is necessary to convert the chain base index into fixed base index number. In that case, the following formula should be applied.

460 Business Statistics

Current year's fixed base index (F.B.I)

$$= \frac{\text{Current year CBI} \times \text{Previous year F.B.I}}{100}$$

(C.B.I. = Chain Base Index)

(F.B.I. = Fixed Base Index)

Illustration 11.12

Convert the following chain base index numbers into fixed base index numbers.

Year	2002	2003	2004	2005	2006	2007
C.B.I.	140	170	180	150	200	210

Solutions

Conversion of C.B.I. into F.B.I

Year	Chain base Index no	Conversion of CBI numbers with 2002 as base	F.B.I. numbers
2002	140		140
2003	170	$\frac{140}{100} \times 170$	238
2004	180	$\frac{140}{100} \times \frac{170}{100} \times 180$	428.4
2005	150	$\frac{140}{100} \times \frac{170}{100} \times \frac{180}{100} \times 150$	642.6
2006	200	$\frac{140}{100} \times \frac{170}{100} \times \frac{180}{100} \times \frac{150}{100} \times 200$	1285.2
2007	210	$\frac{140}{100} \times \frac{170}{100} \times \frac{180}{100} \times \frac{150}{100} \times \frac{200}{100} \times 210$	2698.92

11.3.6 Base Shifting

It means calculation of index number based on new base. The base has to be shifted from old one to new due to the fact that old base may become outmoded. The base can be shifted through the following formula.

Index Number (Based on New Base year)

$$= \frac{\text{Current year old Index Number}}{\text{New base year's old Index Number}} \times 100$$

Illustration 11.13

From the following Index numbers based on 1995. Find out the new Index number based on 1999.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003
Index Number	120	150	160	180	200	200	210	240	270

Solutions

Calculation of New Index Number

Year	Old Index number	New Index number based on 1999
1995	120	$\frac{120}{200} \times 100 = 60$
1996	150	$\frac{150}{200} \times 100 = 75$
1997	160	$\frac{160}{200} \times 100 = 80$
1998	180	$\frac{180}{200} \times 100 = 90$
1999	200	$\frac{200}{200} \times 100 = 100$
2000	200	$\frac{200}{200} \times 100 = 100$
2001	210	$\frac{210}{200} \times 100 = 105$
2002	240	$\frac{240}{200} \times 100 = 120$
2003	270	$\frac{270}{200} \times 100 = 135$

Illustration 11.14

Calculate the Index number based on 2000 from the following Index Number based on 1997.

Year	1997	1998	1999	2000	2001	2002	2003	2004
Index Number	100	145	175	190	200	220	235	250

Solutions

Calculation of Index number based on 1997

Year	Old Index number based on 1997	New Index number based on 2000
1997	100	$\frac{100}{190} \times 100 = 52.6$
1998	145	$\frac{145}{190} \times 100 = 76.3$
1999	175	$\frac{175}{190} \times 100 = 92.1$
2000	190	$\frac{190}{190} \times 100 = 100$
2001	200	$\frac{200}{190} \times 100 = 105.3$
2002	220	$\frac{220}{190} \times 100 = 115.8$
2003	235	$\frac{235}{190} \times 100 = 123.7$
2004	250	$\frac{250}{190} \times 100 = 131.6$

Illustration 11.15

Calculate Index Number based on (i) 2000 and (ii) 2002 from the following Index Number based on 1995.

Year	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Index	195	200	212	234	248	252	267	275	282	295	300

Solutions

Calculation of New Index Number

Year	For base 1995	Index numbers for base	
		2000	For base 2002
1995	195	$\frac{195}{252} \times 100 = 77.4$	$\frac{195}{275} \times 100 = 70.9$

Year	For base 1995	Index numbers for base 2000	For base 2002
1996	200	$\frac{200}{252} \times 100 = 79.4$	$\frac{200}{275} \times 100 = 72.7$
1997	212	$\frac{212}{252} \times 100 = 84.1$	$\frac{212}{275} \times 100 = 77.1$
1998	234	$\frac{234}{252} \times 100 = 92.9$	$\frac{234}{275} \times 100 = 85.1$
1999	248	$\frac{248}{252} \times 100 = 98.4$	$\frac{248}{275} \times 100 = 90.1$
2000	252	$\frac{252}{252} \times 100 = 100$	$\frac{252}{275} \times 100 = 91.6$
2001	267	$\frac{267}{252} \times 100 = 106$	$\frac{267}{275} \times 100 = 97.1$
2002	275	$\frac{275}{252} \times 100 = 109.1$	$\frac{275}{275} \times 100 = 100$
2003	282	$\frac{282}{252} \times 100 = 111.9$	$\frac{282}{275} \times 100 = 102.5$
2004	295	$\frac{295}{252} \times 100 = 117$	$\frac{295}{275} \times 100 = 107.3$
2005	300	$\frac{300}{252} \times 100 = 119$	$\frac{300}{275} \times 100 = 109.1$

11.3.7 Splicing of Two Index Numbers

Splicing means combination of two Index Numbers. The commodity included in an Index Number may become out of fashion. New commodity may come into existence which replaces the old one.

Hence, a new Index Number for the new article may be available. It should be combined with the existing Index Numbers. The procedure of combination of these two Index Numbers is called splicing.

The formula for splicing of Index Number is

$$= \frac{\text{Index No. of current year} \times \text{old Index No. of new base year}}{100}$$

Illustration 11.16

Following are the two sets of Index one with 1990 as base and the other with 1997 as base.

Year I	Index Number	Year II	Index Number
1990	100	1997	100
1991	120	1998	115
1992	145	1999	125
1993	135	2000	120
1994	145	2001	135
1995	160	2002	130
1996	140		
1997	150		

The old Index was discontinued in the year 1997 and New Index Number with 1997 base was calculated. Splice the second Index Numbers with the First Index Numbers to get a continuous series of Index Numbers from 1990 to 2002.

Solutions

Splicing of Index numbers

Year	Index Numbers with 1990 as base	Index numbers with 1997 as base		Index numbers II spliced to I with 1997 as base
		I	II	
1990	100			
1991	120			
1992	145			
1993	135			
1994	145			
1995	160			
1996	140			
1997	150	100	$100 \times \frac{150}{100} = 150$	
1998		115	$115 \times \frac{150}{100} = 172.5$	
1999		125	$125 \times \frac{150}{100} = 187.5$	
2000		120	$120 \times \frac{150}{100} = 180$	
2001		135	$135 \times \frac{150}{100} = 202.5$	
2002		130	$130 \times \frac{150}{100} = 195$	

This process is to be more useful because recent year can be kept as a base. However, much would depend upon the object.

11.3.7 Deflating of Index Numbers

Deflating of Index number means correction of Index number according to changes in price level. It measures the decrease in the purchasing power of money due to the increase in price level. It is usually applied for calculating real wages.

Wages earned by workers should be compared with the price Index in order to find out the real wages. It can be calculated as follows:

$$\text{Real wage} = \frac{\text{Money wage}}{\text{Price Index}} \times 100$$

Real wage of Income Index Number,

$$= \frac{\text{Money wage}}{\text{Price Index}} \times 100$$

or

$$= \frac{\text{Real wage of the current year}}{\text{Real wage of the base year}} \times 100$$

Illustration 11.17

Calculate Real wages and Index number for real wages from the following.

Year	1986	1987	1988	1989	1990	1991	1992	1993
Wages	900	1100	1200	1400	1600	1800	1900	2000
Consumer price Index	100	120	150	175	195	200	210	220

Solutions

Calculation of Real Wage Index Number

Year	Wage Rs.	Price Index Numbers	Index of Wages	Real wage Index	Real Wage
1986	900	100	$\frac{100}{900} \times 900 = 100$	$\frac{100}{100} \times 100 = 100$	$\frac{900}{100} \times 100 = 900$
1987	1100	120	$\frac{100}{900} \times 1100 = 122.2$	$\frac{122.2}{120} \times 100 = 101.8$	$\frac{1100}{120} \times 100 = 916.7$
1988	1200	150	$\frac{100}{900} \times 1200 = 133.3$	$\frac{133.3}{150} \times 100 = 88.9$	$\frac{1200}{150} \times 100 = 800$
1989	1400	175	$\frac{100}{900} \times 1400 = 155.6$	$\frac{155.6}{175} \times 100 = 88.9$	$\frac{1400}{175} \times 100 = 800$
1990	160	195	$\frac{100}{900} \times 1600 = 177.8$	$\frac{177.8}{195} \times 100 = 91.2$	$\frac{1600}{195} \times 100 = 820.5$

Year	Wage	Price	Index of Wages	Real wage Index	Real Wage
	Rs.	Index			
			Numbers		
1991	1800	200	$\frac{100}{900} \times 1800 = 200$	$\frac{200}{200} \times 100 = 100$	$\frac{1800}{200} \times 100 = 900$
1992	1900	210	$\frac{100}{900} \times 1900 = 211.1$	$\frac{211.1}{210} \times 100 = 100.5$	$\frac{1900}{210} \times 100 = 904.8$
1993	2000	220	$\frac{100}{900} \times 2000 = 222.2$	$\frac{222.2}{220} \times 100 = 101$	$\frac{2000}{220} \times 100 = 909$

This method is frequently used to deflate individual values, value series. Its special use is in problems dealing with such diversified things as Rupees Sales of manufacturer's, wholesaler's and relatailer's wages, income and the like.

Consumer Price Index Numbers Consumer price index number was also known as 'Cost of Living Index number' or price of Living Index Number or cost of living price index. In different countries, cost of living index, consumer price index and retail price index are used.

Consumer price index numbers are designed to measure the average change over time in the price paid by the ultimate consumer for a specified quantity of goods and services. Consumer price indices or cost of living Indices measure the change in the cost of living of workers due to change in the retail price. A change in the price level affects the cost of living of different classes of people differently. The general index number fails to reveal this. So, there is a need to construct consumer price index.

The scope of consumer price is necessary to specify the population groups covered; for example, working class, middle class, poor class, rich class etc. and the geographical areas must be covered as urban, rural, city, town etc.

Objectives of Constructing Consumer Price Index Numbers The important objectives in the construction of consumer price index number are stated below:

1. It helps the government to formulate wage policy, price policy and other economic policies.
2. The minimum wage and rate of dearness allowance can be fixed by the government only through the construction of consumer price index number.
3. It helps to measure the purchasing power of money.
4. The real income of the money can be easily measured.
5. It helps to study the rate of changes of price of basic commodities over different periods of time.

Steps or Precautions in the Construction of Consumer Price Index Number The following are the important steps to be followed while constructing consumer Price Index Number:

Determination of the class of people for whom the Index number is to be constructed The people in the society may be classified in to different groups such as low income group, middle income group and high income group. They may be further classified into people living in urban areas, and rural areas. Hence, the class of people for whom the Index number is to be constructed should be determined before constructing the Index Number.

Selection of base period The base period shall be a period of economic stability. The consumption pattern over the long period should remain same. Base period should be taken as one complete year to avoid seasonal fluctuation in that year.

Conducting family budget enquiry The consumption pattern of the class of people for whom index number is to be calculated should be known. The expenses spent by the whole family for various important commodities should be enquired. Only those commodities which are commonly used by all families should be selected.

These enquires are conducted through any one of the different methods of data collection. Since the same class of people may be large in number, sample study may be conducted. While selecting the sample, random sampling technique may be adopted.

Problems in the Construction of Cost of Living Index Number

Determination of the purpose for which index number is being constructed Determination of the purpose for which Index Number is to be constructed is a complicated one. There are various purposes for which Index Number may be constructed. For example, to fix wages, to know the rate of change of price, to know the real income etc.

Problems in selection of the commodities The commodities commonly used by the same group of people for whom study is to be conducted should be selected. Since consumption pattern of each families are not identical, it is difficult to choose the commodities. The selection of commodities should remain uniform in quality from year to year.

Selection of sources of data For collection of data the selection of market should be popular and well known in the localities selected for the study. The markets selected should charge uniform prices for all commodities selected for the study.

Methods of collection of data The data may either be collected from primary source or secondary source. While data are collected from the primary source well qualified and experienced persons should be employed for that purpose. The secondary data can be collected from popular magazines, statistical department, Indian chamber of commerce and industry.

Problems in selection of base year The base period shall be a period of economic stability. It should not be too far in the past. It should be taken as one complete year to avoid seasonal fluctuations in that year.

Problems in the method of combination data The data collected are usually numerous in number due to the number of sample and the number of details. It remains a complicated task to combine these data. Hence, proper care should be needed for totalling and averaging the data so collected.

Problems in the system of weight The system of weight depends upon the commodities, type of weight and time of weight. The economic importance of the commodities for which weight is to be conducted determines the type of weight. The type of weight may either be quantity weight or value weight.

The quantity weight is the amount of commodity consumed. Value weight is the product of price and quantity consumed. The time of weight is the different periods for which weight have to be taken, that is, base period and current period.

Methods of Constructing Consumer Price Index Consumer price index may be constructed through any one of the following two methods.

Aggregate expenditure method or aggregative method Aggregative method is based on the Laspeyre's method. In this method, the quantities of commodities consumed by a particular group in the base year are fixed as weights.

The formula under this method is consumer price index number

$$= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

where P_1 = Price of current year

P_0 = Price of base year

q_0 = Quantity of base year

Family budget method or method of weighted relatives Family budget method is based on the method of weighted average of price relatives.

The formula under this method is consumer price index = $\frac{\sum PV}{\sum V}$

where $P = \frac{P_1}{P_0} \times 100$; V = Value weight ($P_0 q_0$)

P_1 = Price of current year

P_0 = Price of base year

Illustration 11.18

Calculate index number through aggregative expenditure method from the following data.

Commodities	Quantity consumed in 1997	Price per unit in 1997	Price per unit in 2007
A	200	20	25
B	230	15	30
C	240	30	40
D	120	10	20
E	110	5	15
F	80	12	30
G	70	7	25

Solutions

Calculation of Index Number

Commodity	Q_0	P_0	P_1	$P_1 q_0$	$P_0 q_0$
A	200	20	25	5000	4000
B	230	15	30	6900	3450
C	240	30	40	9600	7200
D	120	10	20	2400	1200
E	110	5	15	1650	550
F	80	12	30	2400	960
G	70	7	25	1750	490
			$\sum P_1 q_0 = 29700$	$\sum P_0 q_0 = 17850$	

$$\begin{aligned}
 P_{01} &= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 \\
 &= \frac{29700}{17850} \times 100 = 166.38655
 \end{aligned}$$

Illustration 11.19

Calculate Index number on the basis of family budget method from the following data.

Commodity	Weight	Price per unit 1993 (Rs)	Price per unit 2002 (Rs)
A	60	54	70
B	40	30	60
C	20	26	56
D	30	22	104

Commodity	Weight	Price per unit 1993 (Rs)	Price per unit 2002 (Rs)
E	50	44	120
F	20	60	140

Solutions

Calculation of Index Number

Commodity	Weight	Price per unit in 1993 (P_0)	Price per unit in 2002 (P_1)	$P = \frac{P_1}{P_0} \times 100$	Weighted price PV
A	60	54	70	129.6	7776
B	40	30	60	200	8000
C	20	26	56	215.38	4307.6
D	30	22	104	472.72	14181.6
E	50	44	120	272.72	13636
F	20	60	140	233.330	4666.6
$\Sigma V = 220$		$\Sigma PV = 52567.8$			

$$\text{Index Number } P_{01} = \frac{\Sigma PV}{\Sigma V} = \frac{52567.8}{220} = 238.94$$

Illustration 11.20

Calculate consumer price index through aggregate expenditure method from the following data.

Commodity	Quantity consumed in 1996	Price per unit in 1996	Price per unit in 2006
I	50	30	80
II	40	40	120
III	80	24	100
IV	100	36	150
V	60	50	200

Solutions

Calculation of consumer price index number.

Commodity	q_0	P_0	P_1	$P_1 q_0$	$P_0 q_0$
I	50	30	80	4000	1500
II	40	40	120	4800	1600
III	80	24	100	8000	1920
IV	100	36	150	15000	3600
V	60	50	200	12000	3000
				43800	11620

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

$$= \frac{43800}{11620} \times 100 = 376.94$$

Illustration 11.21

Calculate the consumer price index number through family Budget method for the following data.

Commodity	Weight	Price per unit in 1995 (Rs)	Price per unit in 2005 (Rs)
A	50	30	50
B	60	24	48
C	35	16	25
D	40	10	30
E	20	40	100
F	10	12	60

Solutions

Calculation of consumer price index number

Commodity	Weight (V)	Price per unit in 1995 (P_0)	Price per unit in 2005 (P_1)	$P = \frac{P_1}{P_0} \times 100$	Weighted Relative PV
A	50	30	50	166.66	8333
B	60	24	48	200	12000
C	35	16	25	156.25	5468.75
D	40	10	30	300	12000
E	20	40	100	250	5000
F	10	12	60	500	5000
$\Sigma V = 215$				ΣPV	
				= 47801.75	

$$\text{Index number } P_{01} = \frac{\sum PV}{\sum V} = \frac{47801.75}{215} = 222.33$$

11.4 MISCELLANEOUS ILLUSTRATIONS

1. Construct an index number for the year 2006 taking 2005 as base.

Commodity	Price 2005	Price 2006
Wheat	180	190
Sugar	190	200
Rice	200	205
Cotton	210	215

Solutions

Construction of Price Index

Commodity	Price 2005 P_0	Price 2006 P_1
Wheat	180	190
Sugar	190	200
Rice	200	205
Cotton	210	215
	$\Sigma P_0 = 780$	$\Sigma P_1 = 810$

$$\begin{aligned} P_{01} &= \frac{\Sigma P_1}{\Sigma P_0} \times 100 \\ &= \frac{810}{780} \times 100 = 103.85 \end{aligned}$$

2. Calculate price Index for the following by

- (a) Simple aggregate
- (b) Average of price relative method by using both arithmetic mean and Geometric mean.

Commodity	Price in 2005 (Rs)	Price in 2006 (Rs)
A	30	25
B	35	38
C	25	40
D	40	35
E	50	45
F	60	70

Solutions

$$\text{1. Simple Aggregative Price Index} = \frac{\sum P_1}{\sum P_0} \times 100$$

$$\sum P_0 = 240 ; \sum P_1 = 253$$

$$= \frac{253}{240} = 105.42$$

Calculation for Price Index

Commodity	Price in 2005 (Rs) P_0	Price in 2006 (Rs) P_1	Price relative $= P_1 \{P_1 + P_0\} \times 100$	$\log P$
A	30	25	83.33	1.9208
B	35	38	108.57	2.0357
C	25	40	160.00	2.2041
D	40	35	87.50	1.9420
E	50	45	90.00	1.9542
F	60	70	116.67	2.0669
	$\Sigma p_0 = 240$	$\Sigma p_1 = 253$	$\Sigma p = 646.07$	$\Sigma \log p = 12.1237$

2. (i) Arithmetic mean of price

$$\text{Relatives} = \frac{\sum P_{01}}{N}$$

$$\sum P_{01} = 646.07; N = 6$$

$$\frac{646.07}{6} = 107.68$$

(ii) Geometric mean of price

$$= \text{Antilog} \left[\frac{\sum \log P}{N} \right]$$

$$= \text{Antilog} \left[\frac{12.1237}{6} \right]$$

Relatives Index

$$= \text{Antilog } 2.0206$$

$$= 104.9$$

3. Calculate index number from the following data by

- (a) Laspeyre's Method
- (b) Paasche's Method
- (c) Bowley's Method

- (d) Fisher's Ideal Method and
- (e) Marshall-Edgeworth Method

Commodity	Base year		Current year	
	Kilo	Rate (Rs)	Kilo	Rate (Rs)
Coffee	10	5	15	6
Bread	20	8	18	9
Tea	15	4	16	5
Milk	18	9	20	10

Solutions

Construction of Price Index

Commodity	Base year				$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
	Kilo q_0	Rate (Rs) P_0	Kilo q_1	Rate (Rs) P_1				
Coffee	10	5	15	6	60	50	90	75
Bread	20	8	18	9	180	160	162	144
Tea	15	4	16	5	75	60	80	64
Milk	18	9	20	10	180	162	200	180
				Total	$\Sigma P_1 q_0 = 495$	$\Sigma P_0 q_0 = 432$	$\Sigma P_1 q_1 = 532$	$\Sigma P_0 q_1 = 463$

- (a) Laspeyre's Method

$$\begin{aligned} P_{01} &= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 \\ &= \frac{495}{432} \times 100 = 114.6 \end{aligned}$$

- (b) Paasche's Method

$$\begin{aligned} P_{01} &= \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100 \\ &= \frac{532}{463} \times 100 = 114.9 \end{aligned}$$

- (c) Bowley's Method

$$P_{01} = \frac{\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}}{2} \times 100$$

$$= \frac{\frac{495}{432} + \frac{532}{463}}{2} \times 100 = 114.75 \quad [\text{or}]$$

$$P_{01} = \frac{L+P}{2} = \frac{114.6+114.9}{2} = 114.75$$

(d) Fisher's Ideal Method

$$\begin{aligned} P_{01} &= \sqrt{L \times P} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \\ &= \sqrt{\frac{495}{432} \times \frac{532}{463}} \times 100 \\ &= 1.1475 \times 100 = 114.75 \end{aligned}$$

(e) Marshall-Edgeworth Method

$$\begin{aligned} P_{01} &= \frac{\sum P_1 q_0 + \sum P_1 q_1}{\sum P_0 q_0 + \sum P_0 q_1} \times 100 \\ &= \frac{495 + 532}{432 + 463} \times 100 = 114.75 \end{aligned}$$

4. From the following particulars calculate price index numbers for 2006 with 2000 as base by
- Laspeyre's Method
 - Paasche's Method
 - Marshall-Edgeworth Method
 - Fisher Ideal Index Method

Commodity	2000		2006	
	Price	Qty	Price	Qty
101	20	16	36	20
102	30	24	40	22
103	25	18	48	30
104	35	12	52	20
105	40	22	50	24

Solutions

Calculation of the Price Index

Commodity	2000		2006		$P_1 q_0$	$P_0 q_0$	$P_1 q_1$	$P_0 q_1$
	Price (P_0)	Qty (q_0)	Price (P_1)	Qty (q_1)				
101	20	16	36	20	576	320	720	400
102	30	24	40	22	960	720	880	660
103	25	18	48	30	864	450	1440	750
104	35	12	52	20	624	420	1040	700
105	40	22	50	24	1100	880	1200	960
		Total	$\Sigma P_1 q_0 = 4124$	$\Sigma P_0 q_0 = 2790$	$\Sigma P_1 q_1 = 5280$	$\Sigma P_0 q_1 = 3470$		

(a) Laspeyre's Method

$$\begin{aligned} P_{01} &= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 \\ &= \frac{4124}{2790} \times 100 = 147.8 \end{aligned}$$

(b) Paasche's Method

$$\begin{aligned} P_{01} &= \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100 \\ &= \frac{5280}{3470} \times 100 = 152.2 \end{aligned}$$

(c) Bowley's Method

$$\begin{aligned} P_{01} &= \frac{\frac{\sum P_1 q_1}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}}{2} \times 100 \\ &= \frac{\frac{4124}{2790} + \frac{5280}{3470}}{2} \times 100 \\ &= \frac{1.478 + 1.522}{2} \times 100 \\ &= 150 \end{aligned}$$

(d) Fisher's Ideal Method

$$\begin{aligned} P_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \\ &= \sqrt{\frac{4124}{2790} \times \frac{5280}{3470}} \times 100 = 1.4998 \times 100 = 149.98 \end{aligned}$$

6. Compute price index for the following data by applying weighted average of price relative method, using (i) Arithmetic mean (ii) Geometric mean.

Commodity	Price in 2006 (Rs)	Price in 2007 (Rs)	Qty in 2006 (Rs)
Sugar	10	12	50
Wheat	12	15	30
Milk	13	16	20

Solutions

Construction of Price Index

Item	Price in 2006 (Rs) P_0	Price in 2007 (Rs) P_1	Qty in 2006 (Rs) q_0	V $= P_0 q_0$	P $= P_1 / P_0 \times 100$	PV	logP	logPV
Wheat	10	12	50	500	120	60000	4.7782	2389.1
Sugar	12	15	30	360	125	45000	4.6532	1675.2
Milk	13	16	20	260	123	31980	4.5049	1171.3
			Total	$\sum V = 1120$	$\sum P = 368$	$\sum PV = 136980$	$\sum \log PV = 5235.6$	

(b) Geometric mean

$$P_{01} = \text{Antilog } \frac{\sum \log PV}{\sum V}$$

$$\sum \log PV = 5235.6; \sum V = 1120$$

$$\begin{aligned} P_{01} &= \text{Antilog } \frac{5235.6}{1120} \\ &= \text{Antilog } 4.6746 \\ &= 47271.57 \end{aligned}$$

7. Calculate quantity index by (i) Laspeyre's Method (ii) Paasche's Method (iii) Fisher's Method.

Commodity	Price P_0	2005 Total Value ($P_0 q_0$)	Price P_1	2007 Total value ($P_1 q_1$)
A	12	144	14	196
B	10	100	12	144
C	14	196	18	324
D	16	256	20	400
E	20	400	25	625

Solutions

Construction of Price Index

Commodity	2005		2007		2005 Qty q_0	2007 Qty q_1	$P_0 q_1$	$P_1 q_0$
	Price (P_0)	Total value ($P_0 q_0$)	Price P_1	Total value ($P_1 q_1$)				
A	12	144	14	196	12	14	168	168
B	10	100	12	144	10	12	120	120
C	14	196	18	324	14	18	252	252
D	16	256	20	400	16	20	320	320
E	20	400	25	625	20	25	500	500
	$\sum P_0 q_0 = 1096$		$\sum P_1 q_1 = 1689$				$\sum P_0 q_1 = 1360$	$\sum P_1 q_0 = 1360$

(a) Laspeyre's Method

$$\begin{aligned} L &= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 \\ &= \frac{1360}{1096} \times 100 = 124.1 \end{aligned}$$

(b) Paasche's Method

$$\begin{aligned} P &= \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100 \\ &= \frac{1689}{1360} \times 100 = 124.2 \end{aligned}$$

(c) Fisher's Ideal Method

$$\begin{aligned} &= \sqrt{L \times P} = \sqrt{124.1 \times 124.2} = 124.15 \\ \text{or} \quad &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 = 124.15 \end{aligned}$$

8. From the following construct Fisher's ideal index and prove that the factor reversal test and time reversal test are satisfied by Fisher's ideal formula.

Commodity	2004		2007	
	Price	Qty	Price	Qty
O	10	60	12	70
P	8	65	9	60

Commodity	2004		2007	
	Price	Qty	Price	Qty
<i>Q</i>	12	70	10	75
<i>R</i>	9	65	11	65
<i>S</i>	14	40	13	45

Solutions

Commo- dities	2004		2005		$P_0 q_0$	$P_1 q_0$	$P_1 q_1$	$P_0 q_1$
	Price P_0	Qty q_0	Price P_1	Qty q_1				
<i>O</i>	10	60	12	70	600	720	840	700
<i>P</i>	8	65	9	60	520	585	540	480
<i>Q</i>	12	70	10	75	840	700	750	900
<i>R</i>	9	65	11	65	585	715	715	585
<i>S</i>	14	40	13	45	560	520	585	630
Total [Σ]				3105	3240	3430	3295	

Fisher's Ideal Index

$$\begin{aligned}
 P_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \\
 &= \sqrt{\frac{3240}{3105} + \frac{3430}{3295}} \times 100 \\
 &= \sqrt{1.086 \times 100} = 104.2
 \end{aligned}$$

Factor Reversal test is satisfied if

$$\begin{aligned}
 P_{01} \times Q_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum q_1 P_0}{\sum q_0 P_0} \times \frac{\sum q_0 P_1}{\sum q_1 P_1}} = \frac{\sum P_1 q_1}{\sum P_0 q_0} \\
 P_{01} \times Q_{01} &= \sqrt{\frac{3240}{3105} \times \frac{3430}{3295} \times \frac{3295}{3105} \times \frac{3430}{3240}} \\
 &= \sqrt{\frac{3430 \times 3430}{3105 \times 3105}} = \frac{3430}{3105}
 \end{aligned}$$

Now $\frac{\sum P_1 q_1}{\sum P_0 q_0}$ is also equal to $\frac{3430}{3105}$

Time Reversal test is satisfied if

$$P_{01} \times Q_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_1}{\sum P_1 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0}} = 1$$

$$= \sqrt{\frac{3240}{3105} \times \frac{3430}{3295} \times \frac{3295}{3430} \times \frac{3105}{3240}} = \sqrt{1} = 1$$

Hence, the above data satisfies the time-reversal test and factor-reversal tests.

9. Compute Index numbers using Fishers Ideal formula and show that it satisfies time-reversal test and factor-reversal test.

Commodity	Base Year		Current Year	
	Qty Kilo	Price Rs	Qty Kilo	Price Rs
Apple	14	50	16	60
Pineapple	16	30	14	34
Sugar	20	26	22	30
Orange	18	30	20	26

Solutions

Commodities	Base Price P_0	Year Qty q_0	Current Price P_1	Year Qty q_1	$P_0 q_0$	$P_1 q_0$	$P_1 q_1$	$P_0 q_1$
Apple	50	14	60	16	700	840	960	800
Pineapple	30	16	34	14	480	544	476	420
Sugar	26	20	30	22	520	600	660	572
Orange	30	18	26	20	540	468	520	600
Total [Σ]					2240	2452	2616	2392

Fisher's Ideal Index

$$P_{01} = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100$$

$$= \sqrt{\frac{2452}{2240} + \frac{2616}{2392}} \times 100$$

$$= \sqrt{1.1972} \times 100 = 109.42$$

- (i) Time Reversal Test

$$P_{01} \times P_{10} = 1 = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum P_0 q_0}{\sum P_1 q_0} \times \frac{\sum P_0 q_1}{\sum P_1 q_1}}$$

$$\begin{aligned} P_{01} \times P_{10} &= \sqrt{\frac{2452}{2240} \times \frac{2616}{2392} \times \frac{2240}{2452} \times \frac{2392}{2616}} \\ &= \sqrt{1} = 1 \end{aligned}$$

(ii) Factor Reversal Test

$$\begin{aligned} P_{01} \times Q_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1} \times \frac{\sum q_0 P_1}{\sum q_1 P_0} \times \frac{\sum q_1 P_1}{\sum q_0 P_1}} \\ &= \sqrt{\frac{2452}{2240} \times \frac{2616}{2392} \times \frac{2392}{2240} \times \frac{2616}{2452}} \\ &= \sqrt{\frac{2616}{2240} \times \frac{2616}{2240}} \\ P_{01} \times Q_{01} &= \frac{2616}{2240} \end{aligned}$$

The time reversal test and factor reversal test are satisfied by Fisher's ideal formula.

10. Convert the following fixed base index number into chain base index numbers.

Year	2001	2002	2003	2004	2005	2006
F.B.I.	215	230	245	240	250	247

Solutions

Conversion of F.B.I. to C.B.I.

Year	Fixed base index	Link Relatives	Chain index
2001	215	—	215
2002	230	$\frac{230}{215} \times 100 = 106.98$	$\frac{215 \times 106.98}{100} = 230$
2003	245	$\frac{245}{230} \times 100 = 106.52$	$\frac{230 \times 106.52}{100} = 245$
2004	240	$\frac{240}{245} \times 100 = 97.96$	$\frac{245 \times 97.96}{100} = 240$
2005	250	$\frac{250}{240} \times 100 = 104.17$	$\frac{240 \times 104.17}{100} = 250$
2006	247	$\frac{247}{250} \times 100 = 98.8$	$\frac{250 \times 98.8}{100} = 247$

482 Business Statistics

11. From the following data convert the F.B.I into chain base index numbers.

Year	2001	2002	2003	2004
F.B.I.	100	110	126	120

Solutions

Conversion of F.B.I. to C.B.I.

Year	F.B.I.	Link Relatives	C.B.I.
2001	100	—	100
2002	110	$\frac{110}{100} \times 100 = 110$	$\frac{100 \times 110}{100} = 110$
2003	126	$\frac{126}{110} \times 100 = 114.55$	$\frac{110 \times 114.55}{100} = 126$
2004	120	$\frac{120}{126} \times 100 = 95.24$	$\frac{126 \times 95.24}{100} = 120$

12. Convert the following chain base index numbers into fixed base index numbers.

Year	2002	2003	2004	2005	2006
C.B.I.	100	110	105	115	120

Solutions

Construction of Fixed Base Index

Year	C.B.I.	C.B.I. changed to F.B.I.	F.B.I.
2002	100	—	100
2003	110	$\frac{100}{100} \times 110$	110
2004	105	$\frac{110}{100} \times 105$	115.5
2005	115	$\frac{115.5}{100} \times 115$	132.83
2006	120	$\frac{132.83}{100} \times 120$	159.40

13. Convert the following C.B.I. into F.B.I. numbers.

Year	2002	2003	2004	2005	2006
C.B.I.	120	110	130	140	135

Solutions

Construction of Fixed Base Index

Year	C.B.I.	C.B.I changed to F.B.I.	F.B.I.
2002	120	—	120
2003	110	$\frac{120}{100} \times 110$	132
2004	130	$\frac{132}{100} \times 130$	171.60
2005	140	$\frac{171.6}{100} \times 140$	240.4
2006	135	$\frac{240.4}{100} \times 135$	324.32

14. Reconstruct the following index numbers by shifting the base to (i) 2005
(ii) 2007.

Year	2001	2002	2003	2004	2005	2006	2007
Index Nos.	100	110	130	150	140	170	180

Solutions

Year	Index Numbers	2005 as a base	2007 as a base
2001	100	$\frac{100}{140} \times 100 = 71.43$	$\frac{100}{180} \times 100 = 55.55$
2002	110	$\frac{110}{140} \times 100 = 78.57$	$\frac{110}{180} \times 100 = 61.11$
2003	130	$\frac{130}{140} \times 100 = 92.86$	$\frac{130}{180} \times 100 = 72.22$
2004	150	$\frac{150}{140} \times 100 = 107.14$	$\frac{150}{180} \times 100 = 83.33$
2005	140	$\frac{140}{140} \times 100 = 100$	$\frac{140}{180} \times 100 = 77.78$

484 Business Statistics

Year	Index Numbers	2005 as a base	2007 as a base
2006	170	$\frac{170}{140} \times 100 = 121.43$	$\frac{170}{180} \times 100 = 94.44$
2007	180	$\frac{180}{140} \times 100 = 128.57$	$\frac{180}{180} \times 100 = 100$

15. Reconstruct the following index numbers by shifting base to 2003.

Year	2001	2002	2003	2004	2005
Index No.	120	130	150	140	160

Solutions

Year	Index Numbers	2003 as base
2001	120	$\frac{120}{150} \times 100 = 80$
2002	130	$\frac{130}{150} \times 100 = 86.67$
2003	150	$\frac{150}{150} \times 100 = 100$
2004	140	$\frac{140}{150} \times 100 = 93.33$
2005	160	$\frac{160}{150} \times 100 = 106.67$

16. Two sets of Indices, one with 1998 as base and one units 1990 as base are given below:

(a) Year	Index Numbers	(b) Year	Index Numbers
1988	100		
1989	120		
1990	130	1990	100
1991	200	1991	110
1992	220	1992	120
1993	230	1993	115
1994	250	1994	100
1995	280	1995	120
1996	300	1996	130

The index (a) with 1988 base was discontinued in 1990. You are required to splice the second index numbers (b) with 1990 base to the first index number.

Solutions

Splicing of Index Numbers

Year	Index Number (a) with 1988 as base	Index Number (b) with 1990 as base	Index Number (b) spliced to (a) with 1988 as base
1988	100		
1989	120		
1990	130	100	$100 \times \frac{300}{100} = 300$
1991	200	110	$110 \times \frac{300}{100} = 330$
1992	220	120	$120 \times \frac{300}{100} = 360$
1993	230	115	$115 \times \frac{300}{100} = 345$
1994	250	100	$100 \times \frac{300}{100} = 300$
1995	280	120	$120 \times \frac{300}{100} = 360$
1996	300	130	$130 \times \frac{300}{100} = 390$

17. The following are the index numbers of wholesale prices of a commodity are given below :

Year	2001	2002	2003	2004	2005	2006	2007
Index No.	100	120	130	160	190	240	270

Solutions

Construction of Index Number by Base Shifting

Year	Index Number	Index Numbers Base (2004 = 160)
2001	100	$\frac{100}{160} \times 100 = 62.5$
2002	120	$\frac{120}{160} \times 100 = 75$
2003	130	$\frac{130}{160} \times 100 = 81.25$
2004	160	$\frac{160}{160} \times 100 = 100$
2005	190	$\frac{190}{160} \times 100 = 118.75$
2006	240	$\frac{240}{160} \times 100 = 150$
2007	270	$\frac{270}{160} \times 100 = 168.75$

18. The following data relating to weekly take home pay and consumer price index of a factory.

Year	Weekly take home pay	Consumer price index
2002	1200	125
2003	1300	135
2004	1500	140
2005	1700	145
2006	2100	160
2007	2300	175

- What was the real average weekly wage for each year?
- In which year did the employee have the greater buying power?

Solutions

Year	Weekly take home pay (Rs)	Consumer price index	Real wages
2002	1200	125	$\frac{1200}{125} \times 100 = 960$
2003	1300	135	$\frac{1300}{135} \times 100 = 962.96$

Year	Weekly take home pay (Rs)	Consumer price index	Real wages
2004	1500	140	$\frac{1500}{140} \times 100 = 1071.43$
2005	1700	145	$\frac{1700}{145} \times 100 = 1172.41$
2006	2100	160	$\frac{2100}{160} \times 100 = 1312.5$
2007	2300	175	$\frac{2300}{175} \times 100 = 1314.29$

(i) Real average weekly wage can be obtained by the formula

$$\text{Real wage} = \frac{\text{Money wage}}{\text{Price Index}} \times 100$$

(ii) The employee had the greatest buying power in 2007 [1314.29] as the real wage was maximum in 2007.

19. From the following data calculate the Index Numbers using the Aggregate Expenditure method for the year 2007 with 2005 as base year.

Commodity	Quantity in units	Price per unit in 2005 (Rs)	Price per unit in 2007 (Rs)
A	200	10	13
B	50	8	11
C	300	15	17
D	220	21	25
E	150	16	19
F	270	14	16

Solutions

Calculation of Index Number by using Aggregate Expenditure method.

Commodity	Qty in 2005 q_0	Price in 2005 P_0	Price in 2007 P_1	$P_1 q_0$	$P_0 q_0$
A	200	10	13	2600	2000
B	50	8	11	550	400
C	300	15	17	5100	4500
D	220	21	25	5500	4620
E	150	16	19	2850	2400
F	270	14	16	4320	3780
Total [Σ]				20920	17700

$$\begin{aligned}
 P_{01} &= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 \\
 &= \frac{20920}{17700} \times 100 = 118.19
 \end{aligned}$$

$$P_{01} = 118.19$$

20. Calculate index numbers of prices for 2007 on the basis of 2006 from the data given below.

Commodity	Weights	Price per unit 2006	Price per unit 2007
201	40	20	23
202	50	25	27
203	25	15	18
204	5	12	16
205	15	7	11

Solutions

Calculation of Index Numbers

Commodity	Weights [V]	Price 2006 P_0	Price 2007 P_1	Price relatives [$P = P_1/P_0 \times 100$]	Weighted relatives $P \times V$
201	40	20	23	115	4600
202	50	25	27	108	5400
203	25	15	18	120	3000
204	5	12	16	133.33	666.65
205	15	7	11	157.14	2357.1
$\Sigma V = 135$				$\Sigma PV =$	
				16023.75	

Index number of price for 2007

$$\begin{aligned}
 &= \frac{\sum PV}{\sum V} \\
 &= \frac{16023.75}{135} = 118.69
 \end{aligned}$$

Price index number for 2007 = Rs 118.69

21. An enquiring into the budgets of middle class families in a certain city gave the following information.

Expenses	%	Price (2006)	Price (2007)
Food	30	600	700
Fuel	15	300	400
Clothing	23	200	250
Rent	15	200	275
Miscellaneous	17	100	150

What is the cost of living index number of 2007 as compared with that of 2006?

Expenses	2006 P_0	2007 P_1	$P = \frac{P_1}{P_0} \times 100$	Percent Age = V	$P \times V$
Food	600	700	$\frac{700}{600} \times 100 = 116.67$	30	3500.1
Fuel	300	400	$\frac{400}{300} \times 100 = 133.33$	15	1999.95
Clothing	200	250	$\frac{250}{200} \times 100 = 125$	23	2875
Rent	200	275	$\frac{275}{200} \times 100 = 137.5$	15	2062.5
Miscellaneous	100	150	$\frac{150}{100} \times 100 = 150$	17	2550
Total [Σ]				$\sum V = 100$	12987.55

$$\begin{aligned} C.L. &= \frac{\Sigma PV}{\Sigma V} \\ &= \frac{12987.55}{100} = 129.87 \end{aligned}$$

Cost of living of 2007 = 129.87

22. Construct the consumer price index number for 2006 on the basis of 2005 from the following data using the aggregate expenditure method.

Commodity	Qty consumed 2005 kg	Price in 2005 Rs	Price in 2006 Rs
U	10	15	16
V	8	18	20
W	12	20	23
X	3	6	9
Y	9	5	8
Z	16	14	16

Solutions

Construction of consumer price index

Commodity	Qty q_0 (kg)	Price P_0	Price P_1	$P_1 q_0$	$P_0 q_0$
U	10	15	16	160	150
V	8	18	20	160	144
W	12	20	23	276	240
X	3	6	9	27	18
Y	9	5	8	72	45
Z	16	14	16	256	224
Total [Σ]				951	821

$$\begin{aligned}\text{Consumer price index} &= \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100 \\ &= \frac{951}{821} \times 100 \\ &= 115.83\end{aligned}$$

23. The following are the prices of commodities in 2000 and 2005. Calculate a price index based on price relatives using the geometric mean:

Year	Commodity					
	A	B	C	D	E	F
2000	9	12	4	10	17	24
2005	11	14	6	15	18	26

Solutions

Computation of Price Index

Commodity	2000 P_0	2005 P_1	$P = \frac{P_1}{P_0} \times 100$	Log P
A	9	11	122.22	2.0872
B	12	14	116.67	2.0669
C	4	6	150	2.1761
D	10	15	150	2.1761
E	17	18	105.88	2.0248
F	24	26	108.33	2.0348
Total [Σ]				12.5659

$$\begin{aligned}
 P_{01} &= \text{Antilog} \left[\frac{\sum \log P}{N} \right] \\
 &= \text{Antilog} \left[\frac{12.5659}{6} \right] \\
 &= \text{Antilog } 2.0948 \\
 &= 124.4
 \end{aligned}$$

24. The following table gives the money wages and cost of living index numbers based on 2000.

Year	2000	2001	2002	2003	2004	2005	2006	2007
Wages (Rs)	65	80	85	90	95	110	120	135
C.L.I	100	105	110	120	130	140	145	150

Calculate the real wage index numbers.

Solutions

Real wages for different years can be obtained by using the following formula.

$$\text{Real wage} = \frac{\text{Money wage}}{\text{Cost of Living Index}} \times 100$$

Real wage index numbers are calculated by taking 2000 as base.

Calculation of Real wage Index Numbers

Year	Wages (Rs)	Cost of Living Index	Real wage (Rs)	Real wages Index Numbers
2000	65	100	$\frac{65}{100} \times 100 = 65$	100
2001	80	105	$\frac{80}{105} \times 100 = 76.2$	$\frac{76.2}{65} \times 100 = 117.23$
2002	85	110	$\frac{85}{110} \times 100 = 77.27$	$\frac{77.27}{65} \times 100 = 118.88$
2003	90	120	$\frac{90}{120} \times 100 = 75$	$\frac{75}{65} \times 100 = 115.38$
2004	95	130	$\frac{95}{130} \times 100 = 73.08$	$\frac{73.08}{65} \times 100 = 112.43$
2005	110	140	$\frac{110}{140} \times 100 = 78.57$	$\frac{78.57}{65} \times 100 = 120.88$

Year	Wages (Rs)	Cost of Living Index	Real wage (Rs)	Real wages Index Numbers
2006	120	145	$\frac{120}{145} \times 100 = 82.76$	$\frac{82.76}{65} \times 100 = 127.32$
2007	135	150	$\frac{135}{150} \times 100 = 90$	$\frac{90}{65} \times 100 = 138.46$

25. It is stated that the Marshall–Edgeworth Index is a good approximations to the ideal index numbers. Verify this statement by using the following data.

Commodity	2006		2007	
	Price	Qty	Price	Qty
X	6	124	8	136
Y	7	536	10	510
Z	8	288	11	296

Solutions

Construction of Marshall–Edgeworth and Ideal Index Numbers.

Commodity	2006		2007		$P_0 q_0$	$P_1 q_0$	$P_1 q_1$	$P_0 q_1$
	P_0	q_0	P_1	q_1				
X	6	124	8	136	744	992	1088	816
Y	7	536	10	510	3752	5360	5100	3570
Z	8	288	11	296	2304	3168	3256	2368
			Total		$\sum P_0 q_0 = 6800$	$\sum P_1 q_0 = 9520$	$\sum P_1 q_1 = 9444$	$\sum P_0 q_1 = 6754$

Marshall-Edgeworth Index

$$\begin{aligned}
 &= \frac{\sum P_1 q_0 + \sum P_1 q_1}{\sum P_0 q_0 + \sum P_0 q_1} \times 100 \\
 &= \frac{9520 + 9444}{6800 + 6754} \times 100 \\
 &= \frac{18964}{13554} \times 100 \\
 &= 139.91
 \end{aligned}$$

Fisher's Ideal Index

$$\begin{aligned}
 &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \\
 &= \sqrt{\frac{9520}{6800} \times \frac{9444}{6754}} \times 100 \\
 &= \sqrt{1.4 \times 1.4} \times 100 \\
 &= 140
 \end{aligned}$$

Marshall–Edgeworth Index is a good approximation to the Fisher's Ideal index number.

SUMMARY

Index Number

It is a relative number which expresses the relationship between two figures over a period of time. It is ratio of changes between variables over a period of time.

Classification of Index Numbers

- Price Index numbers
- Quantity Index numbers
- Value Index numbers

Quantity Index Numbers

Deals with the quantity of goods produced or consumed over a period of time.

Price Index Numbers

Used for measuring the changes of prices of commodities over a period of time.

Value Index Numbers

It is concerned with changes in the total value of products in a specific year compared with a past year.

Uses of Index Numbers

- To measure trend values
- To facilitate for policy decisions
- It is used for deflating
- Index numbers are economic barometers
- It helps in comparing the standard of living.

Factors Considered before Constructing Index Numbers

- Purpose of Index Numbers
- Availability of data

- Selection of Items
- Selection of base period
- Selection of Average
- Selection of weights
- Selection of appropriate source of data
- Selection of suitable formula

Methods of construction of Index numbers

Unweighted Method

- Simple Aggregate method
- Simple average price relatives

Weighted Method

- Weighted aggregate method
- Weighted Average of price relatives

Weighted Aggregate Method

- Laspeyre's Method
- Paasche's Method
- Bowley–Doorfish Method
- Fisher's Ideal Method
- Marsall–Edgeworth's Method
- Kelly's Method

Weighted Average of Price Relatives

- Arithmetic Mean Method
- Geometric Mean Method

Tests of Consistency of Index Numbers

- Time Reversal Test : $P_{01} \times P_{10} = 1$
- Factor Reversal Test : $V = P_{01} \times q_{01}$
- Circular Test
- Unit Test

Base-Shifting

Means calculation of Index Number based on new base.

Splicing

Means combination of two Index Numbers.

Deflating

Means correction of Index Number according to changes in price level.

FORMULAE

A. Unweighted Aggregative Index

(a) Simple Aggregative index

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

(b) (i) Simple Average of price relative index

$$P_{01} = \Sigma \left[\frac{P_1}{P_0} \times 100 \right] / N$$

$$P_{01} = \frac{\sum P}{N}$$

(ii) When Geometric mean is used

$$P_{01} = \text{Antilog} \frac{\sum \log P}{N}$$

B. Weighted Index Numbers

(a) Weighted Aggregative Index

$$(i) \text{ Laspeyre's Index (or)} P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_1} \times 100$$

$$(ii) \text{ Paasche's Index (or)} P_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_1} \times 100$$

(iii) Bowley's Index

$$P_{01} = \frac{\frac{\sum P_1 q_0}{\sum P_0 q_0} + \frac{\sum P_1 q_1}{\sum P_0 q_1}}{2} \text{ (or)} = \frac{L + P}{2}$$

(iv) Fisher's Ideal Index

$$\begin{aligned} P_{01} &= \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0} \times \frac{\sum P_1 q_1}{\sum P_0 q_1}} \times 100 \quad (\text{or}) \\ &= \sqrt{L \times P} \times 100 \end{aligned}$$

(v) Marshall-Edgeworth's Index

$$P_{01} = \frac{\sum P_1 (q_0 + q_1)}{\sum P_0 (q_0 + q_1)} \times 100$$

(vi) Kelly's Index

$$P_{01} = \frac{\sum P_1 q}{\sum P_0 q} \times 100 \text{ where } Q = \frac{q_0 + q_1}{2}$$

C. Weighted Average of Price Relative Method

$$P_{01} = \frac{\sum PV}{\sum V}$$

D. Consumer Price Index (CPI)

- (i) Aggregate Expenditure Method

$$\text{CPI} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

- (ii) Family Budget Method

$$\text{CPI} = \frac{\sum PW}{\sum W}$$

- (iii) Consumer Price Index

$$= \frac{\sum IW}{\sum W}$$

E. Time Reversal Test

Time Reversal test is satisfied when $P_{01} \times P_{10} = 1$

$$\text{or } = \sqrt{\frac{\sum P_1 q_0}{\sum P_0 q_0}} \times \sqrt{\frac{\sum P_1 q_1}{\sum P_0 q_1}} \times \sqrt{\frac{\sum P_0 q_1}{\sum P_1 q_1}} \times \sqrt{\frac{\sum P_0 q_0}{\sum P_1 q_0}} = 1$$

F. Factor Reversal Test

Factor Reversal test is satisfied when

$$= P_{01} \times Q_{01} = \frac{\sum P_1 q_1}{\sum P_0 q_0}$$

$$= Q_{01} = \sqrt{\frac{\sum P_1 q_1}{\sum P_0 q_0}} \times \sqrt{\frac{\sum q_1 P_1}{\sum q_0 P_1}}$$

EXERCISES

(a) Choose the best option:

1. Index number is a measure of studying the relationship between
 - (a) Two variable
 - (b) Three variable
 - (c) Three or more variable
2. An Index number which is used for measuring the changes of prices of commodities over a period of time is
 - (a) Price Index number
 - (b) Value Index number
 - (c) Quantity Index number

- 3.** Quantity Index numbers deal with
- Quantity of goods
 - Quality of goods
 - Price of goods
- 4.** Under the fixed base the prices of the base year should be treated as
- 100
 - 110
 - 120
- 5.** Index number based on arithmetic mean is
- $$(a) P_{01} = \frac{\sum \left[\frac{P_1 \times 100}{P_0} \right]}{N}$$
- $$(b) P_{01} = \frac{p_1 \times 100}{\bar{P}_0}$$
- $$(c) P_{01} = \frac{p_1 \times 100}{P_0}$$
- 6.** Under Laspeyre's method price Index is
- $$(a) P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$
- $$(b) P_{01} = \frac{\sum p_1 q_1}{\sum p_1 q_0} \times 100$$
- $$(c) P_{01} = \frac{\sum p_0 q_1}{\sum p_1 q_0} \times 100$$
- 7.** Under Paasche's method price Index is
- $$(a) P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$
- $$(b) P_{01} = \frac{\sum p_0 q_1}{\sum p_1 q_1} \times 100$$
- $$(c) P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_1} \times 100$$
- 8.** Under Bowley–Doorfish Method the price Index is
- $$(a) P_{01} = \frac{L + P}{2}$$
- $$(b) P_{01} = \frac{L - P}{2}$$
- $$(c) P_{01} = \sqrt{\frac{L + P}{2}}$$

9. The circular test is the extension of
 - (a) Factor reversal test
 - (b) Time reversal test
 - (c) Unit test
10. Historically the first index was constructed in
 - (a) 1864
 - (b) 1774
 - (c) 1764
11. The ratio between the value of all commodities in the given period to the value of all commodities in the base period.
 - (a) Quantity Index number
 - (b) Value Index number
 - (c) Price Index number
12. To find out the consistency of Index numbers, time reversal test is
 - (a) $P_{01} \times P_{10} = 1$
 - (b) $P_{01} \times P_{10} = 0$
 - (c) $P_{01} \times P_{10} = 2$
13. The combination of two Index number is
 - (a) splicing
 - (b) Base shifting
 - (c) Shifting
14. The correction of Index number according to changes in price level is
 - (a) splicing
 - (b) Deflating
 - (c) shifting
15. To find out the real wages, wages earned by workers should be compound with
 - (a) Value Index
 - (b) Price-Index
 - (c) Quantity Index

Answers

- | | | | |
|---------|---------|---------|---------|
| 1. (a) | 2. (a) | 3. (a) | 4. (a) |
| 5. (a) | 6. (a) | 7. (a) | 8. (a) |
| 9. (b) | 10. (c) | 11. (b) | 12. (a) |
| 13. (a) | 14. (b) | 15. (b) | |

(b) Theoretical Questions

1. What is an Index number?
2. What are index numbers? Point out their uses.
3. Define Index numbers. Give its advantages.
4. Distinguish clearly herein find base and claim base index numbers and point out their relations merits and demerits.

(B.Com., CHU, BDU, MKU)

5. State and Explain Fisher's ideal formula for price index numbers. Show how it satisfies the time reversal test and the factor reversal tests.

(B.Com., MSU, BDU, MKU)

6. Explain the terms personal index, Budget index, scheduled index and shortage index.
7. What is meant by Family Budget enquiry? Discuss briefly the method of construction of cost of living index numbers.
(B.Com., MKU, MSU, BU)
8. What is fisher's ideal Index?
9. What is an Index number? Give Laspeyre's, Paasche's and Fisher's Index numbers. Which one is the best and why?
10. Explain the steps in the construction a cost of living Index by family Budget method.
(B.Com., MKU, MSU, BU)
11. Index numbers are economic line meters? Explain this statement and mention the utilities of Index numbers?
12. What do you understand by the terms have shifting, splicing and deflating? Explain these concepts in detail with the help of example.
13. What are the manifestations and drawbacks of latest economic indicators? Explain the latest economic indicators of wholesale prices in India?
14. What are the tests of a good index number? Define fisher's ideal index number and show that it satisfies all these tests.
(B.Com., CHU, BDU, MKU)
15. Explain the method of selection of weights for construction of cost of living index numbers.
16. What is need for deflating index numbers?
17. Define the analytical relationship between laspeyre's and Paasche's price index numbers. Under what circumstances are the two index numbers equal?
18. Describe the procedure followed in the construction of Index numbers.
19. Define Index numbers. State the utilities of index numbers.
20. Distinguish between price index numbers, quantity Index numbers and value index numbers.
(B.Com., CHU, BDU, MKU)
21. What do you mean by index number and state the problems arise while constructing index number.
22. Explain the various methods of construction of Index numbers.
(B.Com., CHU, MKU, MSU)
23. Explain the different tests of consistency of Index numbers.
24. What do you mean by chain base Index numbers? State the procedure of construction chain based Index number.
25. State briefly how chain base Index number can be converted into fixed base Index number.
(B.Com., MKU)

- 26.** Write short notes on (a) Base shifting (b) Spliced Index number (c) Deflating of Index number. (B.Com., CHU, BDU, MKU)
- 27.** What do you mean by consumer price index number and explain the objectives of construction consumer price index number?
- 28.** Define consumer price index number and state the precautions for the construction consumer price index number.
- 29.** What are the problems in the construction of cost of living index number.
- 30.** State the important methods of constructing consumer price Index.
- 31.** What do you mean by unweighted index numbers?
- 32.** Explain the methods of constructing an index number. What tests are usually applied to examine the suitability of Index numbers?
- 33.** State the limitations in the construction of index numbers.
- 34.** What do you mean of (a) Aggregate Expenditure method (b) Family Budget method.
- 35.** What do you mean by value Index number?
- 36.** Define Fisher's ideal index number and show that it statistics all these tests.
- 37.** What is an index numbers? Why are Index numbers called economic barometers?
- 38.** "Index numbers are economic barometers". Explain the statement and precautions should be taken in making use of published index number.

(c) Practical Problems

- 39.** Construct Index number for 2006 based on the prices of 2000

Commodities	A	B	C	D	E	F	G
Price for 2000 (Rs)	10	15	24	18	6	22	11
Price for 2006 (Rs)	17	19	25	15	12	30	20

Answer 130

- 40.** The data related to the prices of 7 commodities for 2002, and 2007 are given below.

Commodities	1	2	3	4	5	6	7
Price for 2002 (Rs)	35	48	50	20	15	44	65
Price for 2007 (Rs)	45	55	62	29	26	40	60

Calculate price index on the basis of simple average of price relative method through A.M. and G.M.

Answer 124, 121.3

(B.Com., CHU, BDU, MKU)

41. Construct price Index Number by (a) Laspeyre's method (b) Paasche's method (c) Bowley–Darbish method and (d) Fisher's method from the data given below

Commodity	2005		2007	
	Price	Quantity	Price	Quantity
A	12	8	15	10
B	16	10	25	11
C	10	15	16	14
D	5	6	8	5

- Answer** 151, 149.5; 150.3; 150.2 (B.Com., MKU, MSU, BU)
42. Calculate price index from the following data by weighted average of price relatives through (a) Arithmetic mean method and (b) Geometric mean method.

Sl.No.	Commodity	Quantity	Price in 1996	Price in 1997
1.	A	12	7	8
2.	B	10	14	15
3.	C	8	20	24
4.	D	15	50	48
5.	E	5	12	20

- Answer** 105.3; 104.2
43. Calculate quantity index number through Fisher's method and test the consistency of it by (a) time reversal test and (b) factor reversal test.

Sl. No.	Commodity	1995		1997	
		Price	Quantity	Price	Quantity
1.	Beef	30	20	40	25
2.	Mutton	60	25	120	20
3.	Chicken	25	18	40	20
4.	Pork	15	10	25	8
5.	Fish	30	30	50	32

- Answer** 172.07
44. Convert the following chain base index numbers into fixed base index numbers.

Year	1990	1991	1992	1993	1994	1995	1996
CBI	85	110	140	130	125	175	200

- Answer** 85; 110; 140; 130; 125; 175; 200
(B.Com., BU, BDU, MKU)

45. Find out the new index numbers based on 1994 from the following index number based on 1988.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
Index number	100	140	155	150	165	177	184	195	210	240

Answer 54.3; 76.1; 84; 90; 96; 100; 106; 114; 130

46. Calculate index number through Aggregate Expenditure method from the following data.

Commodities	Quantity consumed in 1986	Price per unit in 1986 Rs	Price per unit in 1986 Rs
A	20	35	60
B	50	25	50
C	30	12	20
D	20	8	10
E	25	10	16
F	15	14	20
G	10	8	15

Answer 178 (B.Com., BU, BDU, CHU)

47. Calculate index number on the basis of family budget method, from the following data.

Commodities	Weight	Price per unit 1987	Price per unit 1997
A	25	12	25
B	15	8	30
C	10	15	40
D	30	20	50
E	40	5	15
F	50	16	35

Answer 245 (B.Com., CHU, BDU, MKU)

48. On the basis of the following information, calculate the Fisher's ideal index number.

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	2	40	6	50
B	4	50	8	40
C	6	20	9	30
D	8	10	6	20
E	10	10	5	20

Answer 149.2

(B.Com., CHU, BDU, BU)

49. Construct the cost of living index number from the following

Group	Index	Weight
A	350	5
B	200	2
C	240	3
D	150	1
E	250	2

Answer 270.8

50. Compute (i) Paasche's (ii) Laspeyre's index numbers for quantity for the following data for the year 1989 taking 1986 as the base year

Commodity	1986		1989	
	Price	Quantity	Price	Quantity
X	7.5	25	8.2	22
Y	16.0	35	18.0	40
Z	12.3	29	14.4	25

Answer (i) 113.3 (ii) 113.5 (B.Com., CHU, BDU, MKU)

51. Construct an index number of prices from the following data:

Commodities	Base year price Rs	Current year price Rs	Weights
A	12	8	38
B	16	15	22
C	27	30	29
D	59	64	9
E	21	29	2

12

CHAPTER

TIME SERIES ANALYSIS

12.1 INTRODUCTION

A Time Series is a series of statistical data recorded in accordance with their time of occurrence. It is a set of observations taken at specified times usually (but not always) at equal intervals.

In time series analysis, current data in a series may be compared with past data in the same series. In economics, statistics and commerce play important roles. A set of data depending on the time (which may be year, quarter, month, week, days etc.) is called a Time Series. Thus, it is a set of quantitative readings of some variables recorded at equal intervals of time.

The analysis of time series is done mainly for the purpose of forecasts and for evaluating the past performances. The techniques of index numbers and time series in statistics deal with changes over time and they are dynamic in nature.

Analysis of time series consists of discovering, measuring and isolating any regular or persistent movements present in the series.

12.2 DEFINITIONS

According to **Croxton and Cowden**, *A time series consists of data arranged chronologically.*

According to **Patterson**, *A Time-series consists of statistical data which are collected, recorded or observed over successive increments.*

12.3 USES OF TIME SERIES ANALYSIS

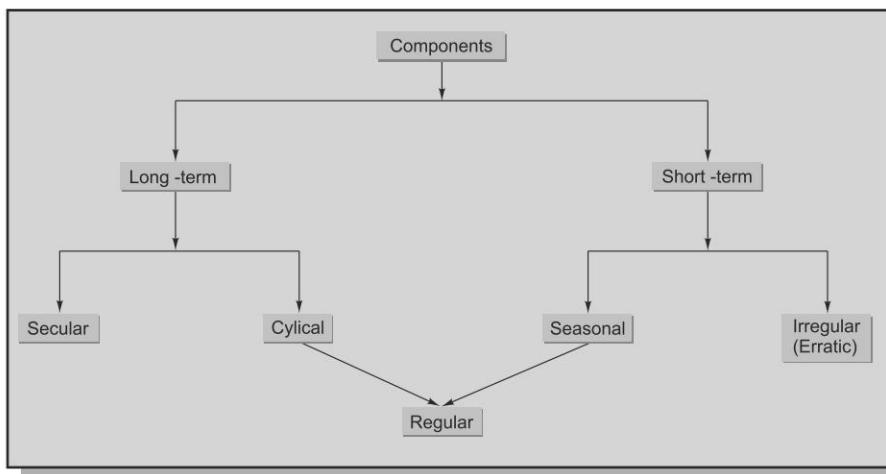
The analysis of time series is important not only to businessmen and economists, but also to scientists, social scientists biologists, management consultants etc. because of the following reasons:

- It helps in the analysis of past behaviour of a variable.
- It helps in forecasting and planning.
- It helps in evaluation of current achievement.
- It helps in making comparative studies.
- It helps in decision-making.

12.4 COMPONENTS OF TIME SERIES

A Time Series is the result of a number of movements which may be caused by numerous economic, political, natural and other factors. There are four basic types of variations and these are called the components or elements of time series. They are:

1. Secular Trend or long-term movement
2. Seasonal variation
3. Cyclical fluctuations
4. Irregular, erratic or random fluctuations.



12.3.1 SECULAR TREND

The word Secular is derived from the Latin word ‘SAECULUM’ which means generation or age. The general tendency of the time series data to increase or decrease or stagnate during a long period of time is called the secular trend, also known as long-term trend.

The concept of trend does not include short-range oscillations, but rather the steady movements over a long time.

This phenomenon is usually observed in most of the series relating to economics and business; for instance, an upward tendency is usually observed in time series relating to population, production prices, income, money in

506 Business Statistics

circulation etc., while a downward tendency is noticed in time series relating to deaths, epidemics etc., due to advancement in medical technology, improved medical facilities, better sanitation etc. Thus, a time series may show fluctuations in the upward or downward directions but there is a distinct tendency for it either to increase or decrease in the long run.

The Secular Trend has the growth factor or the declining factor. It may have either upward or downward movement. There are many types of trend—some trend rise upward; some trend rise upward and some trend fall downward and it is not a rule that the rise or fall of the trend must continue in the same direction throughout the period. The different types of trend are given below.

- A. Linear or straight line Trend.
- B. Non-linear or curvilinear Trend.

12.4.2 Seasonal Variation

In the words of Hirsch, A recurrent pattern of change within the period that results from the operation of forces connected with climate (or customs) at different times of the period.

The seasonal variations are usually measured in an interval within the calendar year. These may be daily, hourly, weekly, monthly or quarterly. The seasonal variation may occur due to—

- a. **Climate and Natural Forces** For instance, the sales of umbrella pick up very fast in rainy season, the sale of ice and ice-cream increases very much in summer; the sale of woollens go up in winter—all being affected by natural, viz., weather or season.
- b. **Customs and Habits** Those variations in a time series within a period of 12 months are due to habits, fashions, customs and conventions of the people in the society. For instance, the people in the society; the sales of jewellery and ornaments go up in marriages; the sales and profits in department stores go up considerably during marriages and festivals like Diwali, Christmas etc.

Seasonal variation is useful to businessmen, agriculturists, sales managers and producers. A study of seasonal variation of a time series facilitates the following:

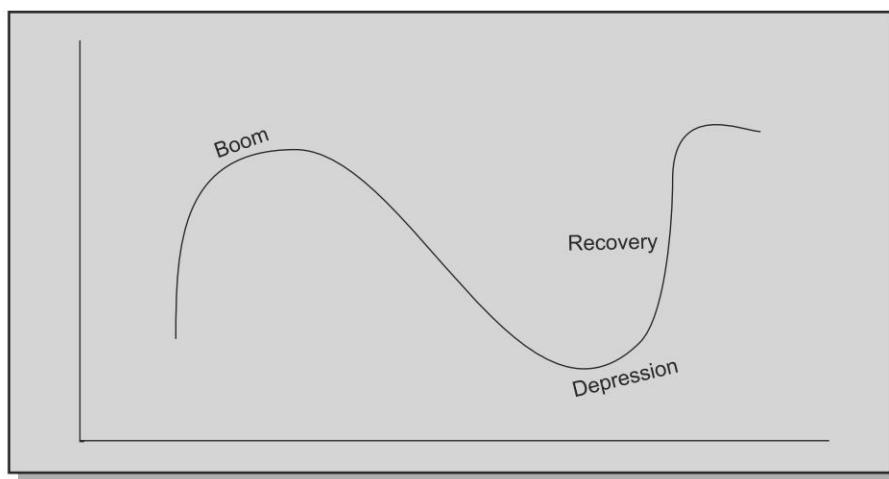
- Planning future operations
- Formulating correct policy decisions through its analysis
- Studying the effects of seasonal variation and to isolate them from the trend.

12.4.3 Cyclical Variation

Cyclical movement is another type of periodical movement, with a period longer than a year. Such movements are fairly regular and oscillatory in nature. Most of the economic and business time series are influenced by the wave-like changes of

prosperity and depression. There is a periodic up and down movement. This movement is known as cyclical variation or oscillation moments which occurs in cycle. One compiled period is called a ‘cycle’.

The movement occurs in every 3 or 10 years. The movement will be like a wave. There are four phases of business cycle. They are prosperity (Boom), recession, depression and recovery.



A study of cyclical variation of a time series facilitates the following:

- To study the character of business fluctuations easily.
- To formulate good policies in order to stabilise the level of business activity.
- To take timely decision in maintaining business during booms and depressions.
- To facilitate a businessman to face the recession period and make them ready to reap the benefits during booms.

12.4.4 IRREGULAR VARIATIONS OR RANDOM VARIATIONS

These variations are completely unpredictable in character. These variations are beyond the control of human hand which are caused by factors either wholly unaccountable, erratic, unforeseen, unpredictable or they are due to numerous non-recurring and irregular circumstances.

These isolated or irregular fluctuations are due to floods, revolution, political upheavals, famines etc., are also called episodic fluctuations. There is no statistical technique for measuring or isolating erratic influences.

Times Series Decomposition Models Before analysing the time-series data, first decide the nature of inter-relationship amongst the different components of time series. Different assumptions are made by different statisticians and this led to two theories, ‘additive’ and ‘multiplicative’.

508 *Business Statistics*

(A) Additive Model In this model, the value of the original data in the series is taken as the sum of four components of the time series and it can be expressed as

$$Y = T + S + C + I$$

where

Y = The original data

T = Trend

S = Seasonal variations

C = Cyclical fluctuations

I = Irregular variations.

Assumptions

- Four components of the time series are independent to each other
- Different components are expressed in original units and are residuals
- Seasonal variations, cyclical and irregular fluctuations are expressed as deviation from the trend.

(B) Multiplicative Model In this model, the product of the values of the time series components are taken as the value of original data and can be expressed as

$$Y = T \times S \times C \times I$$

Here, S , C and I are expressed in ratios and percentages. The multiplicative model is mostly used in practice.

Assumptions

- Only trend is expressed in terms of original values.
- Four components are not necessarily independent and may be dependent on each other so that seasonal variations increase with using trend and the ratio of cyclical fluctuations also remains constant.

12.5 PRELIMINARY ADJUSTMENTS BEFORE ANALYSING TIME SERIES

For analysing Time Series, raw data must be adjusted and the unwanted elements must be removed. They are as follows.

12.5.1 Adjustment of Calendar Variations

While analysing time series data, it must be ensured that the data of one period should be compared with another period—say a week, a month, a quarter or a year. The figures of monthly sales may have to be adjusted on the basis of number of days in different months.

12.5.2 Adjustment for Population Changes

The changes in the size of population lead to have comparisons of many things. For instance, national income may be increasing year after year. Yet per capita income may show a declining trend. So to get the per capita income month by month or year by year, the total figure should be divided for the entire time by the population in each of these time intervals. This adjustment removes the influence of increasing population.

12.5.3 Adjustment of Price Changes

In order to deflate the figures, the effect of price change in raw data must be removed. For this, each weekly figure should be divided by the cost of living index for that week. It must be followed in the entire series.

12.5.4 Adjustment for Comparison Purposes

Two or more series of data are to be compared over a period of time and it is necessary to make such series comparable in a statistical time series. These series can be compared directly with each other if they are converted into percentage of a given past period.

12.5.5 Adjustment for Miscellaneous Changes

The conditions might not remain uniform over a given period of time. Such as quality or type of produces, definition of the unit or item during the given period of time series. Necessary adjustments should be made in order to make a systematic study of time series.

12.6 MEASUREMENT OF TREND

The time series analysis is absolutely essential for planning. It guides the planners to achieve better results. The study of trend enables the planner to project the plan in a better direction. The reason for the measurement of trend is to find out trend characteristics in the series. We can eliminate trend in order to study the other elements by the measurement of trend. It will help us to know the seasonal, cyclical and irregular variations.

The following are the four methods which can be used for determining the trend.

1. Free-hand or Graphic Method
2. Semi-average Method
3. Moving Average Method
4. Method of Least Squares.

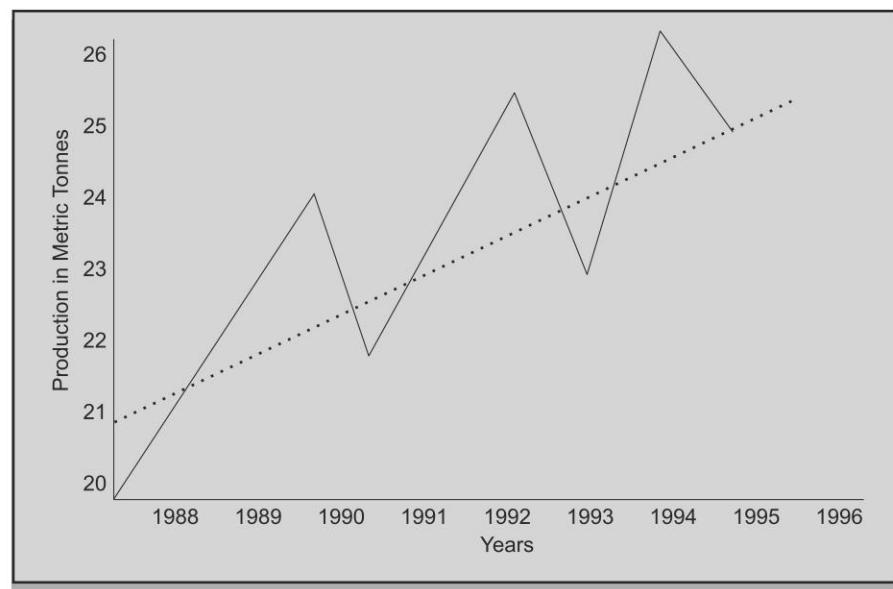
12.6.1 Free-hand or Graphic Method

This is the easiest and simplest method. In this method, we plot the original data on the graph. Draw a smooth curve which will show the direction of the trend. The time and value are shown in horizontal axis and vertical axis respectively.

Illustration 12.1

Fit a trend line to the following data by the free hand method:

Year	Production of Steel (in million tonnes)
1988	20
1989	22
1990	24
1991	21
1992	23
1993	25
1994	23
1995	26
1996	25



The trend line drawn by the free hand method can be extended to project future values. However, the free hand curve fitting is too subjective.

Merits

1. It is the easiest method of measuring trend.
2. It is more flexible than any other methods of measurement.
3. It also gives the picture of the movement of data, in addition to the straight line.

Demerits

1. This method is highly subjective in nature. Different trend lines can be drawn from the curve.
2. It won't accurately project the future.
3. The accuracy depends upon the skill of the analyst.

12.6.2 Method of Semi-Averages

Under this method, the original data is divided into two equal parts and averages are calculated for the both the parts. For example, the value for 1990–1999 can be divided into two parts, 1990–1994 and 1995–1999 and calculate averages for each part.

We can draw the line by the straight line, joining the two points of average. The line may be extended upward or downward. We can get intermediate values or partial values.

Illustration 12.2

(Odd number of years)

Fit a trend line to the following data by the method of semi-averages.

Year	Sales of firm A (thousand Units)
1990	102
1991	105
1992	114
1993	110
1994	108
1995	116
1996	112

Solutions

Since seven years are given, the middle year shall be left out and an average of the first three years and the last three years shall be obtained. The average of the

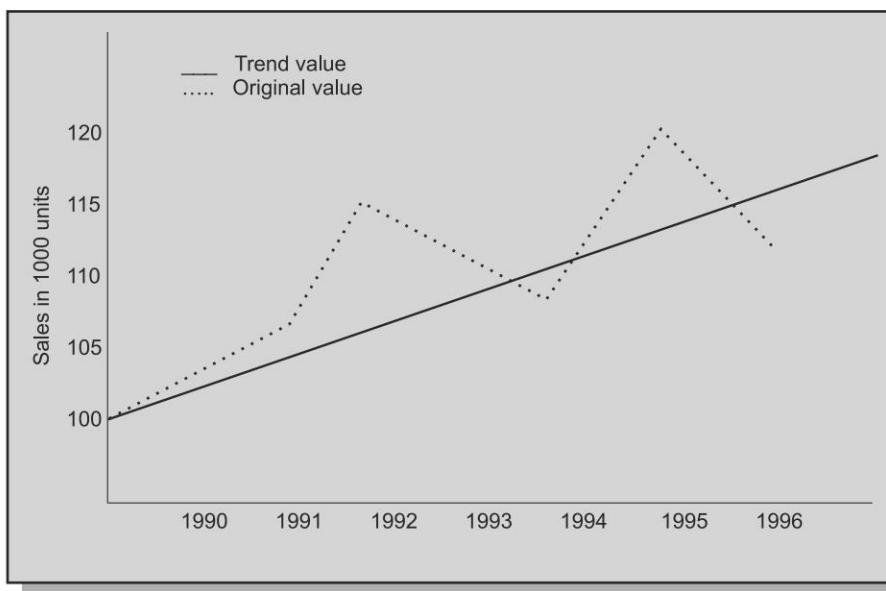
first three years is $\frac{102 + 105 + 114}{3} = \frac{321}{3} = 107$ and the average of last three

years is $\frac{108 + 116 + 112}{3} = \frac{336}{3} = 112$. Thus, we get two points 107 and 112 which shall be plotted corresponding to their respective middle years, that is, 1991 and 1995. By joining these two points, we shall obtain the required trend line. The

512 *Business Statistics*

line can be extended and can be used either for prediction or for determining intermediate values.

The actual data and the trend line are shown in the following graph.

**Even Number of Years**

When there are even number of years like 6, 8, 10, etc., two equal parts can easily be formed and an average of each part is obtained. However, when the average is to be centered, there would be some problem in case the number of years 8, 12, etc. For example, if the data relate to 1992, 1993, 1994 and 1995, which would be the middle year? In such a case, the average will be centered corresponding to 1st July 1993, that is, the middle of 1993 and 1994.

Illustration 12.3

Fit a trend line by the method of semi-averages to the data given below. Estimate the sales for 1997, if the actual sale for that year is Rs 260 lakhs account for the difference between the two figures.

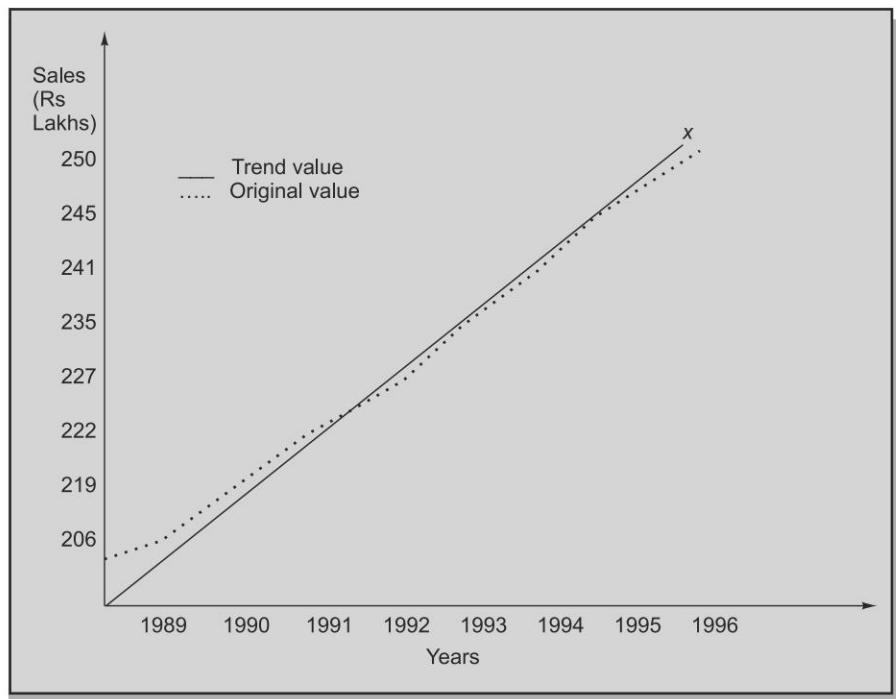
Year	Sales (Rs Lakhs)
1989	206
1990	219
1991	222
1992	227
	$\frac{874}{4} = 218.5$

Year	Sales (Rs Lakhs)
1993	235
1994	241
1995	245
1996	250
	$\frac{971}{4} = 242.75$

Solutions

The average of the first four years is 218.5 and that of the last four years is 242.75. These two points shall be taken corresponding to the middle periods, that is, 1st July 1990 and 1st July 1994.

Trend by the Method of Semi-Averages



The estimated sales for 1997 by projecting the semi-average trend line is the actual figure given to us Rs 260 lakhs. The difference is due to the fact that time series analysis helps us to get the best possible estimates on certain assumptions which may come out to be true (or) not depending upon how far these assumptions have been realised in practice.

Merits

1. It is simple and easy to understand.
2. With the help of trend line, we can get the intermediate values and predict the future values.

Demerits

1. This method assumes straight line relationship whether such relationship exists or not.
2. It is affected by the limitations of arithmetic mean.

12.6.3 Method of Moving Average

It is a logical extension of semi-average method. Under this method, a chain of successive averages should be calculated. Averages can be taken from successive periods. The average value for a number of years, months or weeks is taken into account. Moving averages can be calculated for 3, 4, 5, 6, 7, 8, 9 or 10 yearly period. The moving average trend will be a straight line if there is a regular movement of variables. It will be a smooth curve if there is irregular movement of variables.

Moving averages can be calculated as follows:

(a) Moving Averages for Odd Number of Data

The formula for 3-yearly moving average is

$$\frac{a+b+c}{3}, \quad \frac{b+c+d}{3}, \quad \frac{c+d+e}{3} \dots\dots$$

The formula for 5-yearly moving average is

$$\frac{a+b+c+d+e}{5}, \quad \frac{b+c+d+e+f}{5}, \quad \frac{c+d+e+f+g}{5} \dots\dots$$

(a, b, c, d.... represent variables)

Steps Steps for calculating odd number of years (3, 5, 7, 9). If we want to calculate the three-yearly moving average, then:

1. Compute the value of first three years (1, 2, 3) and place the three years total against the middle year (i.e., 2nd year)
2. Leave the first year's value and add up the values of the next three years, i.e., 2, 3, 4 and place the three-year total against the middle year, i.e., 3rd year.
3. This process must be continued until the last year's value and is taken for calculating moving average.
4. The three-yearly total must be divided by 3 and placed in the next column.
This is the trend value of moving average.

If the trend value is plotted on a graph, it will give us the trend.

Illustration 12.4

Calculate three-yearly moving average of the following data.

Year	No. of Students
1991	15
1992	18
1993	17
1994	20
1995	23
1996	25
1997	29
1998	33
1999	36
2000	40

Solutions

Year	No. of Students	3- yearly Total	Moving Average
1991	15		
1992	18	50	16.7
1993	17	55	18.3
1994	20	60	20.0
1995	23	68	22.7
1996	25	77	25.7
1997	29	87	29.0
1998	33	98	32.7
1999	36	109	36.3
2000	40		

Illustration 12.5

Calculate the 5-yearly moving average from the following data:

Year	No. of Students
1991	705
1992	685

516 Business Statistics

Year	No. of Students
1993	703
1994	687
1995	705
1996	689
1997	715
1998	685
1999	725
2000	730

Solutions

Year	No. of Students	5-yearly Moving Total	Moving Average
1991	705		
1992	685		
1993	703	3485	697.0
1994	687	3469	693.8
1995	705	3499	699.8
1996	689	3481	696.2
1997	715	3519	703.8
1998	685	3544	708.8
1999	725		
2000	730		

Even period of moving average If the period of moving average is 4, 6 or 8, it is an even number. For 4-yearly moving average, the 4-yearly total cannot be placed against any year as the median 2.5 is between the 2nd and the 3rd years. So, the total should be placed in between the 2nd and 3rd years. We must centre the moving average in order to place the moving average against an year.

Steps

1. Compute the values of the first four years and place the total in between the 2nd and the 3rd years.
2. Leave the first year value and compute the value of the next four years and place the total in between the 3rd and 4th year.
3. This process must be continued until the last year is taken into account.

4. Compute the next two four-year totals and place it against the middle year (i.e., 3rd year).
5. Leave the first four-year total and compute the next four-year total and place in the 4th year.
6. This method must be continued until all the four-yearly totals are computed.
7. Divide the above totals by 8 (because it is the total of the two four-yearly totals) and put in the next column. This is the trend value.

Illustration 12.6

Assuming a four-yearly cycle, calculate the trend by the method of moving averages from the following data relating to the production of tea in India.

Year	Production (in million lbs)
1991	464
1992	515
1993	518
1994	467
1995	502
1996	540
1997	557
1998	571
1999	586
2000	612

Solutions

Calculation of moving average

Year	Production	4-yearly Moving Total	Add in pairs (Combined)	4-yearly Moving Average
1991	464			
1992	515			
		1964		
1993	518		3966	495.8
		2002		
1994	467		4029	503.6
		2027		
1995	502		4093	511.6

Year	Production	4-yearly Moving Total	Add in pairs (Combined)	4-yearly Moving Average
		2066		
1996	540		4236	529.5
		2170		
1997	557		4424	553.0
		2254		
1998	571		4580	572.5
		2326		
1999	586			
2000	612			

Merits

1. It is simple and easy to understand.
2. It is more flexible than other methods.
3. It is not only used for the measurement of trend, but also for the measurement of seasonal, cyclical and irregular fluctuations.
4. It is away from personal bias because the period of moving average is determined by the data and not by the personal judgement of the investigator.

Demerits

1. In this method, we cannot get the trend values for all the given observations.
2. The main object of trend value is that it is used for forecasting or predicting future values because this method is not represented by a mathematical function.
3. There is no rule regarding the choice of the number of the moving average, and so the statistician has to use his own judgement.

12.6.4 Method of Least Squares

By the method of least square, a straight line trend can be fitted, to the given time series of data. It is a mathematical and an analytical method. With its help, economic and business time series data can be fitted and this helps in forecasting and predicting. The trend line is called the line of best fit. The sum of deviations of the actual values of Y and the trend value (Y_c) is 0 and the sum of squares of deviations of the actual value and the trend value is the least.

$$(Y - Y_c) = 0 \text{ and}$$

$$(Y - Y_c)^2 = \text{least.}$$

So, this method is called the least square method or the line of best fit.

The method of least square can be used to explain the linear and non-linear trend, that is, a straight line trend or parabolic trend.

The straight line trend or the first degree parabola is represented by the mathematical equation,

$$Y_c = a + bx$$

Y_c = required trend value

X = unit of time

a and b are constants or unknowns.

The equation for second degree parabola is

$$Y_c = a + bx + cx^2$$

For third degree parabola, the equation is

$$Y_c = a + bx + cx^2 + dx^3 \dots\dots$$

The straight line method or the first degree parabola will give a straight line and the other equations will give curved line or non-linear trend which will depend upon the degree of parabola.

In the equation for the first degree parabola $Y_c = a + bx$, the values of the unknowns or constants can be calculated by the following two normal equations:

$$\sum Y = Na + b \sum x$$

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

N = the number of years or months for which data are given.

When $X = 0$ (when middle year is taken as the origin), the equation will take the form of:

$$\sum Y = Na + b \sum x \quad @ b \sum x = 0$$

$$\sum XY = a \sum x + b \sum x^2 @ b \sum x^2 = 0$$

By these equations, we can know the values of a and b , i.e.,

$$a = \frac{\sum Y}{N} \text{ and } b = \frac{\sum XY}{\sum x^2}$$

a = the mean value of Y values

b = rate of change.

We can use this method when we are given odd number of years. It is easy and is widely used in practice. If the number of items is odd, we can follow the steps given below:

1. Denote time as the X variable and values of Y .
2. Middle year is assumed as the period of origin and find out deviations.
3. Square the time deviation and find X^2 .
4. Multiply the given value of Y by the respective deviation of X and find the total $\sum XY$.

520 Business Statistics

5. Find out the values of Y ; get ΣY .
6. The values so obtained are placed in the two equations
(1) $\sum Y = Na + b \sum x$ (2) $\sum XY = a \sum x + b \sum x^2$; find out the values of a and b .
7. The calculated values of a and b are substituted and the trend value of Y_c are found for various values of X .

Illustration 12.7

Calculate trend values by the method of least square from the data given below and estimate the sales for 2009.

Year	Sales of Co. A (Rs Lakhs)
2002	140
2003	144
2004	160
2005	172
2006	180

Solutions

Calculation of trend values by the method of Least Square

Year	Sales Y	X 2004: 0X	X^2	XY	Y_c
2002	140	-2	4	-280	137.6
2003	144	-1	1	-144	148.4
2004	160	0	0	0	159.2
2005	172	1	1	172	170
2006	180	2	4	360	180.8
$N = 5$	$\sum y = 796$	$\sum x = 0$	$\sum x^2 = 10$	$\sum xy = 108$	796

Since

$$a = \frac{\sum y}{N}$$

$$= \frac{796}{5} = 159.2$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{108}{10} = 10.8$$

Hence,

$$\begin{aligned} Y_c &= a + bx = 159.2 + 10.8(-2) \\ &= 159.2 - 21.6 = 137.6 \end{aligned}$$

$$Y_{2002} = 137.6$$

∴ The other values can be obtained by adding the value of b to the preceding value for 2009, x will be 5, putting $x = 5$ in the equation.

$$\begin{aligned} Y_{2009} &= 159.2 + 10.8(5) \\ &= 159.2 + 54 \end{aligned}$$

$$Y_{2009} = 213.2$$

Fitting a Straight Line Trend for Even Number of Years The method of least square for even numbers in a time series is almost the same as that for odd number of years for computing the trend value. There will be no middle year against which the value of $X = 0$ can be assigned. But origin is taken as the mid-point of two middle years of the series. We must give negative signs to the one proceeding the origin and must give positive signs to the one succeeding the origin year. The sum of X values will be 0, as in the odd number of years.

In order to avoid the decimal points in the deviations from the value of X , we can use another method, i.e.,

$$X = \frac{(t - \text{Mid point of } X)}{1/2}$$

or

$$X = 2(t - \text{mid points of } X)$$

where

t = time or year

Illustration 12.8

Calculate trend value from the following data using the Method of Least Square.

Year	Production
1994	7
1995	9
1996	12
1997	15
1998	18
1999	23

Solutions

Computation of Trend by Least Square method

Year	Production Y	Time deviation from 1996.5 x	x^2	xy	Trend Value (computed)
1994	7	-2.5	6.25	-17.5	6.15
1995	9	-1.5	2.25	-13.5	9.29
1996	12	-0.5	0.25	-6.0	12.43
1997	15	0.5	0.25	7.5	15.57
1998	18	1.5	2.25	27.0	18.71
1999	23	2.5	6.25	57.5	21.85
N = 6	$\sum y = 84$		$\sum x^2 = 17.50$	$\sum xy = 55$	

Since

$$x = 0,$$

$$a = \frac{\sum xy}{N} = \frac{84}{6} = 14$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{55}{17.50} = 3.14$$

Hence,

$$Y_c = a + bx$$

$$Y_{1994} = 14 + 3.14 (-2.5)$$

$$= 14 - 7.85 = 6.15$$

$$Y_{1995} = 14 + 3.14 (-1.5) = 14 - 4.71 = 9.29$$

$$Y_{1996} = 14 + 3.14 (-0.5) = 14 - 1.57 = 12.43$$

$$Y_{1997} = 14 + 3.14 (0.5) = 14 + 1.57 = 15.57$$

$$Y_{1998} = 14 + 3.14 (1.5) = 14 + 4.71 = 18.71$$

$$Y_{1999} = 14 + 3.14 (2.5) = 14 + 7.85 = 21.85$$

Second Degree Parabola**Illustration 12.9**

Fit a Parabola of the second degree to the data given below.

Year	Sales ('000)
1995	16
1996	18
1997	19
1998	20
1999	24

Solutions

Year	γ_{1997}	Year x	x^2	x^3	x^4	xy	x^2y	Trend Value
1995	16	-2	4	-8	16	-32	64	16.08
1996	18	-1	1	-1	1	-18	18	17.46
1997	19	0	0	0	0	0	0	19.12
1998	20	1	1	1	1	20	20	21.06
1999	24	2	4	8	16	48	96	23.28
$N = 5$	$\sum y = 97$	$\sum x = 0$	$\sum x^2 = 10$	$\sum x^3 = 0$	$\sum x^4 = 34$	$\sum xy = 18$	$\sum x^2y = 198$	

$$Y_c = a + bx + cx^2$$

Since

$$\sum x = 0,$$

$$a = \frac{\sum y - c \sum x^2}{N}$$

$$a = \frac{97 - c(10)}{5} = \frac{97 - 10c}{5}$$

$$5a = 97 - 10c$$

$$5a + 10c = 97 \quad (1)$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{18}{10} = 1.8$$

$$c = \frac{\sum x^2y - a \sum x^2}{\sum x^4} = \frac{198 - a(10)}{34} = c$$

$$c = \frac{198 - 10a}{34} = 34c = 198 - 10a$$

$$34c + 10a = 198 \quad (2)$$

524 Business Statistics

Multiply Eq. (1) by 2 and subtract it from Eq. (2)

$$10a + 34c = 198$$

$$10a + 20c = 194$$

$$14c = 4$$

$$c = 4/14 = 0.29$$

Substitute the value of c in Eq. (1)

$$5a + 10(0.29) = 97$$

$$5a = 97 - 2.9 = 94.1$$

$$a = 94.1/5 = 18.82$$

$$\text{Value of } a = 18.82; b = 1.8; c = 0.29$$

Substitute the value in the equation $Y_c = a + bx + cx^2$

$$\begin{aligned} Y_{1995} &= 18.82 + 1.8(-2) + 0.29(-2)^2 \\ &= 18.82 - 3.6 + 1.16 = 16.38 \end{aligned}$$

$$\begin{aligned} Y_{1996} &= 18.82 + 1.8(-1) + 0.29(-1)^2 \\ &= 18.82 - 1.8 + 0.29 = 17.31 \end{aligned}$$

$$Y_{1997} = 18.82 + 1.8(0) + 0.29(0)^2 = 18.82$$

$$Y_{1998} = 18.82 + 1.8(1) + 0.29(1)^2 = 20.91$$

$$\begin{aligned} Y_{1999} &= 18.82 + 1.8(2) + 0.29(2)^2 \\ &= 18.82 + 3.6 + 1.16 = 23.58 \end{aligned}$$

Merits

1. As it is a mathematical method of measuring trend, it eliminates the element of subjective judgement or personal bias of the investigator.
2. We can compute the trend values for all the given time periods in the series.
3. We can predict the values of the variable for the future or intermediate periods of the given series.
4. The principle of Least Square is called the line of best fit; we can get the rate of growth for all the periods.

Demerits

1. It is difficult for a non-mathematical person to calculate; and it will take more time to calculate than the other methods.

2. It ignores the cyclical, seasonal and irregular fluctuations and is based on the long term variations or trend.
3. If we add new observations, then the calculation must be done once again.

12.7 MEASUREMENT OF SEASONAL VARIATION

The economic and business activities mostly depend upon seasonal variations. The seasonal variations could not be identified when the data are expressed annually. It can be identified only when the data are expressed monthly or quarterly. For example, the sales of cool drinks for hot seasons are greater than the sales in other seasons. Hence, the sales for the whole one year as compared with the sales of other years won't show much difference. It also won't explain the seasonal variations. Hence, the value for various seasons for different years should be calculated to find out the seasonal variations.

12.7.1 Methods of Measuring Seasonal Variations

The important methods for measuring the seasonal variations are as follows:

1. Simple averages method
2. Ratio to trend method
3. Ratio to moving average method
4. Link relative method

Simple Averages Method This is the easiest method for calculating seasonal variations. This method can be applied to monthly, bi-monthly, quarterly, half-yearly etc. Under this method, the individual totals for all the same months or period of all the given years should be obtained first. For example, if 10 years data are available for 12 months, then the 12 months total for all the 10 years should be obtained. Then the average monthly total should be calculated. After that, the averages of the average of 12 months so obtained should be calculated. To ascertain the seasonal variations, index of each month or period of the individual averages connected with the month or season should be divided by the average of the averages; and then it can be multiplied by 100.

Illustration 12.10

Compute seasonal indices by the method of monthly averages to determine the monthly indices for the following data of production of a commodity for the years 1989, 1990, 1991:

Month	1989	1990	1991
	(Production in Lakhs of tonnes)		
January	12	15	16
February	11	14	15

Month	1989	1990	1991
	(Production in Lakhs of tonnes)		
March	10	13	14
April	14	16	16
May	15	16	15
June	15	15	17
July	16	17	16
August	13	12	13
September	11	13	10
October	10	12	10
November	12	13	11
December	15	14	15

Solutions

Computation of Seasonal Indices

Month	1989 (Production in Lakhs of tonnes)	1990 (Production in Lakhs of tonnes)	1991 (Production in Lakhs of tonnes)	Total	Average Monthly	Seasonal Index
(1)	(2)	(3)	(4)	(5)	(6)	(7)
January	12	15	16	43	14.33	104.886
February	11	14	15	40	13.33	97.566
March	10	13	14	37	12.33	90.247
April	14	165	16	46	15.33	112.205
May	15	16	15	46	15.33	112.205
June	15	15	17	47	15.66	114.620
July	16	17	16	49	16.33	119.524
August	13	12	13	38	12.66	92.662
September	11	13	10	34	11.33	82.928
October	10	12	10	32	10.66	78.024
November	12	13	11	36	12.00	87.832
December	15	14	15	44	14.66	107.301
		Total	492	163.95	1200	
		Average	41	13.6625	100	

Average of Averages,

$$\bar{x} = 1/12 (14.33 + 14.33 + \dots + 14.66)$$

$$= \frac{163.95}{12} = 13.6625$$

$$\text{Seasonal Index for January} = \frac{14.33}{13.6625} \times 100 = 104.884$$

$$\text{Seasonal Index for February} = \frac{13.33}{13.6625} \times 100 = 97.566 \text{ and so on.}$$

Illustration 12.11

Calculate the seasonal index from the following data using the average method.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1994	72	68	80	70
1995	76	70	82	74
1996	74	66	84	80
1997	76	74	84	78
1998	78	74	86	82

Solutions

Computation of Seasonal Indices

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1994	72	68	80	70
1995	76	70	82	74
1996	74	66	84	80
1997	76	74	84	78
1998	78	74	86	82
Total	376	352	416	384
Average	75.2	70.4	83.2	76.8
Seasonal Index	98.43	92.15	108.9	100.52

$$\text{Average of Averages} = \frac{1}{4} (75.2 + 70.4 + 83.2 + 76.8)$$

$$= \frac{305.6}{4} = 76.4$$

$$\text{Seasonal Index} = (\text{A.M.} \div \text{Grand average}) \times 100$$

$$(75.2 \div 76.4) \times 100 = 98.43; ((70.4 \div 76.4) \times 100 = 92.15)$$

$$(83.2 \div 76.4) \times 100 = 108.9; (76.8 \div 76.4) \times 100 = 100.52$$

Note

Average seasonal movements can be obtained by subtracting the grand average from the A.M. of each quarter. Slight adjustments are sometimes necessary to make the total seasonal movement zero.

Merits This method is a simple and the easiest method of measuring seasonal variations.

Demerits

1. The assumption that there is no trend component in the series is not a reasonable assumption.
2. The effects of cyclical periods on the original values may or may not be eliminated by the process of averaging.

Ratio to Trend Method This method is an improvement over the method of simple average method. It assumes that seasonal variation for a given month is constant fraction of trend. Under this method, the trend is eliminated when the ratios are calculated. This method isolates the seasonal factors as follows:

$$\frac{T \times S \times C \times I}{T} = S \times C \times U$$

where, T —Trend S —Seasonal factor
 C —Cyclical factor I —Irregular factor

The steps for calculating seasonal index under this method are as follows:

1. First of all, the trend values should be calculated through the method of least squares.
2. Then, divide the actual data for all periods (months or quarter by the corresponding trend values. These values must be multiplied with 100. These values are free from trend).
3. The next step is to free these values from irregular and cyclical movements also. The values so obtained, for the various years of the periods (months or quarters) should be averaged either by mean or by median.
4. The sum of the average values of the periods should be equal to 1200 if the periods represent months. It should be equal to 400, if the periods represent quarters. If it is not equal to 1200, then each average value should be multiplied by,

$$\frac{1200}{\text{Sum of the 12 values (when the periods represent months)}}$$

If it is not equal to 400, then each average values should be multiplied by,

$$\frac{1400}{\text{Sum of the 4 values}} \text{ (when the periods represent quarters)}$$

Illustration 11.12

Find the seasonal variation by the ratio to trend method from the data given below.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1997	86	95	96	99
1998	96	102	104	110
1999	103	108	106	107

Solutions

Computation of Trend by Method of Least Square

Year	Yearly Total	Yearly Average y	x	x^2	xy	Trend Value
1997	376	94	-1	1	-94	95
1998	9	103	0	0	0	101
1999	424	106	1	1	106	107
N = 3	$\sum y = 303$			$\sum x^2 = 2$	$\sum xy = 12$	

$$Y = a + bx$$

Since

$$x = 0$$

$$a = \frac{\sum y}{N} = \frac{303}{3} = 101$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{12}{2} = 6$$

Trend Equation :

$$Y = 101 + 6x$$

$$Y_{1997} = 101 + 6(-1) = 101 - 6 = 95$$

$$Y_{1998} = 101 + 6(0) = 101 + 0 = 101$$

$$Y_{1999} = 101 + 6(1) = 101 + 6 = 107$$

Yearly increment is = 6

Quarterly increment = $6/4 = 1.5$

Merits

1. This method is more rational when compared to other methods of calculating seasonal indices.
2. Unlike the moving average method, this method won't exclude the extreme values.

530 Business Statistics

3. This method gives trend values for all the seasons. Hence, it can be applied even for the calculation of trend values for a short period.

Demerits

1. This method requires complex mathematical calculations.
2. The cyclical variations affect the result obtained under this method.

Ratio to Moving Average Method This method is also known as the percentage of moving average method. This method is the most commonly used method for measuring seasonal variations. The important steps for measuring seasonal variations are as follows:

1. The seasonal variations should be eliminated by taking a centered-moving average with a period of 12 months, if monthly data are available.
If quarterly data are available, a centered 4-quarter moving average must be used. This moving averages will also eliminate irregular movements.
2. Calculate the ratios of original data to the corresponding moving averages. It should be expressed in percentage, i.e.,

$$\frac{\text{Actual Value}}{\text{Moving Average}} \times 100$$

3. These percentages are then arranged by months and the average for each month is calculated. The total of these averages should be equal to 1200. If it is not equal to 1200 then each average value should be multiplied by

$$\frac{1200}{\text{Sum of the values}}$$

If the periods represent quarters, the total of the 4 averages should be equal to 400. If it is not equal to 400, then each average. Values should be multiplied by

$$\frac{400}{\text{Sum of the 4 values}}$$

Illustration 12.13

Using 4-quarterly moving average in respect of the following data, find

- (a) the trend
- (b) short-term fluctuations
- (c) seasonal variations

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1991	35	86	67	124
1992	38	109	91	176
1993	47	158	104	226

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1994	61	177	134	240
1995	72	206	141	307

Solutions

Computation of 4-quarter moving averages and short term fluctuations

Year/ Quarter (1)	Y (2)	4-Quarter Moving Total (3)	Sum of Two 4-Quarter Moving Total (4)	4-Quarter Moving Average (5) = (4) ÷ 8	Short-term Fluctuations (5) = (4) ÷ 8
1991					
1st	35				
2nd	86				
		312			
3rd	67		627	78.37	- 11.37
		315			
4th	124		653	81.62	42.38
1992					
	338				
1st	38		700	87.50	- 49.50
		362			
2nd	109		776	97.00	12.00
		414			
3rd	91		837	104.62	- 13.62
		423			
4th	176		895	111.87	64.13
1993					
	472				
1st	47		957	119.62	- 72.62
		485			
2nd	158		1020	127.50	30.50
		535			
3rd	104		1084	135.50	- 31.50
		549			
4th	226		1117	139.62	86.38
1994					
	568				
1st	61		1166	145.75	- 84.75

Contd.

Year/ Quarter (1)	Y (2)	4-Quarter Moving Total (3)	Sum of Two 4-Quarter (4)	4-Quarter Moving Average (5) = (4) ÷ 8	Short-term Fluctuations (6) = (2) – (5)
		598			
2nd	177		1210	151.25	25.75
		612			
3rd	134		1235	154.37	- 20.37
		623			
4th	240		1275	159.37	80.63
1995		652			
1st	72		1311	163.87	- 91.87
		659			
2nd	206		1385	173.12	32.88
		726			
3rd	141				
4th	307				

Merits

1. This method is the most widely used method for calculating seasonal variations.
2. The seasonal index calculated under this method does not fluctuate like other methods.
3. The cyclical variations won't affect much of the calculation of the seasonal index under this method.

Demerits

1. In this method, the seasonal indices cannot be calculated for all the months even though the data are available.
2. The values of the months at the two extremes are left out in this method.

4. Link Relatives Method This method is the most difficult one among the methods of measuring seasonal variation. The following are the important steps for calculating seasonal variations under this method.

1. Calculate the link relatives for the given values for different seasons.

It can be calculated as, $\frac{\text{Current Season's figure}}{\text{Previous Season's figure}} \times 100$

In these link relatives, link each month to proceeding month, or link each quarter to preceding quarter.

2. Arrange the link relatives by months or quarters. Average value of the link relatives for the season (months or quarters) should be calculated, on the basis of mean or median.
3. Convert these averages into chain relatives on the basis of the first season (month or quarter). The chain relative of the 1st quarter will be,

$$\frac{\text{C.R of the 1st quarter} \times \text{L.R arithmetic mean of the 2nd quarter}}{100}$$

The chain relative of the 3rd quarter will be,

$$\frac{\text{C.R of the 2nd quarter} \times \text{L.R arithmetic mean of the 3rd quarter}}{100}$$

The chain relative of the 4th quarter will be,

$$\frac{\text{C.R of the 3rd quarter} \times \text{L.R arithmetic mean of the 4th quarter}}{100}$$

4. Then calculate the chain relative of the 1st quarter, for correction, as,

$$\frac{\text{1st quarter L.R} \times \text{4th quarter C.R}}{100}$$

The difference between the chain relative of 1st quarter taken previously and calculated now should be divided by the number of seasons. The resulting figure should be multiplied by 1, 2 and 3 respectively. Then the product so obtained should be deducted respectively from the chain relatives of 2nd, 3rd and 4th quarters. These are called the corrected chain relatives.

5. The corrected chain relatives should be represented in percentage to give the seasonal indices.

Illustration 12.14

Calculate the seasonal indices by the Link Relative Method from the data given below.

Year	Output of Wheat in Million Tonnes			
	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1995	30	35	31	35
1996	31	34	35	34
1997	35	35	32	31
1998	30	37	34	33
1999	32	40	35	39

Solutions

Computation of seasonal index by link relative method

Year	Link Relative			
	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1995	—	116.7	88.6	112.9
1996	88.6	109.7	102.9	97.1
1997	102.9	100.0	91.4	96.9
1998	96.8	123.3	91.9	97.1
1999	97.0	125.0	87.5	111.4
Total of Link Relatives	385.3	574.7	462.3	515.4
Arithmetic Mean	96.3	114.9	92.5	103.1
Chain Relatives	100	114.9	106.1	109.4
Adjusted Chain Relatives	100	113.6	103.4	105.4
Seasonal Index	92.9	105.57	96.09	97.95

Link Relative for any Quarter

$$\frac{\text{Current quarter's value}}{\text{Previous quarter's value}} \times 100$$

Link Relative for 2nd quarter of first year, i.e., 1995

$$\frac{35}{30} \times 100 = 116.7$$

Link Relative for 3rd quarter

$$\frac{31}{35} \times 100 = 88.6$$

$$\text{Chain Relative} = \frac{\text{Average L.R of Relative Quarters} \times \text{C.R of Previous Quarter}}{100}$$

Calculation of Correction factor

C.R for I Q (on the basis of IV Q),

$$\frac{96.3 \times 109.4}{100} = 105.35$$

Chain relative for Ist Quarter (on basis of Ist Quarter) = 100

The difference between these chain relatives,

$$105.35 - 100 = 5.35$$

$$\text{Hence, correction factor} = \frac{5.35}{4} = 1.34$$

Average of corrected chain relatives

$$\frac{100 + 113.56 + 104.76 + 108.06}{4} = 106.59$$

$$\text{Seasonal Index} = \frac{\text{Corrected Chain Relatives}}{\text{Averages of Corrected Chain Relatives}} \times 100$$

12.8 MISCELLANEOUS ILLUSTRATIONS

Illustration 12.15

Fit a straight line trend by the method of least square to the following data, relating to the net profits of a public concern.

Year	Profits (Rs ,000) (y)
1990	300
1991	700
1992	600
1993	800
1994	900
1995	700
1996	1000

Solutions

Year	Profits (Rs '000) (y)	(x) 1993	x^2	xy	Trend Values Y_c
1990	300	-3	9	-900	457.2
1991	700	-2	4	-1400	542.9
1992	600	-1	1	-600	628.6
1993	800	0	0	0	714.3
1994	900	1	1	900	800.0
1995	700	2	4	1400	885.7
1996	1000	3	9	3000	971.4
$N = 7$	$\sum y = 5000$	$\sum x = 0$	$\sum x^2 = 28$	$\sum xy = 2,400$	

536 Business Statistics

Equation:

$$Y_c = a + bx$$

Since $x = 0$,

$$a = \frac{\sum y}{N}$$

$$b = \frac{\sum xy}{\sum x^2}$$

Substituting

$$a = \frac{5000}{7} = 714.3$$

$$b = \frac{2400}{28} = 85.7$$

So, the equation of the straight line trend is

$$Y_c = 714.3 + 85.7x$$

Trend values are computed as follows:

When

$$x = -3$$

$$y = 714.3 + 85.7x$$

$$x = 714.3 + 85.7 \times (-3)$$

$$= 714.3 - 257.1$$

$$= 457.2$$

$$x = -2, y = 714.3 - 85.7x$$

$$= 714.3 + 85.7 \times (-2)$$

$$= 714.3 - 171.4$$

$$= 542.0$$

$$x = -1, y = 714.3 - 85.7x$$

$$= 714.3 + 85.7 \times (-1)$$

$$= 714.3 - 857$$

$$= 628.6$$

$$x = 0, y = 714.3 + 85.7 \times 0$$

$$= 714.3$$

$$x = 1, y = 714.3 - 85.7 \times 1$$

$$= 714.3 + 85.7$$

$$= 800$$

$$x = 2, y = 714.3 - 85.7 \times 2$$

$$= 714.3 + 171.4$$

$$= 885.7$$

$$x = 3, y = 714.3 - 85.7 \times 3$$

$$= 714.3 + 257.1$$

$$= 971.4$$

Illustration 12.16

Obtain the straight line trend equation and tabulate against each year after estimation of the trend and short-term fluctuations.

Year	Value
1990	380
1991	400
1992	650
1993	720
1994	690
1995	620
1996	670
1997	950
1998	1040

Solutions

Original Number	Year x	Value Y	x^2	xy	Trend Values Y_c	Short -term fluctuations ($Y - Y_c$)
1990	-4	380	16	-1520	398.0	-18.0
1991	-3	400	9	-1200	468.5	-68.5
1992	-2	650	4	-1300	539.0	111.0
1993	-1	720	1	-720	609.5	110.5
1994	0	690	0	0	680.0	10.0
1995	1	620	1	620	750.5	-130.5
1996	2	670	4	1340	821.0	-151.0
1997	3	550	9	2850	891.5	58.5
1998	4	1040	16	4160	962.0	78.0
	0	6120	60	4230		

Substituting the values in the two equations

$$\begin{aligned}
 \sum y &= Na + b \sum x \\
 \sum y &= 6120 \\
 N &= 9 \\
 \sum x &= 0 \\
 Na + b \sum x &= \sum y \\
 9 \times a + b \times 0 &= 6120 \\
 9a &= 6120 \\
 a &= 6120/9 \\
 a &= 680 \\
 \sum xy &= a \sum x + b \sum x^2 \\
 \sum xy &= 4230, \quad \sum x = 0, \quad \sum x^2 = 60 \\
 a \times 0 + b \times 60 &= 4230 \\
 b60 &= 4230 \\
 b &= 4230/60 \\
 b &= 70.5
 \end{aligned}$$

The trend equation

$$\begin{aligned}
 Y &= a + bx \\
 Y &= 680 + 70.5x
 \end{aligned}$$

The required trend values are computed from the trend equations as follows:

$$\begin{aligned}
 1990: \quad x &= -4, \\
 y &= 680 + 70.5(-4) \\
 &= 680 - 282 \\
 &= 398.0 \\
 1991: \quad x &= -3, \\
 y &= 680 + 70.5(-3) \\
 &= 680 - 211.5 \\
 &= 468.5 \\
 1992: \quad x &= -2, \\
 y &= 680 + 70.5(-2) \\
 &= 680 - 141 \\
 &= 539.0 \\
 1993: \quad x &= -1, \\
 y &= 680 + 70.5(-1) \\
 &= 680 - 70.5 \\
 &= 609.5
 \end{aligned}$$

1994:	$x = 0,$ $y = 680 + 70.5 \times 0$ = 680
1995:	$x = 1,$ $y = 680 + 70.5(1)$ = 680 + 70.5 = 750.5
1996:	$x = 2,$ $y = 680 + 70.5(2)$ = 680 + 141.0 = 821.0
1997:	$x = 3,$ $y = 680 + 70.5(3)$ = 680 + 211.5 = 891.5
1998:	$x = 4,$ $y = 680 + 70.5(4)$ = 680 + 282 = 962.0

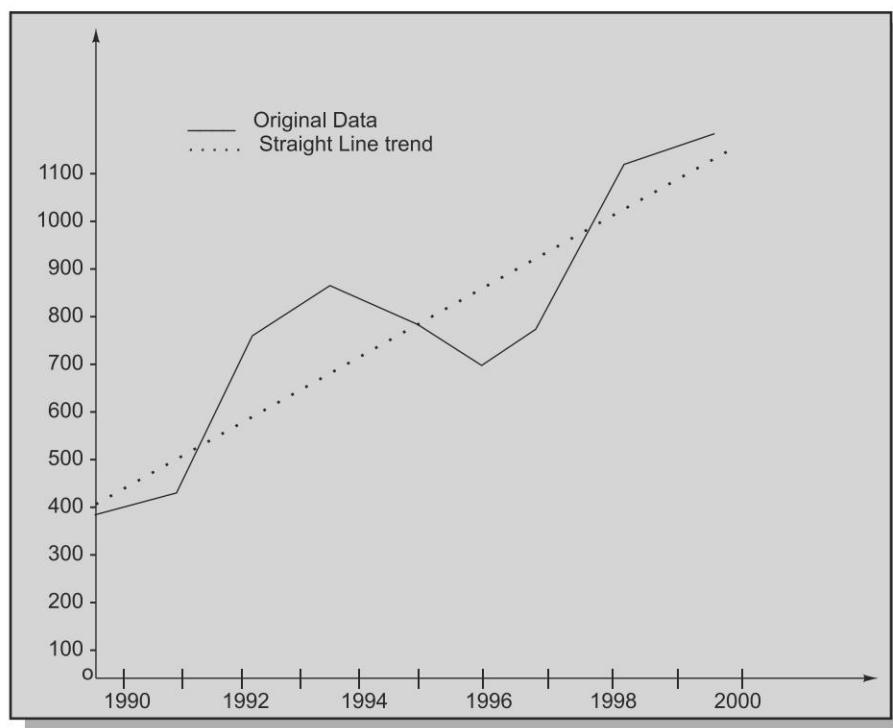


Illustration 12.17

Take a five-yearly period of moving average and determine short-term oscillations from the following data.

Year	Production ('000)
1989	14
1990	17
1991	22
1992	28
1993	26
1994	18
1995	20
1996	24
1997	25
1998	29
1999	30
2000	23

Solutions

Year (1)	Production (‘000) (2)	5-yearly Moving Total (3)	5-yearly Moving Average (3 ÷ 5)	Short-Term Oscillation (Y – Y _c)
1989	14	–	–	–
1990	17	–	–	–
1991	22	107	21.4	0.6
1992	28	111	22.2	5.8
1993	26	114	22.8	3.2
1994	18	116	23.2	– 5.2
1995	20	113	22.6	– 2.6
1996	24	116	23.2	0.8
1997	25	128	25.6	– 0.6
1998	29	131	26.2	2.8
1999	30	–	–	–
2000	23	–	–	–

Illustration 12.18

For the following table:

- Fit a straight line trend by the method of Least Square
- Calculate the trend values:

Year	Production
1990	12
1991	10
1992	14
1993	11
1994	13
1995	15
1996	16

Solutions

Year	Production y	x	x^2	xy	Y_c
1990	12	-3	9	-36	10.75
1991	10	-2	4	-20	11.50
1992	14	-1	1	-14	12.25
1993	11	0	0	0	13.00
1994	13	1	1	13	13.75
1995	15	2	4	30	14.50
1996	16	3	9	48	15.25
$N = 7$	$\sum y = 91$	$\sum x = 0$	$\sum x^2 = 28$	$\sum xy = 21$	$\sum Y_c = 91$

The Trend Equation:

$$Y_o = a + bx$$

$$a = \frac{\sum y}{N} = \frac{91}{7} = 13$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{21}{28} = 0.75$$

$$Y = 13 + 0.75x$$

$$\begin{aligned} Y_{1990} &= 13 + 0.75(-3) \\ &= 10.75 \end{aligned}$$

542 Business Statistics

The other trend value can be calculated by adding the value of b to the proceeding value.

Illustration 12.19

Fit a straight line trend by the method of Least Square to the following data.

Year	Production (Tonnes)
1991	24
1992	25
1993	29
1994	26
1995	22
1996	24

Solutions

Year	Production in Tonnes y	Deviations from 1993.5. $2x$	x^2	xy
1991	24	-5	25	-120
1992	25	-3	0	-75
1993	29	-1	1	-29
1994	26	1	1	26
1995	22	3	9	66
1996	24	5	25	120
N = 6	$\sum y = 150$	$\sum x = 0$	$\sum x^2 = 70$	$\sum xy = -12$

The Trend Equation

$$Y_o = a + bx$$

Since

$$\sum x = 0,$$

$$a = \frac{\sum y}{N} = \frac{150}{6} = 25$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{-12}{70} = -0.171$$

Hence,

$$Y = 25 - 0.171x$$

For 1998 x would be 9

Substituting the value of $x = 9$, we get

$$\begin{aligned} Y &= 25 - 0.171(9) \\ &= 25 - 1.539 \\ &= 23.461 \end{aligned}$$

Hence, the likely production for the year 1998 is 23.461 tonnes.

Illustration 12.20

Calculate trend by four year moving average of the following data given below.

Year	Production
1986	614
1987	615
1988	652
1989	678
1990	681
1991	655
1992	717
1993	719
1994	708
1995	779
1996	757

Solutions

Year	Production	4-yearly Moving Total	4-yearly Moving Average	4-yearly Moving Average centered
(1)	(2)	(3)	(4)	(5)
1986	614	—	—	—
1987	615	2559	639.75	648.125
1988	652	2626	656.50	661.500
1989	678	2666	666.50	674.625
1990	681	2731	682.75	687.875
1991	655	2772	693.00	696.375

Year	Production	4-yearly Moving Total	4-yearly Moving Average	4-yearly Moving Average centered
(1)	(2)	(3)	(4)	(5)
1992	717	2799	699.75	715.250
1993	719	2923	730.75	735.750
1994	708	2963	740.75	—
1995	779	—	—	—
1996	757	—	—	—

Illustration 12.21

Year	No. of students
1991	332
1992	317
1993	357
1994	392
1995	402
1996	405
1997	410
1998	427
1999	405
2000	438

Calculate five-yearly moving averages.

Solutions

Year	No. of students	5-yearly Moving Total	5-yearly Moving Average
1991	332	—	—
1992	317	—	—
1993	357	1800	360.0
1994	392	1873	374.6 or 375
1995	402	1966	393.2 or 393

Year	No. of students	5-yearly Moving Total	5-yearly Moving Average
1996	405	2036	407.2 or 407
1997	410	2049	409.8 or 410
1998	427	2085	417.0
1999	405	—	—
2000	438	—	—

Illustration 12.22

Using three-year moving averages, determine the trend and short-term fluctuations.

Year	Production in ('000 tonnes)
1983	21
1984	22
1985	23
1986	25
1987	24
1988	22
1989	25
1990	26
1991	27
1992	26

Solutions

Calculation of Trend and Short-Term Fluctuations

Year	Production ('000)	3-yearly Moving Total	3-yearly Moving Average (Trend Values) Y_c	Short-Term Fluctuations ($Y - Y_c$)
1983	21	—	—	—
1984	22	66	22.00	0
1985	23	70	23.33	-0.33
1986	25	72	24.00	1.00

546 Business Statistics

Year	Production ('000)	3-yearly Moving Total	3-yearly Moving Average (Trend Values) Y_c	Short-Term Fluctuations ($Y - Y_c$)
1987	24	71	23.67	0.33
1988	22	71	23.67	-1.67
1989	25	73	24.33	0.67
1990	26	78	26.00	0
1991	27	79	26.33	0.67
1992	26	-	-	-

Illustration 12.23

Calculate four-yearly moving averages from the following data.

Year	Value
1982	41
1983	61
1984	55
1985	48
1986	53
1987	67
1988	62
1989	60
1990	67
1991	73
1992	78
1993	76
1994	84

Solutions

Calculation of 4-yearly moving average

Year	Imported Consumption	4-yearly Moving Total	4-yearly Moving Average	4-yearly Moving Average centered
1982	41	—	—	—
1983	61	205	51.25	—
1984	55	217	54.25	52.75
1985	48	223	55.75	55.00
1986	53	230	57.50	56.62
1987	67	242	60.50	59.00
1988	62	256	64.00	62.25
1989	60	262	65.50	64.75
1990	67	278	69.50	67.50
1991	73	294	73.50	71.50
1992	78	311	77.75	75.62
1993	76	—	—	—
1994	84	—	—	—

Illustration 12.24

Fit a straight line trend to the following data on the domestic demand for motor fuel:

Year	Average monthly Demand (million barrels)
1985	61
1986	66
1987	72
1988	76
1989	82
1990	90
1991	96
1992	100
1993	103
1994	110
1995	114

Solutions

Fitting straight line trend

Year	Average monthly Demand (million barrels)	Deviation from 1990 <i>x</i>	<i>x</i> ²	<i>xy</i>
	<i>y</i>			
1985	61	-5	25	-305
1986	66	-4	16	-264
1987	72	-3	9	-216
1988	76	-2	4	-152
1989	82	-1	1	-82
1990	90	0	0	0
1991	96	1	1	96
1992	100	2	4	200
1993	103	3	9	309
1994	110	4	16	440
1995	114	5	25	570
<i>N</i> = 11	$\sum y = 970$	$\sum x = 0$	$\sum x^2 = 110$	$\sum xy = 596$

Since

$$\sum x = 0,$$

$$a = \frac{\sum y}{N} = \frac{970}{11} = 88.18$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{596}{110} = 5.418$$

Hence,

$$Y = 88.18 + 5.418x$$

For 2000 *x* shall be 10

$$\begin{aligned} Y_{2000} &= 88.18 + 5.418(10) \\ &= 88.18 + 54.18 = 142.36 \end{aligned}$$

Illustration 12.25

Compute the trend value for the following data using method of least squares.

Year	<i>y</i>
1990	83
1991	60
1992	54
1993	21
1994	22
1995	13
1996	23

Year	y	Deviations from 1993 x	xy	x^2	Y_o
1990	83	-3	-249	9	72.22
1991	60	-2	-120	4	61.29
1992	54	-1	-54	1	50.36
1993	21	0	0	0	39.43
1994	22	1	22	1	28.50
1995	13	2	26	4	17.57
1996	23	3	69	9	6.64
N = 7	$\sum y = 276$	$\sum x = 0$	$\sum xy = -306$	$\sum x^2 = 28$	$\sum Y_o = 276.01$

The Trend Equation:

$$Y_o = a + bx$$

Since,

$$\sum x = 0,$$

$$a = \frac{\sum y}{N} = \frac{276}{7} = 39.43$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{-306}{28} = -10.93$$

Hence,

$$Y = 39.43 - 10.93x$$

When

$$x = -3,$$

$$\begin{aligned} Y_{1990} &= 39.43 - 10.93(-3) \\ &= 39.43 + 32.79 = 72.22 \end{aligned}$$

The other trend values can be obtained by similar manner.

Illustration 12.26

The following data related to number of passenger car in million sold from 1989 to 1996:

Year	y
1989	6.7
1990	5.3
1991	4.3
1992	6.1
1993	5.6

Year	y
1994	7.9
1995	5.8
1996	6.1

- (a) Fit a straight line trend to the data through 1994 only.
 (b) Use your result in (a) to estimate production in 1996 and compare with the actual production.

Solutions

Fitting straight line trend

Year	No. of passenger cars y	Deviation from 1994 x	xy	x ²
1989	6.7	-5	-33.5	25
1990	5.3	-4	-21.2	16
1991	4.3	-3	-12.9	9
1992	6.1	-2	-12.2	4
1993	5.6	-1	-5.6	1
1994	7.9	0	0	0
1995	5.8	1	5.8	1
1996	6.1	2	12.2	4
N = 8	$\sum y = 47.8$	$\sum x = -212$	$\sum xy = -67.4$	$\sum x^2 = 60$

- (a) The equation of the straight line trend is

$$Y = a + bx$$

$$\sum y = Na + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

$$47.8 = 8a - 12b$$

$$-67.4 = -12a + 60b$$

Multiplying Eq. (i) by 3 and Eq. (ii) by 2

$$143.4 = 24a - 36b$$

$$-134.8 = -24a + 120b$$

$$\hline 8.6 = 84b$$

$$84b = 8.6;$$

$$\mathbf{b = 8.6 / 84 = .102}$$

Putting the value of (b) in Eq. (i)

$$47.8 = 8a - 12 \times .102$$

$$8a - 1.224 = 47.8$$

or

$$8a = 47.8 + 1.224$$

$$8a = 49.024$$

$$\mathbf{a = 6.128}$$

Thus, the required equation is $Y = 6.128 + .102x$

(b) Estimate for 1996

For 1996 x is 2

$$\begin{aligned} Y &= 6.128 + .102(2) \\ &= 6.128 + .204 \\ &= 6.332 \end{aligned}$$

Thus, the estimated sale for 1996 is 6.332 million cars. There is some difference in the actual sale figure which is 6.1 million passenger cars and the estimated figure. Some difference is likely to be there between the actual and estimated figures because estimates are based on certain assumptions. It may be a rare chance when actual and may completely coincide.

Illustration 12.27

Fit a straight line trend by the method of Least Squares for the following data.

Year	Reserves y
1987–88	612
1988–89	719
1989–90	820
1990–91	907
1991–92	1001
1992–93	1106
1993–94	1231

Solutions

Fitting straight line trend

Year	Reserves y	x	xy	x²
1987–88	612	– 3	– 1836	9
1988–89	719	– 2	– 1438	4
1989–90	820	– 1	– 820	1
1990–91	907	0	0	0
1991–92	1001	1	1001	1
1992–93	1106	2	2212	4
1993–94	1231	3	3693	9
N = 7	$\sum y = 6396$		$\sum xy = 2812$	$\sum x^2 = 28$

$$Y = a + bx$$

Since,

$$\sum x = 0,$$

$$a = \frac{\sum y}{N} = \frac{6396}{7} = 913.71$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{2812}{28} = 100.43$$

Hence,

$$Y = 913.71 + 100.43x$$

SUMMARY**Times Series**

Sequence of values of some variables corresponding to successive points in time.

Importance of Time Series

- Helps in understanding past behaviour.
- Facilitates for forecasting and planning.

Components of Time Series

Changes in the value of variables in different periods of time are due to several factors. They are: (a) Trend (b) Seasonal Changes (c) Cyclical changes (d) Irregular or Random.

Methods of Measuring Secular Trend

- Graphic Method
- Semi-Average Method

- Moving Average Method
- Method of Least Squares

De-seasonalisation

Means elimination of the seasonal effects from the given value. It can be done for multiplication model as well as addition model.

De-seasonalisation of Data for Multiplication Model

$$\frac{O}{S} = \frac{TCSI}{S} = TCI$$

De-seasonalisation of Data for Addition Model

$$O - S = T + C + I$$

Methods of Measuring Seasonal Variations

- Simple averages method
- Ratio to trend method
- Ratio to moving average method
- Link relative method

FORMULAE

Method of Least Squares

The equation of the straight line trend or linear trend in the first degree parabola is

$$Y_c = a + bx$$

Y_c = required trend value

X = unit of time

a and b = constants or unknowns

The values of the constants a and b can be calculated by the following two normal equations.

$$\sum y = Na + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

N = Number of years or months for which date are given.

When $x = 0$ (When middle year is taken as origin), the equation to know the values of a and b is

$$a = \frac{\sum y}{N}, b = \frac{\sum xy}{\sum x^2}$$

The equation of the second degree parabola is :

$$Y = a + bx + cx^2$$

EXERCISES

(a) Choose the best option

1. _____ method is the easiest method for calculating seasonal variations.
 - (a) Simple averages method
 - (b) Ratio to trend method
 - (c) Link relative method
2. _____ method is an improvement over the method of simple average method.
 - (a) Link relative method
 - (b) Ratio to moving average method
 - (c) Ratio to trend method
3. _____ method is also known as the percentage of moving average method.
 - (a) Ratio to trend method
 - (b) Ratio to moving average method
 - (c) Link relative method
4. _____ method requires complex mathematical calculations.
 - (a) Simple average method
 - (b) Ratio to trend method
 - (c) Ratio to moving average method
5. _____ method is the most difficult one among the methods of measuring seasonal variation.
 - (a) Link Relatives method
 - (b) Simple average method
 - (c) Ratio to trend method

Answers

1. a 2. c 3. b 4. b 5. a

(b) Fill in the blanks

1. A time series consists of data arranged _____.
2. The line obtained by method of least squares is known as the line of _____.
3. A polynomial of the form $y = a + by + cx^2$ is called a _____.
4. The active model of a time series is expressed as _____.
5. Time series is the _____ of statistical data on the basis of time.

6. The trend may be defined as the _____ over a long period of time.
7. Seasonal variations are measured for _____.
8. Irregular movements are otherwise called as _____ variations.
9. The graphic method is otherwise called as _____.
10. Moving average method is logical extension of _____ method.

Answers

- | | |
|---------------------------|------------------------|
| 1. Chronologically | 2. Best fit |
| 3. Second degree equation | 4. $y = T + S + C + I$ |
| 5. arrangement | 6. changes |
| 7. one calendar year | 8. random |
| 9. free hand method | 10. semi-average |

(c) Theoretical Questions

1. What is a time series? What are the main components? Give illustrations for each of them. (B.Com., CHU, MKU, BDU)
2. Discuss briefly the importance of time series analysis in business and economics. What are the components of a time series? Give an example of each component. (B.Com., CHU, MSU, BDU)
3. Distinguish between additive model and multiplicative model in the analysis of time series.
4. Give the addition and multiplication models of the time series equations and explain briefly the components of a time series. (B.Com., CHU, MKU, BDU)
5. Define trend. Enumerate the different methods of measuring secular trend in a given time series. (B.Com., B.B.A., MSU, MKU, BDU)
6. Discuss the statistical procedure you would adopt in the analysis of time series and explain how you will isolate the secular trend. (B.Com., CHU, MKU, BDU)
7. Distinguish between the seasonal component and trend component of a time series. (B.Com., CHU, BU, MSU, BDU)
8. Distinguish between secular trend, seasonal variations and cyclical fluctuations. How would you measure secular trend in any given data?
9. What are secular trend and cyclical seasonal and irregular fluctuations? Describe the methods of isolation of trend.

10. Explain the principle of least squares. How is it used in trend fitting? What are the relative merits and demerits of trend fitting by the principle of least squares?
11. Distinguish between seasonal variation, cyclical variation and secular trend. **(B. Com., CHU, MKU, BDU)**
12. Give an account of the common components of time series data. Explain any one method of obtaining long-term trend. **(B. Com., MKU, MSU, BU, CHU)**
13. Explain the importance of time series analysis in business forecasting. **(B. Com., B.B.A., MKU, MSU, BDU)**
14. What is a Time Series? What is the main object of constructing a time series? Explain fully the components of a time series.
15. Distinguish between trend, seasonal variations and cyclical fluctuations in a time series. How can trend be isolated from fluctuations?
16. What is ‘moving average’? What are its uses in time series?
17. What do you mean by seasonal variation? Explain with a few examples the utility of such a study.

(d) Practical Problems

18. Outline the nature of seasonal variations. How do they differ from cyclical variations? Compute the seasonal index for the following data assuming that there is no need to adjust the data for the trend.

Quarter	1990	1991	1992	1993	1994	1995
I	3.5	3.5	3.5	4.0	4.1	4.2
II	3.9	4.1	3.9	4.6	4.4	4.6
III	3.4	3.7	3.7	3.8	4.2	4.3
IV	3.6	4.8	4.0	4.5	4.5	4.7

(B. Com., CHU, BU, BDU)

19. Compute the average seasonal movement for the following series;

Year	Quarterly Production			
	I	II	III	IV
1978	3.5	3.9	3.4	3.6
1979	3.5	9.1	3.7	4.0
1980	3.5	3.9	3.7	4.2
1981	4.0	4.6	3.8	4.5
1982	4.1	4.4	4.2	4.5

Answer Seasonal Indices = 94.17, 105.8, 95.19, 105.3

20. Analyse the seasonal variation in the world production of coal

Quarter / Year	Production of Coal (Million metric tonnes)			
	1970	1971	1972	1973
I	518	536	507	547
II	518	525	527	528
III	506	510	505	521
IV	533	481	531	502

(B. Com., CHU, MKU, BDU, BU)

21. The number (in hundreds) of letters posted in a certain city on each day in a typical period of five weeks was as follows:

Week	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1st	18	161	170	164	153	181	76
2nd	18	165	179	157	168	195	85
3rd	21	162	169	153	139	185	82
4th	24	171	182	170	162	179	95
5th	27	162	186	170	170	182	120

Calculate seasonal variations.

(B. Com., CHU, BU, MSU, BDU)

Answer Seasonal indices: 15.75; 119.71; 129.18; 118.68; 115.48; 134.43; 66.78

22. Find seasonal variations by the ratio of trend method from the data given below:

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1995	30	40	36	34
1996	34	52	50	44
1997	40	58	54	48
1998	54	75	68	62
1999	80	92	86	82

Answer Straight line trend is given by $Y = 56 + 12x$, Origin: 1995 (1st July); x unit = 1 year; y units = Average quarterly values.

Seasonal Indices : 92.0; 117.4, 102.1, 88.5

(B. Com., MKU, BDU)

23. Determine the cyclical variations for the following data assuming that

- (a) data are based on additive model and
- (b) data are based on multiplicative model.

Year	1972	1973	1974	1975	1976	1977	1978
Production (in million tonnes)	70	72	80	76	84	88	90

Answer (a) 0.29, -1.14, 3.43, -4.00, 0.57, 1.14, -0.29,

(b) 0.42, -1.56, 4.48, -5.00, 0.68, 1.31, -0.32

- 24.** You are given the annual profit figures for a certain firm for the year 1980 to 1986. Fit a straight line trend to the data and estimate the expected profit for the year 1987.

Year	1980	1981	1982	1983	1984	1985	1986
Profit (Rs Lakhs)	60	72	75	65	80	85	95

Answer 1987: 95.44 Lakhs) (B.A., BBA., CHU, MKU)

- 25.** The following are the annual profits in thousands of rupees in an industrial concern:

Year	1982	1983	1984	1985	1986	1987	1988	1989	1990
Profit (Rs '000)	65	77	80	70	89	95	109	102	105

Use the method of least squares to fit a straight line trend to the above data. Also make an estimate of profit on 2000.

Answer 162.2 thousands (B. Com., MKU, MSU, BDU)

- 26.** Fit a straight line trend by the method of least squares to the following data. Assuming that the same rate of change continues, what would be the earning for the year 1992?

Year	1983	1984	1985	1986	1987	1988	1989	1990
Earnings (Rs Lakhs)	38	40	65	72	69	60	87	95

Answer 106.12 lakhs

- 27.** Population figures for a city are as given below. Fit a curve of the type $Y = a + bx$ and estimate the population for 1987.

Year	1981	1982	1983	1984	1985
Population (in '000s)	132	142	157	170	191

Answer 226.8 thousand

- 28.** The following table shows the number of salesman working for a certain concern.

Year	1990	1991	1992	1993	1994
Number	28	38	46	40	56

Use the method of Least Square to fit a straight line and estimate the number of salesman in 1995. (B. Com., MKU, MSU, BU)

Answer $Y = 41.6 + 5.8x$, 59

29. Calculate seasonal indices by the ratio to moving average method from the following data.

Year	I Quarter	II Quarter	III Quarter	IV Quarter
1985	68	62	61	63
1986	65	58	66	61
1987	68	63	63	67

Answer 105.30, 95.21, 100.97, 98.52

30. From the data given below, determine the line of trend and hence find the expected value for 1992.

Year	1982	1983	1984	1985	1986	1987	1988	1989	1990
Production (in tonnes)	110.2	143.3	143.3	134.5	138.55	74	129	150	140

Answer 121.07 tonnes **(B. Com., BDU, BU, MSU)**

31. Obtain seasonal fluctuations from the following time series:
Quarterly output of coal for four years.

Quarter	I	II	III	IV
Year				
1998	65	58	56	61
1999	58	63	63	67
2000	70	59	56	52
2001	60	55	51	58

Answer 5.37; -1.29; -2.96; -0.96

(B. Com., CHU, MSU, BU, BDU)

13

CHAPTER

INTERPOLATION AND EXTRAPOLATION

13.1 INTRODUCTION

Interpolation and extrapolation are important statistical estimations of a variable when a set of variables are given. It is used to find out a missing figure for past or future period. Interpolation or extrapolation is a technique of obtaining the most likely estimate of a certain quantity (dependent variable) from the given relevant facts under certain assumptions.

For example, the sales for the year 2006 can be estimated from the sales of 2000–2005 and 2007–2012 period. If sales for 2000–2005 are available, sale for 1998, 1999, 2006, 2007 etc., can be estimated through interpolation and extrapolation techniques.

13.2 DEFINITIONS

The important definitions of interpolation and extrapolation are stated as below:

13.2.1 Interpolation

Interpolation is the estimation of a most likely estimate in given conditions. The technique of estimating a past figure is termed as interpolation while that of estimating a probable figure for future is called extrapolation.

—Hirach

Interpolation is the art of leading between the lines of the table.

—Theile

Interpolation consists in reading a value which lies between two extreme points extrapolation means reading a value that lies outside to two extreme points.

—W.M. Harper

It is clear from the above definitions that interpolation is a measure of a value in between a set of variables; for example, calculation of the value of June 2000 when, the value of the other 11 months of the year are available.

13.2.2 Extrapolation

Extrapolation is a measure to find out a value in future on the basis of the values of past. For example, calculation of the value of 2005 when the values for 1997, 1998, 1999, 2000, 2001, 2002 are available.

13.2.3 Assumptions

The following are the important assumptions for the interpolation and extrapolation method:

- (i) The variables should not raise and fall over a period of time. It means that the variables should not assume sudden jump or fall from period to period.
- (ii) The rate of change of figures should be uniform from one period to another.
- (iii) It is advisable that the time span for which data are available should be uniform one.

13.3 USES

1. It is the method available to fill the missing figures or estimating intermediate figures or finding projections or destroyed figures. For example, population of any intermediate year of decennial census.
2. It is freely used to compute the value of median and mode in continuous series.
3. Continuity of information can be received through the technique of interpolation, in case, figures are missing or destroyed.
4. Prediction of the future or estimating the future, in economic planning, policy formulation, etc. can be gainfully attained through the scientific technique of interpolation and extrapolation.

13.4 METHODS OF INTERPOLATION AND EXTRAPOLATION

The methods of interpolation or extrapolation may be broadly classified as follows.

1. Graphic method and
2. Algebraic method.

The important algebraic methods are as follows:

1. Binomial Expansion method
2. Newton's method:
 - (a) Newton's Gauss (forward) method
 - (b) Newton's Gauss (backward) method

- (c) Newton's method of backward differences
- (d) Newton's divided difference method
- 3. Lagrange's method and
- 4. Parabolic curve method.

13.4.1 Graphic Method

In this method, the available data should be plotted on a graph paper. The plotted points should be joined so as to get a curve. If there are only two points, then a straight line can be drawn instead of a curve. The missing figure for a period can easily be obtained with the help of this curve or straight line.

Steps The following are the steps to be followed for graphic method.

- (i) Draw X axis and Y axis in a graph paper.
- (ii) On the X axis, mark the periods conveniently.
- (iii) Plot the points for respective variables corresponding to the periods.
- (iv) The points should be joined to get a straight line or curve.

Merits

- (i) This is the easiest method for measuring interpolation.
- (ii) It is time consuming than any other method.
- (iii) It can be used for non-linear situations also.

Demerits

- (i) It is not an accurate method for extrapolation.
- (ii) The shape of the curve may change due to changes in the space provided for years and for variables.
- (iii) This method could not be a clear one when the number of variables are more.

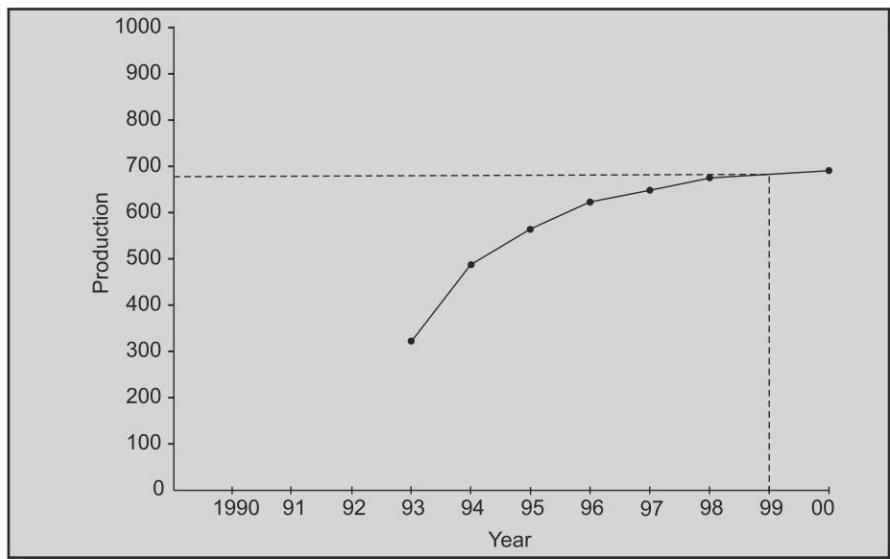
Illustration 13.1

Find out the production of a factory for the year from the following data through graphic method 1999.

Year	1993	1994	1995	1996	1997	1998	2000
Production (in tonnes)	325	490	570	620	650	680	695

Solutions

Calculation of production of a factory.



The estimated production for 1999 is 690 tonnes.

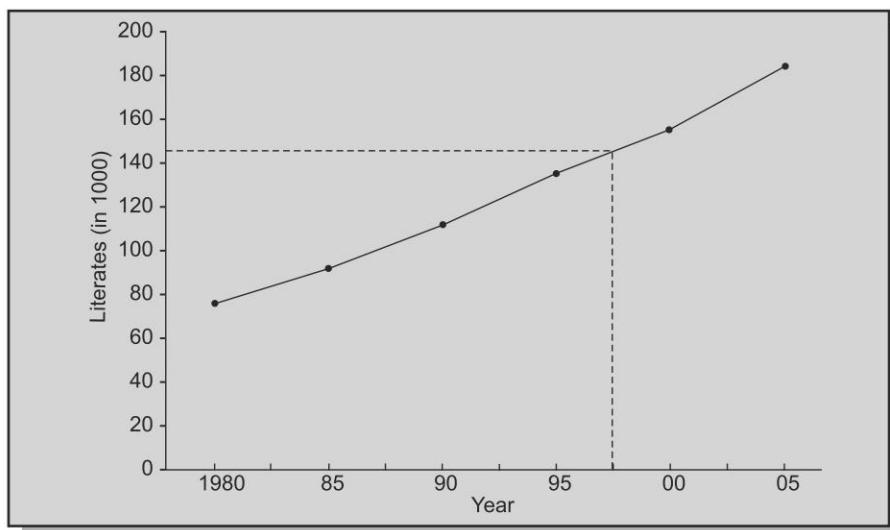
Illustration 13.2

Interpolate literate in a village by graphic method for the year 1997 from the following data:

Year	1980	1985	1990	1995	2000	2005
Literates (in 1000)	78	92	112	136	156	186

Solutions

Interpolation of literates for the year 1997.



The estimated number of literates in 1997 is 148.

13.4.2 Algebraic Method

Binomial Expansion Method This method is one of the easiest methods of interpolation. It can be applied only for specific type of problems. The independent variables (x) should advance by equal intervals. For example, 10, 20, 30, 40, 50. The value of X to be interpolated should be anyone of the variables in between the given values. For example, if the variables for x series are 10, 20, 30, 40, 50 then the value of Y for any of the X variables can be calculated under this method.

If all the values of Y for the X variables are available except for variable 20, then the value of Y for the variable 20 can be calculated. But the value of Y for 25 or 32 for X (independent variable) can be calculated through this method.

The formula for interpolation under the binomial expansion method is given below.

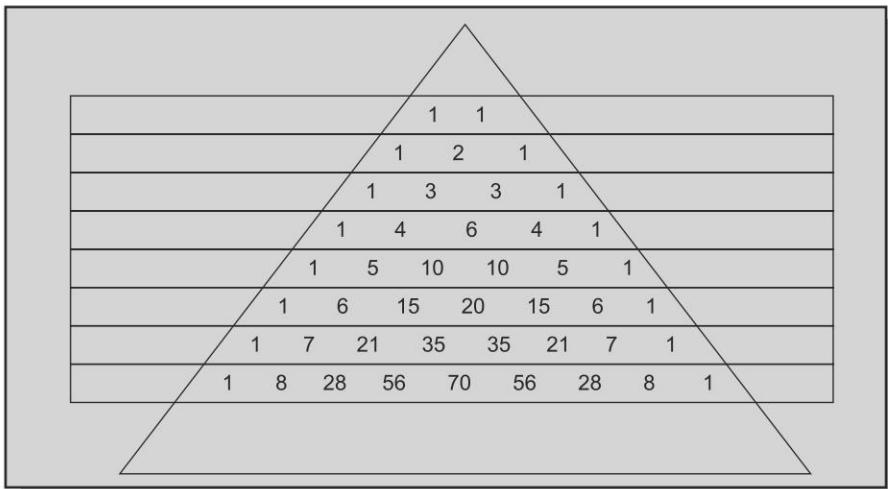
$$(y - 1) - y^n - ny^{n-1} + \frac{n(n-1)}{2!} y^{n-2} - \frac{n(n-2)}{3!} Y^{n-3} + \dots = 0$$

where y is the dependent variable and n is the number of known values of Y . The above equation is based on the number of known values of Y . Hence, the equations for the known values of Y can be framed as follows:

No. of known values	Equation for determining the unknown values
2 or Δ_0^2	$Y_2 - 2Y_1 + Y_0 = 0$
3 or Δ_0^3	$Y_3 - 3Y_2 + 3Y_1 - Y_0 = 0$
4 or Δ_0^4	$Y_4 - 4Y_3 + 6Y_2 - 4Y_1 + Y_0 = 0$
5 or Δ_0^5	$Y_5 - 5Y_4 + 10Y_3 - 10Y_2 + 5Y_1 - Y_0 = 0$
6 or Δ_0^6	$Y_6 - 6Y_5 + 15Y_4 - 20Y_3 + 15Y_2 - 6Y_1 + Y_0 = 0$
7 or Δ_0^7	$Y_7 - 7Y_6 + 21Y_5 - 35Y_4 + 35Y_3 - 21Y_2 + 7Y_1 - Y_0 = 0$
8 or Δ_0^8	$Y_8 - 8Y_7 + 28Y_6 - 56Y_5 + 70Y_4 - 56Y_3 + 28Y_2 - 8Y_1 + Y_0 = 0$
9 or Δ_0^9	$Y_9 - 9Y_8 + 36Y_7 - 84Y_6 + 126Y_5 - 126Y_4 + 84Y_3 - 56Y_2 + 9Y_1 - Y_0 = 0$

Procedure

- (i) Arrange the dependent variables in a continuous order and give symbol for it as y_1, y_2, y_3, y_4 etc.
- (ii) The first value of the equation would be Y_{n-r} .
- (iii) After the first value, the second and subsequent values of the equation would be y_{n-2}, y_{n-3} etc.
- (iv) The minus symbol should be given after the first value and subsequently (+) and (-) should be given, alternatively for all the values.
- (v) The numerical coefficients for the first value and last values of y in the equation are one. The other numerical coefficients of equations can be calculated through the Pascal's triangle which is given below.



If the known variables of Y series are 8 in number then to know the unknown variable, the binomial equation $(y - 1)^8 = 0$ should be applied. The co-efficient at Y can be found out from the above Pascals triangle which are 1, 8, 28, 56, 70, 56, 28, 8, 1.

Illustration 13.3

Find out the missing item of the following data by binomial expansion method.

Year	2000	2001	2002	2003	2004	2005
Sales (in tonnes)	40	49	52	?	73	89

Solutions

Calculation of the value for 2003.

X	2000 X_0	2001 X_1	2002 X_2	2003 X_3	2004 X_4	2005 X_5
Y	40 Y_0	49 Y_1	52 Y_2	0 Y_3	73 Y_4	89 Y_5

The known variables are five.

Hence, $\Delta_0^5 = 0$

$$\begin{aligned}\Delta_0^5 &= Y_5 - 5Y_4 + 10Y_3 - 10Y_2 + 5Y_1 - Y_0 = 0 \\ \Rightarrow 89 - 5(73) + 10(Y_3) - 10(52) + 5(49) - 40 &= 0\end{aligned}$$

$$89 - 365 - 10Y_3 - 520 + 245 - 40 = 0$$

$$10Y_3 - 591 = 0$$

$$10Y_3 = 591 \quad Y_3 = 59.1 = 59$$

$$\therefore Y_3 = 59$$

Calculation of Two or More Missing Values through the Binomial Expansion Method Binomial expansion method is the most suitable method to find out two or more missing values. The number of constants are formed. For example, if two variables are missing then the binomial equation $\Delta^n = 0$ and $\Delta^{n-1} = 0$ should be constructed and the two unknown variables can be calculated by solving the equations.

Illustration 13.4

From the following data, estimate the average profit of an industry for the year 2002 and 2005.

Year	2000	2001	2003	2004	2006	2007
Profit (Rs in lakhs)	84	91	98	105	112	119

Solutions

Year	2000	2001	2002	2003	2004	2005	2006	2007
Sales (in tonnes)	84	91	—	98	105	—	112	119
	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7

Two equations should be framed. There are 2 missing terms. They are Δ_0^7 and Δ_0^6

$$\Delta_0^7 = y_7 - 7y_6 + 21y_5 - 35y_4 + 35y_3 - 21y_2 + 7y_1 - y_0 = 0$$

$$119 - 7(112) + 21y_5 - 35(105) + 35(98) - 21y_2 + 7(91) - 84 = 0$$

$$119 - 784 + 21y_5 - 3675 + 3430 - 21y_2 + 637 - 84 = 0$$

$$21y_5 - 21y_2 - 357 = 0$$

$$21y_5 - 21y_2 = 357$$

$$\% \text{ by 21} \quad \Rightarrow \quad y_5 - y_2 = 17 \quad (1)$$

$$\Delta_0^6 = y_6 - 6y_5 + 15y_4 - 20y_3 + 15y_2 - 6y_1 + y_0 = 0$$

$$112 - 6y_5 + 15(105) - 20(98) + 15y_2 - 6(91) + 84 = 0$$

$$112 - 6y_5 + 1575 - 1960 + 15y_2 - 546 + 84 = 0$$

$$- 6y_5 + 15y_2 - 735 = 0$$

$$- 6y_5 + 15y_2 = 735$$

$$\% \text{ by 3} \quad \Rightarrow \quad - 2y_5 + 5y_2 = 245 \quad (2)$$

Multiply Eq. 1 divide by Eq. 2

$$\begin{array}{r}
 2y_5 - 2y_2 = 34 \\
 - 2y_5 + 5y_2 = 245 \\
 \hline
 3y_2 = 279 \\
 y_2 = 93
 \end{array}$$

Substitute y_2 in Eq. 1

$$y_5 - y_2 = 17$$

$$\begin{aligned}y_5 - 93 &= 17 \\y_5 &= 17 + 93 \\y_5 &= 110\end{aligned}$$

Newton's Method

(a) Newton's Method of Advancing Differences Sometimes the difference between two independent variables is a constant number which is same for all the differences of the two adjacent independent variables.

For example, (i) 25, 40, 55, 70, 85 (ii) 5, 10, 15, 20, 25, 30

In the example (i) difference between two variables is 15 which is same for all the differences and in example (ii) the difference is 5 which is same for all the differences of adjacent variables in the series.

When the differences in the independent variables are same, then Newton's method of advancing differences can be used for interpolation and extrapolation.

The formula under this method to find out the value of a dependent variable (Y) for an independent variable (X) is

$$y_x = y_0 + x\Delta_0 + \frac{x(x-1)}{1*2}\Delta_0^2 + \frac{x(x-1)(x-2)}{1*2*3}\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{1*2*3*4}\Delta_0^4 + \dots +$$

where y_x = Dependent variable to be calculated for a given independent variable

y_0 = Value of Y in the origin

Δ = Difference between various values of Y

The value of X can be calculated through the following formula.

$$X = \frac{\text{The value to be interpolated} - \text{The value of } X \text{ in the origin}}{\text{The difference of two adjacent variables of } X}$$

If the value of X is given in years then

$$X = \frac{\text{Year of Interpolation} - \text{Year of origin}}{\text{Difference between two adjoining years}}$$

The difference between the various values of Y (Δ) can be calculated as follows.

$$\begin{aligned}\Delta_0^1 &= y_1 - y_0 \\ \Delta_1^1 &= y_2 - y_1 \\ \Delta_2^1 &= y_3 - y_2 \text{ etc.} \\ \Delta_0^2 &= \Delta_1^1 - \Delta_0^1 \\ \Delta_1^2 &= \Delta_2^1 - \Delta_1^1 \text{ etc.} \\ \Delta_0^3 &= \Delta_1^2 - \Delta_0^2 \\ \Delta_1^3 &= \Delta_2^2 - \Delta_1^2 \text{ etc.} \\ \Delta_0^4 &= \Delta_1^3 - \Delta_0^3 \\ \Delta_1^4 &= \Delta_2^3 - \Delta_1^3 \text{ etc.}\end{aligned}$$

Illustration 13.5

From the following data, interpolate the value for the year 1989.

Year	1985	1990	1995	2000	2005
Sales (in tonnes)	280	310	360	380	410

Solutions

Calculation of the difference of Y .

X	1985 X_0	1990 X_1	1995 X_2	2000 X_3	2005 X_4
Y	280 Y_0	310 Y_1	360 Y_2	380 Y_3	410 Y_4
X	Y	$\Delta^1 Y$	$\Delta^2 Y$	$\Delta^3 Y$	$\Delta^4 Y$
1985	280				
1990	310	30			
1995	360	50	20	-50	
2000	380	20	-30	40	90
2005	410	30			

$$x = \frac{\text{Year of Interpolation} - \text{Year of origin}}{\text{Difference between two adjoining years}} = \frac{X - X_0}{n} = \frac{1989 - 1985}{5} = 4/5$$

$$x = 0.8$$

$$\begin{aligned} y_x &= y_0 + x\Delta_0 + \frac{x(x-1)}{2 \times 1} \Delta_0^2 y + \frac{x(x-1)(x-2)}{3 \times 2 \times 1} \Delta_0^3 y + \frac{x(x-1)(x-2)(x-3)}{4 \times 3 \times 2 \times 1} \Delta_0^4 y \\ &= 280 + (0.8)(30) + \frac{0.8(0.8-1)}{2 \times 1} (20) + \frac{0.8(0.8-1)(0.8-2)}{3 \times 2 \times 1} (-50) + \\ &\quad \frac{0.8(0.8-1)(0.8-2)(0.8-3)}{4 \times 3 \times 2 \times 1} (90) \\ &= 299.216 \approx 299 \end{aligned}$$

∴ The value for the year 1989 = 299

Illustration 13.6

Interpolate the value of Y for the value of x as 32 from the following data.

X	10	20	30	40	50
y	47	52	74	82	95

Solutions

X	10	20	30	40	50
	X_0	X_1	X_2	X_3	X_4
Y	47	52	74	82	95
	Y_0	Y_1	Y_2	Y_3	Y_4
X	Y	Δ_0	Δ_0^2	Δ_0^3	Δ_0^4
10	47				
20	52	5	17		
30	74	22	-14	-31	50
40	82	8	5	19	
50	95	13			

$$X = \frac{\text{Year of Interpolation} - \text{Year of Origin}}{\text{Diff between two adjoining years}} = \frac{X - X_0}{n} = \frac{32 - 10}{10} = 22/10$$

$$\therefore x = 2.2$$

$$\begin{aligned}
 Y_x &= Y_0 + X \Delta_0^1 + \frac{x(x-1)}{2 \times 1} \Delta_0^2 + \frac{x(x-1)(x-2)}{3 \times 2 \times 1} \Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{4 \times 3 \times 2 \times 1} \Delta_0^4 \\
 &= 47 + 2.2(5) + \frac{2.2(2.2-1)}{2}(17) + \frac{2.2(2.2-1)(2.2-2)}{6}(-31) \\
 &\quad + \frac{2.2(2.2-1)(2.2-2)(2.2-3)}{24}(50)
 \end{aligned}$$

$$= 47 + 11 + 22.44 - 2.728 - 0.88$$

$$= 76.832$$

$$\therefore y_{32} = 76.832 \text{ or } 77$$

Illustration 13.7

Find out the total number of students who got less than 54 marks from the following data.

Marks	0–30	30–50	50–70	70–90
No. of Students	30	25	20	15

Solutions

X	30	50	70	90
	X_0	X_1	X_2	X_3
Y	30	55	75	90
	Y_0	Y_1	Y_2	Y_3

X	Y	Δ_0	Δ_0^2	Δ_0^3
30	30			
50	55	25		
70	75	20	-5	
90	90	15	-5	0

$$x = \frac{X - X_0}{n} = \frac{54 - 30}{20} = 1.2$$

$$\begin{aligned} y_x &= y_0 + x \Delta_0^1 + \frac{x(x-1)}{2 \times 1} \Delta_0^2 + \frac{x(x-1)(x-2)}{3 \times 2 \times 1} \Delta_0^3 \\ &= 30 + 1.2(25) + \frac{1.2(1.2-1)}{2}(-5) + \frac{(1.2)(1.2-1)(1.2-2)}{6}(0) \\ &= 30 + 30 - 0.6 + 0 \\ &= 59.4 \\ \therefore y_{54} &= 59.4 \end{aligned}$$

(b) **Newton-Gauss (forward) Method** This method can be applied only when the independent variable x increases by equal intervals and the variable to be interpolated lies in the middle of the series. The formula under this method is

$$\begin{aligned} y_x &= y_0 + x \Delta_0^1 y - 0 + \frac{x(x-1)}{1 \times 2} + \Delta_0^2 y - 1 + \frac{x(x-1)(x-2)}{1 \times 2 \times 3} \Delta_0^3 y - 1 \\ &\quad + \frac{x(x-1)(x-2)(x-3)}{1 \times 2 \times 3 \times 4} \Delta_0^4 y - 2 \end{aligned}$$

$$x = \frac{\text{Variable } x \text{ to be interpolated} - \text{Preceding variable to the item to be interpolated}}{\text{Difference between any two adjoining variables } x}$$

Steps

- The next preceding variable to the item to be interpolated of the x series should be denoted by x_0 .
- The previous variables to x_0 should be denoted as x_1, x_2, x_3 etc.
- The succeeding variables to x_0 should be denoted as x_1, x_2, x_3 etc.
- The corresponding values of Y variables to X variables should be denoted by y_0, y_1, y_2, y_3 , etc., (For x_0, x_1, x_2, x_3 etc.,) and y_{-1}, y_{-2}, y_{-3} , etc., (for x_{-1}, x_{-2}, x_{-3} etc.)
- Prepare the difference table and find out the value of y_{x_0} .

Illustration 13.8

Calculate the value of y when the value of X is 14 from the following data.

X	4	8	12	16	20	24
y	12	18	22	36	45	62

Solutions

Y	X	Δ_1	Δ_2	Δ_3
$x_{-2} 4$	$12y_{-2}$	$6\Delta^1 y_{-2}$		
$x_{-1} 8$	$18y_{-1}$	$4\Delta^1 y_{-1}$	$-2\Delta^2 y_{-2}$	$12\Delta^3 y_{-2}$
$x_0 12$	$22y_0$	$14\Delta^1 y_0$	$10\Delta^2 y_{-1}$	$-15\Delta^3 y_{-1}$
$x_1 16$	$36y_1$	$9\Delta^1 y_1$	$-5\Delta^2 y_0$	$13\Delta^3 y_0$
$x_2 20$	$45y_2$		$8\Delta^2 y_1$	
$x_3 24$	$62y_3$	$17\Delta^1 y_2$		

Here, 14 is selected for X_0 , since it is the next proceeding item to $12x =$

$$\frac{14 - 12}{4} = 0.5$$

$$Y_x = y_0 + X \Delta^1 y_0 + \frac{x(x-1)}{2} \Delta^2 y_1 - 1 + \frac{x(x-1)(x-2)}{3 \times 2 \times 1} \Delta^3 y_2 - 1$$

$$y_{14} = 22 + (0.5 \times 14) + \frac{0.5(0.5-1)}{2 \times 1} (10) + \frac{0.5(0.5-1)(0.5-2)}{3 \times 2 \times 1} (-15)$$

$$= 22 + 7 - 1.25 - 0.94$$

$$= 26.8$$

(c) **Newton-Gauss (Backward) Method** This method of interpolation can be applied only when the variable to be interpolated lies at the end of the series and the difference between all the two adjoining variables x are the same. In this method, the next succeeding item of the x series to the item to be interpolated should be taken as x_0 .

Steps

- The next succeeding variable to the item to be interpolated of the X series should be denoted by X_0 .
- The previous variables to x_0 should be denoted as x_{-1}, x_{-2}, x_{-3} etc.
- The succeeding variables to X_0 should be denoted as X_1, X_2, X_3 , etc.
- The corresponding value of y variables to X variable should be denoted by y_0, y_1, y_2, y_3 , etc. (for x_0, x_1, x_2, x_3 etc.) and y_{-1}, y_{-2}, y_{-3} etc. (for x_{-1}, x_{-2}, x_{-3} etc.)
- Prepare the difference table and find out the value of y_x .

Then the value so calculated should be substituted in the following formula.

$$y_x = y_0 + X\Delta^1 y_{-1} + \frac{x(x+1)}{1 \times 2} \Delta^2 y_{-1} + \frac{x(x+1)(x-1)}{1 \times 2 \times 3} \Delta^3 y_{-2} + \frac{x(x+1)(x-1)(x-2)}{1 \times 2 \times 3 \times 4} \Delta^4 y_{-2}$$

$$x = \frac{\text{Variable } x \text{ succeeding the item to be interpolated} - \text{Variable } x \text{ to be interpolated}}{\text{Difference between any two adjoining variables } x}$$

Illustration 13.9

From the following data find out the sales for the month of September.

Months	Feb	Apr	June	Aug	Oct	Dec
Sales (in units)	250	310	450	570	680	750

Month	X	Y	Differences				
			Δ_1	Δ_2	Δ_3	Δ_4	Δ_5
Feb	$x_{-4} 2$	$x_{-4} 250$	$60\Delta^1 y_{-4}$				
Apr	$x_{-3} 4$	$x_{-3} 310$		$80\Delta^2 y_{-4}$			
Jun	$x_{-2} 6$	$x_{-2} 450$	$140\Delta^1 y_{-3}$	$-20\Delta^2 y_{-3}$	$-100\Delta^3 y_{-4}$		
Aug	$x_{-1} 8$	$x_{-1} 570$	$120\Delta^1 y_{-2}$	$-10\Delta^2 y_{-2}$	$10\Delta^3 y_{-3}$	$-150\Delta^4 y_{-4}$	
Oct	$x_0 10$	$x_0 680$	$110\Delta^1 y_{-1}$		$-30\Delta^3 y_{-2}$	$-40\Delta^4 y_{-3}$	
Dec	$x_1 12$	$x_1 750$	$70\Delta^1 y_0$	$-40\Delta^2 y_{-1}$			

$$X = \frac{\text{Variable } X \text{ succeeding item to be interpolated} - \text{Variable } x \text{ to be interpolated}}{\text{Difference between any two adjoining variables } x}$$

$$x = \frac{\text{Oct} - \text{Sep}}{\text{Feb} - \text{Apr}} = \frac{10 - 9}{2} = \frac{1}{2} = 0.5$$

$$Y_x = Y_0 - X\Delta^1 y_{-1} + \frac{x(x-1)}{2!} \Delta^2 y_{-1} - \frac{x(x+1)(x-1)}{3!} \Delta^3 y_{-2} + \frac{x(x+1)(x-1)(x-3)}{4!} \Delta^4 y_{-2}$$

$$= 680 - 0.5(110) + \frac{0.5(0.5+1)}{2} (-40) - \frac{0.5(0.5+1)(0.5-1)}{3 \times 2 \times 1} (-30)$$

$$= 680 - 55 - 15 - 1.875$$

$$= 608.125$$

Sale in September = 608 units

(d) **Newton's Method of Backward Differences** This method is most suitable when the figure to be interpolated is at the end of the tabulated values.

The formula under this method is

$$Y_x = Y_0 + X \Delta^1 0 + \frac{x(x-1)}{1 \times 2} \Delta^2 0 + \frac{x(x-1)(x-2)}{1 \times 2 \times 3} \Delta^3 0 \\ + \frac{x(x-1)(x-2)(x-3)}{1 \times 2 \times 3 \times 4} \Delta^4 0$$

Illustration 13.10

Calculate the value of Y , when x is 45 from the following data

X	10	20	30	40	50
Y	115	128	135	145	175

Solutions

X	Y	Differences			
		Δ_1	Δ_2	Δ_3	Δ_4
X_4 10	$115y_4$				
X_3 20	$128y_3$	$13\Delta^1 y_3$	$-6\Delta^2_2$		
X_2 30	$135y_2$	$7\Delta^1 y_2$	$3\Delta^2_1$	$9\Delta^3_1$	$8\Delta^4_0$
X_1 40	$145y_1$	$10\Delta^1 y_1$	$20\Delta^2_0$	$17\Delta^3_0$	
X_0 50	$175y_0$	$30\Delta^1 y_0$			

$$x = \frac{45 - 50}{10} = -5/10 = -0.5$$

$$Y_x = Y_0 + X \Delta^1_0 + \frac{x(x-1)}{2} \Delta^2_0 + \frac{x(x-1)(x-2)}{3 \times 2 \times 1} \Delta^3_0 + \frac{x(x-1)(x-2)(x-3)}{4 \times 3 \times 2 \times 1} \Delta^4_0 \\ = 175 + (-0.5 \times 30) + \frac{(-0.5)(-0.5-1)}{2} (20) + \frac{(-0.5)(-0.5-1)(-0.5-2)}{6} (17) \\ + \frac{(-0.5)(-0.5-1)(-0.5-2)(-0.5-3)}{24} (8) \\ = 175 - 15 + 7.5 - 5.31 + 2.19 \\ = 164.38$$

(e) **Newton's Divided Difference Method** When the value of independent variable 'x' advances by unequal intervals, this method can be adopted to find out the missing figures. To calculate the missing figure first of all the "Table of Divided Differences" should be prepared before applying the formula. Following is the method of preparing the table of divided differences.

Table of Divided Differences

X	Y	Divided Differences		
		Δ^1	Δ^2	Δ^3
x_0	y_0	$\frac{y_1 - y_0}{x_1 - x_0} = \Delta_0^1$	$\frac{\Delta_1^1 - \Delta_0^1}{x_1 - x_0} = \Delta_0^2$	$\frac{\Delta_2^2 - \Delta_0^2}{x_3 - x_0} = \Delta_0^3$
x_1	y_1	$\frac{y_2 - y_1}{x_2 - x_1} = \Delta_1^1$	$\frac{\Delta_2^2 - \Delta_0^2}{x_3 - x_1} = \Delta_1^2$	
x_2	y_2	$\frac{y_3 - y_2}{x_3 - x_2} = \Delta_2^1$		
x_3	y_3			

After calculating the differences, it should be applied in the following formula.

$$y_x = y_0 + (x - x_0) \Delta_0^1 + (x - x_0)(x - x_1) \Delta_0^2 + (x - x_0)(x - x_1)(x - x_2) \Delta_0^3 + \dots$$

Where $\Delta_0^1, \Delta_0^2, \Delta_0^3$ are the first, second and third leading divided differences respectively.

Illustration 13.11

Following are the details relating to the production in a factory for various months. Estimate the production for the month September.

Month	January	March	July	December
Production	300	330	470	560

Solutions

Month	X	Y	Δ_1	Δ_2	Δ_3
Jan	x_{01}	y_{0300}	$330 - 300/3 - 1 = 15 = \Delta_0^1$	$35 - 15/7 - 1 = 3.3 = \Delta_0^2$	$-1.89 - 3.3/12 - 1 = -0.47 = \Delta_0^3$
Mar	x_{13}	y_{1330}	$470 - 330/7 - 3 = 35 = \Delta_1^1$	$18 - 35/12 - 3 = -1.89 = \Delta_1^2$	
July	x_{27}	y_{2470}	$560 - 470/1 - 7 = 18 = \Delta_2^1$		
Dec	x_{312}	y_{3560}			

X is September (x) $x = 9$

$$\begin{aligned} y_x &= y_0 + (x - x_0) \Delta_0^1 + (x - x_0)(x - x_1) \Delta_0^2 + (x - x_0)(x - x_1)(x - x_2) \Delta_0^3 \\ &= 300 + (9 - 1)(15) + (9 - 1)(9 - 3)(3.3) + (9 - 1)(9 - 3)(9 - 7)(-0.47) \end{aligned}$$

$$\begin{aligned}
 &= 300 + (8 \times 15) + (8 \times 6 \times 3.3) + (8 \times 6 \times 2 (-0.47)) \\
 &= 300 + 120 + 158.4 - 45.12 \\
 y_x &= 533.28
 \end{aligned}$$

Lagrange Method A famous mathematician named Lagrange advocated this method for interpolation. It can be applied both for regular and irregular intervals of the independent variable x . This method can also be applied both for interpolation in the beginning and in the end.

The formula under this method is,

$$\begin{aligned}
 y_x = & y_0 \frac{(x - x_1)(x - x_2)(x - x_3)(x - x_n)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)(x_0 - x_n)} \\
 & + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)(x - x_n)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)(x_1 - x_n)} \\
 & + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)(x - x_n)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)(x_2 - x_n)} + y_3 \text{ etc.} \\
 & + y_n \frac{(x - x_0)(x - x_1)(x - x_2)(x - x_{n-1})}{(x_n - x_0)(x_n - x_1)(x_n - x_2)(x_n - x_{n-1})}
 \end{aligned}$$

Illustration 13.12

From the following data, find out the number of employed below 40 years of age in a village based on Lagrange method.

Age	No. of employed
Up to 15 years	85
Up to 25 years	104
Up to 35 years	119
Up to 45 years	125

Solutions

Calculation of the number of employed below 40 years old.

4	X_0	X_1	X_2	X_3
Y	85	104	119	125
	Y_0	Y_1	Y_2	Y_3

Calculation of number of employed below 40 years of age $x = 40$

$$y_x = y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)}$$

$$\begin{aligned}
& + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} \\
& = 85 \frac{(40 - 25)(40 - 35)(40 - 45)}{(15 - 25)(15 - 35)(15 - 45)} + 104 \frac{(40 - 15)(40 - 35)(40 - 45)}{(25 - 25)(25 - 35)(25 - 45)} \\
& \quad + 119 \frac{(40 - 15)(40 - 25)(40 - 45)}{(35 - 15)(35 - 25)(35 - 45)} + 125 \frac{(40 - 15)(40 - 25)(40 - 35)}{(45 - 15)(45 - 25)(45 - 35)} \\
& = 85 \left(\frac{-375}{-6000} \right) + 104 \left(\frac{-625}{-2000} \right) + 119 \left(\frac{-1875}{-2000} \right) + 125 \left(\frac{-1875}{-6000} \right) \\
& = 5.3125 - 32.5 + 111.5625 + 39.0625 \\
y_{40} & = 123.4375
\end{aligned}$$

Illustration 13.13

Calculate the number of students who obtained marks up to 45 from the following data on the basis of Lagrange's method.

Marks	0–10	10–20	20–30	30–40	40–50	50–60
No. of students	8	12	9	13	10	12

Solutions

Calculation of number of students who obtained up to 45 marks.

X	X ₀	X ₁	X ₂	X ₃	X ₄	X ₅
Y	8	12	9	13	10	12
y ₀	y ₁	y ₂	y ₃	y ₄	y ₅	

$$\begin{aligned}
Y &= y_0 \frac{(x - x_1)(x - x_2)(x - x_3)(x - x_4)(x - x_5)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)(x_0 - x_4)(x_0 - x_5)} \\
&+ y_1 \frac{(x - x_0)(x - x_2)(x - x_3)(x - x_4)(x - x_5)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)(x_1 - x_4)(x_1 - x_5)} \\
&+ y_2 \frac{(x - x_0)(x - x_1)(x - x_3)(x - x_4)(x - x_5)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)(x_2 - x_4)(x_2 - x_5)} \\
&+ y_3 \frac{(x - x_0)(x - x_1)(x - x_2)(x - x_4)(x - x_5)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)(x_3 - x_4)(x_3 - x_5)} \\
&+ y_4 \frac{(x - x_0)(x - x_1)(x - x_2)(x - x_3)(x - x_5)}{(x_4 - x_0)(x_4 - x_1)(x_4 - x_2)(x_4 - x_3)(x_4 - x_5)}
\end{aligned}$$

$$\begin{aligned}
& + y_5 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_5-x_0)(x_5-x_1)(x_5-x_2)(x_5-x_3)(x_5-x_4)} \\
& = 8 \frac{(45-20)(45-30)(45-40)(45-50)(45-60)}{(10-20)(10-30)(10-40)(10-50)(10-60)} \\
& + 20 \frac{(45-10)(45-30)(45-40)(45-50)(45-60)}{(20-10)(20-30)(20-40)(20-50)(20-60)} \\
& + 29 \frac{(45-10)(45-20)(45-40)(45-50)(45-60)}{(30-10)(30-20)(30-40)(30-50)(30-60)} \\
& + 42 \frac{(45-10)(45-20)(45-30)(45-50)(45-60)}{(40-10)(40-20)(40-30)(10-50)(10-60)} \\
& + 52 \frac{(45-10)(45-20)(45-30)(45-40)(45-60)}{(50-10)(50-20)(50-30)(50-50)(50-60)} \\
& + 64 \frac{(45-10)(45-20)(45-30)(45-40)(45-50)}{(60-10)(60-20)(60-30)(60-40)(60-50)} \\
& = 8 + \frac{(25 \times 15 \times 5 \times (-5) \times (-15))}{(-10 \times -20 \times -30 \times -40 \times -50)} + 20 \frac{(35 \times 15 \times 5 \times (-5) \times (-15))}{(-10 \times -10 \times -20 \times -30 \times -40)} \\
& + 29 \frac{(35 \times 15 \times 5 \times (-5) \times (-15))}{(20 \times 10 \times -10 \times -20 \times -30)} + 42 \frac{(35 \times 25 \times 15 \times (-5) \times (-15))}{(30 \times 20 \times 10 \times -10 \times -20)} \\
& + 52 \frac{(35 \times 25 \times 15 \times 5 \times (-15))}{(40 \times 30 \times 20 \times 10 \times -10)} + 64 \frac{(35 \times 25 \times 15 \times 5 \times (-5))}{(50 \times 40 \times 30 \times 20 \times 10)} \\
& = -0.09375 + 1.640625 - 7.9296875 + 34.453125 + 21.328125 - 1.75 \\
& = 47.648 = 47.65
\end{aligned}$$

Illustration 13.14

Find out the number of students who obtained marks from 40 to 55 through Lagrange method from the data given below.

Marks	0–30	30–40	40–50	50–60	60–70
No. of students	15	11	8	9	7

Solutions

X	X ₀	X ₁	X ₂	X ₃	X ₄
Y	15	26	34	43	50
	y ₀	y ₁	y ₂	y ₃	y ₄

$$\begin{aligned}
Y &= y_0 \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)(x_0-x_4)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_4)} \\
&\quad + y_2 \frac{(x-x_0)(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)(x_2-x_4)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)(x_3-x_4)} \\
&\quad + y_4 \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_0)(x_4-x_1)(x_4-x_2)(x_4-x_3)} \\
&= 15 \frac{(55-40)(55-50)(55-60)(55-70)}{(30-40)(30-50)(30-60)(30-70)} \\
&\quad + 26 \frac{(55-30)(55-50)(55-60)(55-70)}{(40-30)(40-50)(40-60)(40-70)} \\
&= 34 \frac{(55-30)(55-40)(55-60)(55-70)}{(50-30)(50-40)(50-60)(50-70)} \\
&\quad + 43 \frac{(55-30)(55-40)(55-50)(55-70)}{(60-30)(60-40)(60-50)(60-70)} \\
&\quad + 50 \frac{(55-30)(55-40)(55-50)(55-60)}{(70-30)(70-40)(70-50)(70-60)} \\
&= 15 \frac{(15)(5)(-5)(-15)}{(-10)(-20)(-30)(-40)} + 26 \frac{(25)(5)(-5)(-15)}{(10)(-10)(-20)(-30)} \\
&\quad + 34 \frac{(25)(15)(-5)(-15)}{(20)(10)(-10)(-20)} + 43 \frac{(25)(15)(5)(-15)}{(30)(20)(10)(-10)} \\
&\quad + 50 \frac{(25)(15)(-5)(-15)}{(40)(30)(20)(10)} \\
&= 15 \left(\frac{5625}{240000} \right) + 26 \left(\frac{9375}{-60000} \right) + 34 \left(\frac{28125}{40000} \right) + 43 \left(\frac{28125}{60000} \right) \\
&\quad + 50 \left(\frac{-9375}{240000} \right) \\
&= (0.3515625 - 4.0625 + 23.90625 + 20.15625 + 20.1525 - 1.953125) \\
&= 38.398
\end{aligned}$$

Parabolic Curve Method (Method of Simultaneous Equation) This method can be applied for interpolating any value of a dependent variable (y) for a given value of an independent variable (x). This method is based on the principle of simultaneous equation. The equation under parabolic curve method is

$$y = a + bx + cx^2 + dx^3 + \dots + nx^n$$

This equation should be continued for the number of variables known minus one, that is, x should be written for up to the power of x^{n-1} where n means number of known values.

For example, if number of known variables are five then x should continue up to the power of four, that is, x^4 . Hence, the parabolic curve equation should be

$$Y = a + bx + cx^2 + dx^3 + ex^4$$

Accordingly, the parabolic equation would be formed based on the number of known variables. Following are some of the equations based on the known variables.

Number of known variable	Equation
2	$y = a + bx$
3	$y = a + bx + cx^2$
4	$y = a + bx + cx^2 + dx^3$
5	$y = a + bx + cx^2 + dx^3 + ex^4$
n	$y = a + bx + cx^2 + dx^3 + ex^4 + \dots + nx^{n-1}$

The parabolic equation should be solved to get the value for a . The value of a is the value for unknown variable.

Illustration 13.15

Find out the sale for 2002 through parabolic curve method from the following data.

Year	2000	2001	2003	2004
Sales (in units)	175	190	220	260

Solutions

Since only 4 variables of y are known the parabolic curve for third order can be applied for interpolator the value for 2002.

The equation for the third order

$$y_0 = a + bx + cx^2 + dx^3$$

X	2000	2001	2002	2003	2004
Const	-2	-1	0	1	2
Y	175	190	y_0	220	260

$$y = a + bx + cx^2 + dx^3 \quad (1)$$

Put $x = -2$ $y = 175$ in Eq. (1)

$$175 = a - 2b + 4c - 8d \quad (2)$$

Put $x = -1$ $y = 190$ in Eq. (1)

$$190 = a - b + c - d \quad (3)$$

Put $x = 0$ $y = y_0$ in Eq. (1)

$$y_0 = a \quad (4)$$

Put $x = 1$ $y = 220$ in Eq. (1)

$$220 = a + b + c + d \quad (5)$$

Put $x = 2$ $y = 260$ in Eq. (1)

$$260 = a + 2b + 4c + 8d \quad (6)$$

Adding Eqs. 3 and 5

$$190 = a - b + c - d$$

$$\begin{array}{r} 220 = a + b + c + d \\ 410 = 2a + 2c \end{array} \quad (7)$$

Adding Eqs. 3 and 5

$$\begin{array}{r} 175 = a - 2b + 4c - 8d \\ 260 = a + 2b + 4c + 8d \\ \hline 435 = 2a + 8c \end{array} \quad (8)$$

Multiplying equation 7 by 4 and deducting Eq. 8 from it
equation $7 \times 4 \Rightarrow 1640 = 8a + 8c$

$$\begin{array}{r} 435 = 2a + 8c \\ (-) \quad (-) \quad (-) \\ \hline 1205 = 6a \\ a = 200.83 \end{array}$$

\therefore The sales for the year 2002 = 200.83
 $\cong 201$ units.

Extrapolation Extrapolation means estimation of the value for future period.

Illustration 13.16

Extrapolate the production for the year 2006 form the following data.

Year	2001	2002	2003	2004	2005
Sales (Rs in '000)	80	87	90	100	115

Solutions

Binomial expansion method can be adopted. The following is the equation for the five known variables.

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

X	2001	2002	2003	2004	2005
Y	80	87	90	100	115
y_0	y_1	y_2	y_3	y_4	

$$Y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

$$Y_5 - 5(115) + 10(100) - 10(90) + 5(87) - 80 = 0$$

$$Y_5 - 575 + 1000 - 900 + 435 - 80 = 0$$

$$y_5 - 120 = 0$$

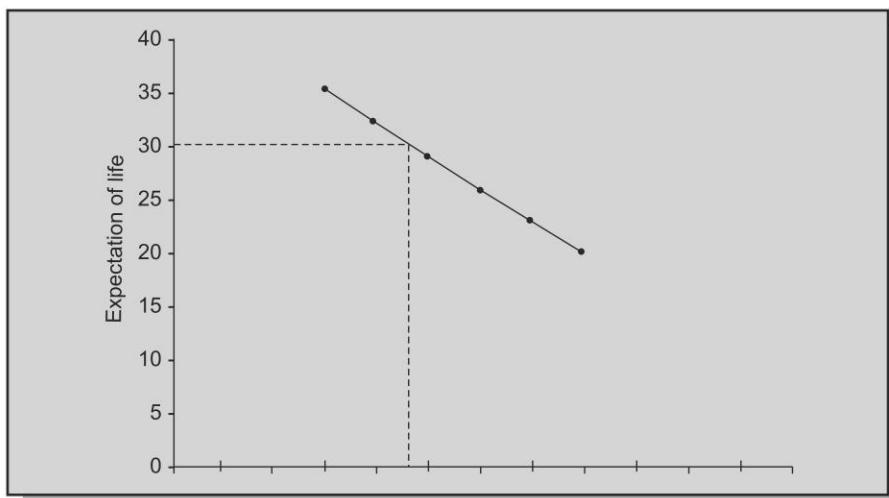
$$y_5 = 120$$

13.5 MISCELLANEOUS ILLUSTRATIONS

Illustration 13.17

From the following data, interpolate graphically the expectation of life at the age of 23:

Age	15	20	25	30	35	40
Expectation of life	35.4	32.2	29.1	26.0	23.1	20.4

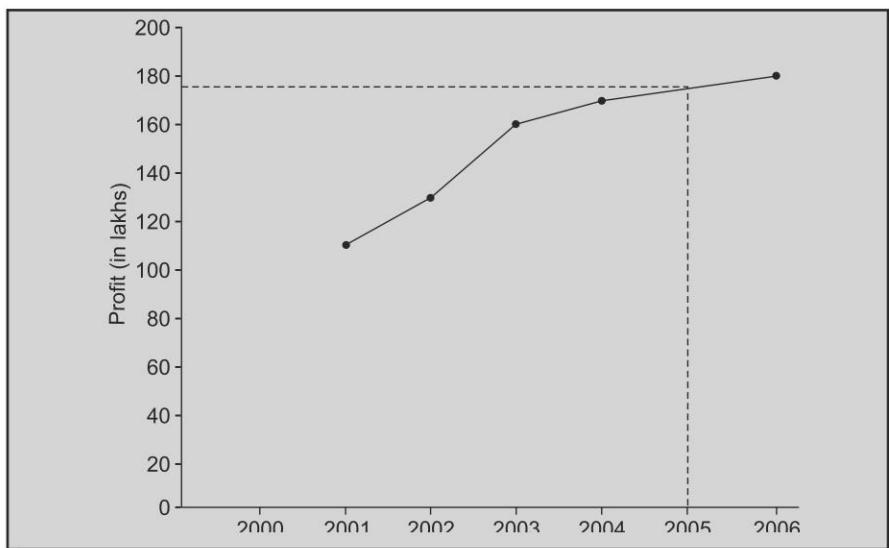
Solutions

Expectation of the life at the age of $x = 23$ is $y = 30.5$.

Illustration 13.18

From the following data find the value for 2005

Year	2001	2002	2003	2004	2005	2006
Profit in lakhs	110	130	160	170	?	180

Solutions

Profit for 2005 is 176 lakhs.

13.5.1 Binomial Expansion Method

Illustration 13.19

Estimate the missing term in the following table.

X	1	2	3	4	5
Y	3	9	27	?	243

Since $f(x)$ has 4 value

Shifting operator

$$\Delta^4(f(x)) = 0$$

$$\Delta = E - 1$$

$$(E - 1)^4(y_0) = 0$$

$$F(x) = y$$

$$(E^4 - 4E^3 + 6E^2 - 4E + 1)(y_0) = 0$$

$$E^4(y_0) - 4E^3(y_0) + 6E^2(y_0) - 4E(y_1) + y_0 = 0$$

$$Y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$$

$$243 - 4(y_3) + 6(27) - 4(9) + 3 = 0$$

$$243 - 4y_3 + 162 - 36 + 3 = 0$$

$$372 - 4y_3 = 0$$

$$372 = 4y_3$$

$$372/4 = y_3$$

$$y_3 = 93$$

Illustration 13.20

The following table gives the amount of cement (y) in thousands of tonnes manufactured in India. Find the missing figure.

Year	Output
2000	39
2001	85
2002	?
2003	151
2004	264
2005	388

X	2000	2001	2002	2003	2004	2005
Y	39	85	—	151	264	388
	y_0	y_1	y_2	y_3	y_4	y_5

Solutions

Since $f(x)$ has 5 values

$$\Delta^5 f(x_0) = 0 \quad (E - 1)^5 f(x_0) = 0$$

$$\begin{aligned}
 & (E^5 - 5E^4 + 10E^3 - 10E^2 + 5E - 1)y_0 = 0 \\
 & E^5(y_0) - 5E^4(y_0) + 10E^3(y_0) - 10E^2(y_0) + 5E(y_0) = 0 \\
 & y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0 \\
 & 388 - 5(264) + 10(151) - 10(y_2) + 5(85) - 39 = 0 \\
 & 388 - 1320 + 1510 - 10y_2 + 425 - 39 = 0 \\
 & 964 - 10y_2 = 0 \\
 & 964 = 10y_2 \\
 & 964/10 = y_2 \\
 & y_2 = 96.4
 \end{aligned}$$

Illustration 13.21

Use any suitable method of interpolation to find the value of y for $x = 3$ from the table below.

X	1	2	3	4	5
Y	220	236	?	256	362

Solutions

X	1	2	3	4	5
f(x)	220	236	?	256	362
	Y_0	Y_1	Y_2	Y_3	Y_4

Since $f(x)$ has 4 values $\Delta^4(f(x)) = 0$ $(E - 1)^4(f(x)) = 0$

$$\begin{aligned}
 & (E^4 - 4E^3 + 6E^2 - 4E + 1)y_0 = 0 \\
 & E^4(y_0) - 4E^3(y_0) + 6E^2(y_0) - 4E(y_0) + y_0 = 0 \\
 & y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0 \\
 & 362 - 4(256) + 6(y_2) - 4(236) + 220 = 0 \\
 & 362 - 1024 + 6y_2 - 944 + 220 = 0 \\
 & 6y_2 - 1386 = 0 \\
 & 6y_2 = 1386 \\
 & y_2 = 1386/6 \\
 & y_2 = 231
 \end{aligned}$$

Illustration 13.22

Find by interpolation the number of policies in force in 2004 from the following data.

Year	2001	2002	2003	2005	2006	2007
Number in force	66	73	81	99	106	112

Solutions

X	2001	2002	2003	2004	2005	2006	2007
f(x) Y	66	73	81	-	99	106	112
	y_0	y_1	y_2	y_3	y_4	y_5	y_6

Since $f(x)$ has 6 values

$$\begin{aligned}
 \Delta^6(f(x_0)) &= 0 \\
 (E-1)^6 f(x) &= 0 \\
 (E^6 - 6E^5 + 15E^4 - 20E^3 + 15E^2 - 6E + 1)y_0 &= 0 \\
 E^6(y_0) - 6E^5(y_0) + 15E^4(y_0) - 20E^3(y_0) + 15E^2(y_0) - 6E(y_0) + y_0 &= 0 \\
 y_6 - 6y_5 + 15y_4 - 20y_3 + 15y_2 - 6y_1 + y_0 &= 0 \\
 112 - 6(106) + 15(99) - 20y_3 + 15(81) - 6(73) + 66 &= 0 \\
 112 - 636 + 1485 - 20y_3 + 1215 - 438 + 66 &= 0 \\
 1804 - 20y_3 &= 0 \\
 1804 - 20y_3 &= 0 \\
 y_3 &= \frac{1804}{20} = 90.2 \\
 y_3 &= 90.2
 \end{aligned}$$

Illustration 13.23

The number of member of international statistical society are:

x	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
y	845	867	?	846	821	772	?	757	761	796

Make the best estimate you can of the members in 2000 and 2004.

Solutions

$F(x)$ has 8 values

$$\begin{aligned}
 \Delta^8(f(x_0)) &= 0 \\
 \Delta^8(y_0) &= 0 \\
 (E-1)^8(y_0) &= 0 \\
 (E^8 - 8E^7 + 28E^6 - 56E^5 + 70E^4 - 56E^3 + 28E^2 \\
 &\quad - 8E + 1)(y_0) = 0 \\
 (E^8 - (y_0) - 8E^7(y_0) + 28E^6(y_0) - 56E^5(y_0) + 70E^4(y_0) \\
 &\quad - 56E^3(y_0) + 28E^2(y_0) - 8E(y_0) + (y_0)) = 0 \\
 y_8 - 8y_7 + 28y_6 - 56y_5 + 70y_4 - 56y_3 + 28y_2 - 8y_1 + y_0 &= 0 \\
 761 - 8(757) + 28y_6 - 56(772) \\
 &\quad + 70(821) - 56(846) + 28y_2 - 8(867) + 845 = 0
 \end{aligned}$$

$$\begin{aligned}
 & 761 - 6056 + 28y_6 - 43232 + 57470 \\
 & - 47376 + 28y_2 - 6936 + 845 = 0 \\
 & 28y_6 + 28y_2 - 44524 = 0 \\
 & 28y_6 + 28y_2 = 44524 \quad (1) \\
 & \Delta^8(f(x_1)) = 0 \\
 & \Delta^8(y_1) = 0 \\
 & (E - 1)^8(y_1) = 0 \\
 & (E^8 - 8E^7 + 28E^6 - 56E^5 + 70E^4 - 56E^3 + 28E^2 - 8E + 1)y_1 = 0 \\
 & E^8(y_1) - 8E^7(y_1) + 28E^6(y_1) - 56E^5(y_1) + 70E^4(y_1) \\
 & - 56E^3(y_1) + 28E^2(y_1) - 8E(y_1) + y_1 = 0 \\
 \Rightarrow & y_9 - 8y_8 + 28y_6 - 56y_7 + 70y_5 - 56y_4 + 28y_3 - 8y_2 + y_1 = 0 \\
 \Rightarrow & 796 - 8(761) + 28(757) - 56y_6 + 54040 - 45976 \\
 & + 23688 - 8y_2 + 867 = 0 \\
 & 796 - 6088 + 21196 - 56y_6 + 54040 - 45976 \\
 & + 23688 - 8y_2 + 867 = 0 \\
 & - 56y_6 - 8y_2 + 48523 = 0 \\
 & - 56y_6 - 8y_2 = - 48523 \quad (2)
 \end{aligned}$$

Solve Eqs. (1) and (2)

$$\begin{aligned}
 \text{Equation 1} \times 2 & \quad 56y_6 + 56y_2 = 89048 \\
 \text{Equation 2} & \quad - 56y_6 - 8y_2 = - 48523 \\
 & \quad 48y_2 = 40525 \\
 & \quad y_2 = \frac{40525}{48} = 844.28
 \end{aligned}$$

Substitute y_2 in Eq. (2)

$$\begin{aligned}
 & - 56y_6 - 8y_2 = - 48523 \\
 & - 56y_6 - 8(844.28) = - 48523 \\
 & - 56y_6 - 6754.24 = - 48523 \\
 & - 56y_6 = - 48523 + 6754.24 \\
 & - 56y_6 = - 41768.76 \\
 & y_6 = 41768.76/56 \\
 & y_6 = 745.87
 \end{aligned}$$

Illustration 13.24

Given the following table, construct a difference table and from it estimate y when $x = 0.35$ by using Newton's backward interpolation formula.

x	0	0.1	0.2	0.3	0.4
y	1	1.095	1.179	1.251	1.310

Solutions

x	0	0.1	0.2	0.3	0.4
y	1	1.095	1.179	1.251	1.310
	y_0	y_1	y_2	y_3	y_4

Newton's backward formula

$$f(x) = Y_n + \frac{\mu}{1!} \Delta y_n + \frac{\mu(\mu+1)}{2!} \Delta^2 y_n + \dots + \frac{\mu(\mu+1)\dots\mu+(n-1)}{n!} \Delta^n y_n$$

$$\mu = \frac{x - x_n}{\lambda} = \text{Interval between 2 numbers} = 0.1 - 0 = 0.1$$

$$f(x) = y_4 + \frac{\mu}{1!} \Delta y_4 + \frac{\mu(\mu+1)}{2!} \Delta^2 y_4 + \frac{\mu(\mu+1)(\mu+2)}{3!} \Delta^3 y_4 \\ + \frac{\mu(\mu+1)(\mu+2)(\mu+3)}{4!} \Delta^4 y_4$$

$$x = 0.35$$

$$\mu = \frac{x - x_n}{\lambda} = \frac{0.35 - x_4}{0.1} = \frac{0.35 - 0.4}{0.1} = \frac{0.05}{0.1} = -0.5$$

$$\mu = -0.5$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
0	1	0.095			
0.1	1.095	0.084	-0.011	0.001	
0.2	1.179	0.072	-0.012	0.001	0
0.3	1.251	0.059	0.013		
0.4	1.310				

$$f(x) = 1.310 + \frac{(-0.5)}{1!} (0.059) + \frac{(-0.05)(-0.5+1)}{2} (-0.013) \\ + \frac{(-0.5)(-0.5+1)(0.5+2)}{3 \times 2 \times 1} (0.001) + 0 \\ = 1.310 - 0.0295 + 0.0004875 - 0.00032 = 1.2801$$

Illustration 13.25

The following are annual premium charged by the Life Insurance Corporation of India for a policy of Rs 1 lakh. Calculate the premium payable at the age of 26.

Age in Year	20	25	30	35	40
Premium in Rs.	230	260	300	350	420

Solutions

x	20	25	30	35	40
y	230	260	300	350	420

y_0 y_1 y_2 y_3 y_4

$$X = 26$$

Newton's forward formula

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)(\mu-2)}{2!} \Delta^2 y_0 + \dots$$

$$+ \frac{\mu(\mu-1)(\mu-2)\dots\mu-n+1}{n!} \Delta^n y_0$$

$$\mu = \frac{x - x_0}{\lambda}$$

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)(\mu-2)}{2!} \Delta^2 y_0 + \frac{\mu(\mu-1)(\mu-2)}{3!} \Delta^3 y_0$$

$$+ \frac{\mu(\mu-1)(\mu-2)(\mu-3)}{4!} \Delta^4 y_0$$

$$\mu = \frac{26 - 25}{5} = \frac{1}{5} = 0.2$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
20	230				
25	260	30			
30	300	40	10	0	
35	350	50	10	10	
40	420	70			

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)}{2!} \Delta^2 y_0 + \frac{\mu(\mu-1)(\mu-2)}{3!} \Delta^3 y_0$$

$$+ \frac{\mu(\mu-1)(\mu-2)(\mu-3)}{4!} \Delta^4 y_0$$

$$= 230 + 0.2(30) + \frac{0.2(0.2-1)}{2} (10) + \frac{0.2(0.2-1)(0.2-2)}{3 \times 2 \times 1} (0)$$

$$+ \frac{0.2(0.2-1)(0.2-2)(0.2-3)}{4 \times 3 \times 2 \times 1} (10)$$

$$= 230 + 6 - 0.8 + 0 + \frac{0.2(-0.8)(-1.8)(-2-8)}{24} (10)$$

$$= 230 + 6 - 0.8 + 0 - 0.336$$

$$= 234.864$$

Illustration 13.26

Expectation of life at different ages for males in India is shown in table below:

Age (year)	20	25	30	35	40
Expectation of life (year)	36	32	31	28	26

Use Newton's formula and estimate the expectation of life at the age of 33 years.

Solutions

x	20	25	30	35	40
y	36	32	31	28	26
	y_0	y_1	y_2	y_3	y_4

Newton's backward formula

$$f(x) = y_n + \frac{\mu}{1!} \Delta y_n + \frac{\mu(\mu+1)}{2!} \Delta^2 y_n + \dots + \frac{\mu(\mu+1)\dots(\mu+(n-1))}{n!} \Delta^n y_n$$

$$\therefore \mu = \frac{x - x_1}{\lambda}$$

$$f(x) = y_4 + \frac{\mu}{1!} \Delta y_4 + \frac{\mu(\mu+1)}{2!} \Delta^2 y_4 + \frac{\mu(\mu+1)(\mu+2)}{3!} \Delta^3 y_4 + \frac{\mu(\mu+1)(\mu+2)(\mu+3)}{4!} \Delta^4 y_4$$

$$\mu = \frac{x - x_1}{\lambda} = \frac{x - x_4}{\lambda} = \frac{33 - 40}{5} = -1.4$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
20	36	-4	3	-5	8
25	32	-1	-2	3	
30	31	-3	1		
35	28	-2			
40	26				

$$\begin{aligned}
 f(x) &= 26 + \frac{(-1.4)}{1!} (-2) + \frac{(-1.4)(-1.4+1)}{2} (1) + \frac{(-1.4)(-1.4+1)(-1.4+2)}{3 \times 2} (3) \\
 &\quad + \frac{(-1.4)(-1.4+1)(-1.4+2)(-1.4+3)}{4 \times 3 \times 2 \times 1} (8) \\
 &= 26 + 2.8 + \frac{(1.4)(-0.4)}{2} + \frac{(1.4)(-0.4)(0.6)}{6} (3) + \frac{(1.4)(-0.4)(0.6)(1.6)}{24} (8) \\
 &= 26 + 2.8 + 0.28 + 0.168 + 0.1792 \\
 &= 29.4272 = 29.4
 \end{aligned}$$

Illustration 13.27

Given the following pairs of corresponding values of x and y .

x	35	40	45	50	55
y	73	198	573	1198	1450

Estimate the value of y for $x = 37$.

Solutions

x	35	40	45	50	55
y	73	198	573	1198	1450
	y_0	y_1	y_2	y_3	y_4

Newton's forward formula

$$y - y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)}{2!} \Delta^2 y_0 + \frac{\mu(\mu-1)(\mu-2)}{3!} \Delta^3 y_0 + \dots \\ + \frac{\mu(\mu-1)\dots(\mu-(n-1))}{n!} \Delta^n y_0 + \dots$$

$$\mu = \frac{x - x_1}{\lambda}$$

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)}{2!} \Delta^2 y_0 + \frac{\mu(\mu-1)(\mu-2)}{3!} \Delta^3 y_0 \\ + \frac{\mu(\mu-1)(\mu-2)(\mu-3)}{4!} \Delta^4 y_0$$

$$\mu = \frac{x - x_1}{\lambda} \quad x = 37; x_1 = 40; n = 5$$

$$\mu = \frac{37 - 40}{5} = -0.6$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
35	73		125		
40	198		375	250	
45	573		625	250	0
50	1198		252	-373	-623
55	1450				

$$y = 73 + \frac{(0.6)}{1}(125) + \frac{(-0.6)(-0.6-1)}{2}(250) + \frac{(-0.6)(-0.6-1)(-0.6-2)}{3 \times 2 \times 1}(0) \\ + \frac{(-0.6)(-0.6-1)(-0.6-2)(-0.6-3)}{4 \times 3 \times 2 \times 1}(-623) \\ = 73 - 75 + 120 + 0 - 233.2512 \\ = -115.2512$$

Illustration 13.28

Below are given the wages earned by workers per month in a certain factory. Calculate the numbers of workers earning more than Rs 750 per week.

Weekly wages	No. of workers
Upto Rs 500	50
Upto Rs 600	150
Upto Rs 700	300
Upto Rs 800	500
Upto Rs 900	700
Upto Rs 1000	800

Solutions

x	x_0 500	x_1 600	x_2 700	x_3 800	x_4 900	x_5 1000
y	50 y_0	150 y_1	300 y_2	500 y_3	700 y_4	800 y_5

Newton's forward formula,

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \dots + \frac{\mu(\mu-1)\cdots(\mu-(n-1))}{n!} \Delta^n y_0$$

where $\mu = \frac{x - x_1}{\lambda}$

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)}{2!} \Delta^2 y_0 + \frac{\mu(\mu-1)(\mu-2)}{3!} \Delta^3 y_0 + \frac{\mu(\mu-1)(\mu-2)(\mu-3)}{4!} \Delta^4 y_0 + \frac{\mu(\mu-1)(\mu-2)(\mu-3)(\mu-4)}{5!} \Delta^5 y_0$$

$$x = 750 \quad x_1 = 600$$

$$\mu = \frac{750 - 600}{100} = 150/100 = 1.5$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
500	50	100	50			
600	150	150	50	0	-50	
700	300	200	0	-50	-50	0
800	500	200	-100	-100		
900	700	100				
1000	800					

$$\begin{aligned}
 y &= 50 + \frac{1.5}{1!}(100) + \frac{1.5(1.5-1)}{2!}(50) + \frac{1.5(1.5-1)(1.5-2)}{3 \times 2 \times 1}(0) \\
 &\quad + \frac{1.5(1.5-1)(1.5-2)(1.5-3)}{4 \times 3 \times 2 \times 1}(-50) + \frac{1.5(1.5-1)(1.5-2)(1.5-3)(1.5-4)}{5 \times 4 \times 3 \times 2 \times 1}(0) \\
 &= 50 + 150 + 18.75 + 0 - 1.171875 + 0 \\
 &= 217.57813 = 217.58
 \end{aligned}$$

Illustration 13.29

Find out the number of students securing second division in the university examination from the following figures.

Marks obtained (out of 100)	0–20	20–40	40–60	60–80	80–100
No. of students	5	26	85	54	30

(48% and above but less than 60% marks makes second division)

Solutions

Newton's forward formula

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)}{2!} \Delta^2 y_0 + \dots + \frac{\mu(\mu-1)\dots(\mu-(n-1))}{n!} \Delta^n y_0$$

$$\text{where } \mu = \frac{x - x_0}{n}$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$
20	5				
40	31	26	59		
60	116	85	-31	-90	
80	170	54	-24	7	97
100	200	30			

$$\mu = \frac{x - x_0}{n} = \frac{48 - 20}{20} = 28/20 = 1.4$$

$$\begin{aligned}
 y &= 5 + \frac{1.4}{1!}(26) + \frac{1.4(1.4-1)}{2!}(59) + \frac{1.4(1.4-1)(1.4-2)}{3 \times 2 \times 1}(-90) \\
 &\quad + \frac{1.4(1.4-1)(1.4-2)(1.4-3)}{4 \times 3 \times 2 \times 1}(97) \\
 &= 5 + 36.4 + 16.52 + 5.04 + 2.17
 \end{aligned}$$

$$= 65.13$$

$$= 65$$

Number of students getting less than 60% = 116 (from the table)

Number of students getting less than 48% = 65

Number of students getting second division

$$\text{above } 48\% \text{ and less than } 60\% = 116 - 65 = 51$$

Illustration 13.30

From the data given below, estimate the number of persons living between the age of 35 and 42.

Age (in years)	20	30	40	50
No. of persons	1026	878	692	486

Solutions

x	20	30	40	50
y	1026	878	692	486

$$x = 35$$

Newton's forward formula

$$y = y_0 + \frac{\mu}{1!} \Delta y_0 + \frac{\mu(\mu-1)}{2!} \Delta^2 y_0 + \dots + \frac{\mu(\mu-1)\cdots(\mu-(n-1))}{n!} \Delta^n y_0$$

$$\mu = \frac{x - x_0}{n} = \frac{35 - 20}{10} = 15/10 = 1.5$$

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
20	1026			
30	878	-148	-38	
40	692	-186	-20	18
50	486	-206		

$$y = 1026 + \frac{1.5}{1!} (-148) + \frac{1.5(1.5-1)}{2!} (-38) \frac{1.5(1.5-1)(1.5-2)}{3 \times 2 \times 1} (18)$$

$$= 1026 - 222 - 14.25 - 1.125 = 788.625$$

$$x = 42$$

Newton's backward formula

$$y = y_n + \frac{\mu}{1!} \Delta y_n + \frac{\mu(\mu+1)}{2!} \Delta^2 y_n + \dots + \frac{\mu(\mu+1)\cdots(\mu(n-1))}{n!} \Delta^n y_n$$

$$\mu = \frac{x - x_n}{n} = \frac{42 - 50}{10} = \frac{-8}{10} = -0.8$$

$$\begin{aligned}
 y &= y_3 + \mu/1! \Delta y_3 + \frac{\mu(\mu+1)}{2!} \Delta^2 y_3 + \frac{\mu(\mu+1)(\mu-2)}{3!} \Delta^3 y_3 \\
 &= 486 + \frac{(-0.8)}{1!} (-206) + \frac{(-0.8)(-0.8+1)}{2} (-20) \\
 &\quad + \frac{(-0.8)(-0.8+1)(-0.8+2)}{3 \times 2 \times 1} (18) \\
 &= 486 + 164.8 + 1.6 - 0.576 \\
 &= 651.824
 \end{aligned}$$

Number of person living between 35 and 42 is

$$\begin{aligned}
 &= 788.625 - 651.824 \\
 &= 136.801 \\
 &= 136.8
 \end{aligned}$$

Illustration 13.31

The following table gives the normal weight of a baby during the first six months of life.

Age in months	Weight in lbs
0	10
2	14
3	16
5	20

Estimate the weight of a baby at the age of 4 months.

Solutions

x	x_0	x_1	x_2	x_3
y	10	14	16	20
	y_0	y_1	y_2	y_3

$$x = 4$$

Lagrange's formula

$$\begin{aligned}
 y &= y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\
 &\quad + y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\
 &= 10 + \frac{(4-0)(4-2)(4-5)}{(0-0)(0-2)(0-5)} + 14 \frac{(4-0)(4-3)(4-5)}{(2-0)(2-3)(2-5)}
 \end{aligned}$$

$$\begin{aligned}
 & + 16 \frac{(4-0)(4-2)(4-5)}{(3-0)(3-2)(3-5)} + 20 \frac{(4-0)(4-2)(4-3)}{(5-0)(5-2)(5-3)} \\
 & = 8 - 9.3334 + 21.3334 + 5.3334 \\
 & = 25.33165 = 25 \text{ lbs}
 \end{aligned}$$

Illustration 13.32

The values of x and y are given below:

x	5	6	9	11
y	12	10	14	16

Find the value of y when $x = 10$ by using Lagrange's method.

Solutions

x	x_0 5	x_1 6	x_2 9	x_3 11
y	12 y_0	10 y_1	14 y_2	16 y_3

$$x = 10$$

Lagrange's formula

$$\begin{aligned}
 y &= y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\
 &\quad + y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\
 y &= 12 \frac{(10-6)(10-9)(10-11)}{(5-6)(5-9)(5-11)} + 10 \frac{(10-5)(10-9)(10-11)}{(6-5)(6-9)(6-11)} \\
 &\quad + 14 \frac{(10-5)(10-6)(10-11)}{(9-5)(9-6)(9-11)} + 16 \frac{(10-5)(10-6)(10-9)}{(11-5)(11-6)(11-9)} \\
 &= 12 \left(\frac{-4}{-24} \right) + 10 \left(\frac{-5}{-15} \right) + 14 \left(\frac{-20}{-24} \right) + 16 \left(\frac{20}{60} \right) \\
 &= 1.9999 - 3.3334 + 11.6667 + 5.3334 \\
 &= 15.6666 \\
 &= 15.67
 \end{aligned}$$

Illustration 13.33

Given the following table, find \log_{656} .

No.	654	658	659	661
Log	2.8156	2.8182	2.8189	2.8202

Solutions

$$x = 656$$

X	x_0 654	x_1 658	x_2 659	x_3 661
Y	2.8156 y_0	2.8182 y_1	2.8189 y_2	2.8202 y_3

Using Lagrange's formula

$$\begin{aligned}
 y &= y_0 \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} + y_1 \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} \\
 &\quad + y_2 \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + y_3 \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} \\
 &= 2.8156 \frac{(656 - 658)(656 - 659)(656 - 661)}{(654 - 658)(654 - 659)(654 - 661)} \\
 &\quad + 2.8182 \frac{(656 - 654)(656 - 659)(656 - 661)}{(658 - 654)(658 - 659)(658 - 661)} \\
 &\quad + 2.8189 \frac{(656 - 654)(656 - 658)(656 - 661)}{(659 - 654)(659 - 658)(659 - 661)} \\
 &\quad + 2.8202 \frac{(656 - 654)(656 - 658)(656 - 659)}{(661 - 654)(661 - 658)(661 - 659)} \\
 &\quad + 2.8156 \frac{(-2)(-3)(-5)}{(-4)(-5)(-7)} + 2.8182 \frac{2(-3)(-5)}{4(-1)(-3)} \\
 &\quad + 2.8189 \frac{2(-2)(-5)}{5(1)(-2)} + 2.8202 \frac{2(-2)(-3)}{7(3)(2)} \\
 &= 0.6033 + 7.0455 - 5.6378 + 0.8058 \\
 &= 2.8168
 \end{aligned}$$

Illustration 13.34

Use Lagrange interpolation formula to find y when $x = 0$, given the following table.

x	-1	-2	2	4
y	-1	-9	11	69

Solutions

X	x_0 -1	x_1 -2	x_2 2	x_3 4
y	-1 y_0	-9 y_1	11 y_2	69 y_3

$$x = 0$$

Using Lagrange's formula

$$\begin{aligned}
 y &= y_0 \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} + y_1 \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} \\
 &\quad + y_2 \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} + y_3 \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} \\
 &= (-1) \frac{(0+2)(0-2)(0-4)}{(-1+2)(-1-2)(-1-4)} + (-9) \frac{(0+1)(0+2)(0-4)}{(-2+1)(-2-2)(-2-4)} \\
 &\quad + 11 \frac{(0+1)(0+2)(0-4)}{(2+1)(2-2)(2-4)} + 69 \frac{(0+1)(0+2)(0-2)}{(4+1)(4+2)(4-2)} \\
 &= (-1) \frac{2(-2)(-4)}{1(-3)(-5)} + (-9) \frac{1(-2)(-4)}{-1(-4)(-6)} + 11 \frac{1(2)(-4)}{3(4)(-2)} + 69 \frac{1(2)(-2)}{5(6)(2)} \\
 &= -1.07 + 3.00 + 3.67 - 4.60 \\
 &= 1
 \end{aligned}$$

Illustration 13.35

The pressure of wind in pounds per square foot, corresponding to the velocity in miles per hour has been determined by experiment to be approximately as follows:

Velocity	20	30	40	50
Pressure	4	6	9	12

Estimate the pressure for a velocity of 35 miles per hour.

Solutions

Velocity	20	30	35	40	50
Deviation from 35 x	-3	-1	0	1	3
Pressure y	4	6	?	9	12

$$y = a + bx + cx^2 + dx^3 \quad (1)$$

Put $x = -3$ $y = 2$

Substitute x and y value in Eq. (1)

$$2 = a - 3b + 9c - 27d \quad (2)$$

Put $x = -1$ $y = 4.4$

$$\text{Equation 1} \Rightarrow 4.4 = a - b + c - d \quad (3)$$

Put $x = 0$; $y = -$

$$\text{Equation 1} \Rightarrow y = a \quad (4)$$

Put $x = 1$; $y = 7.9$

$$\text{Equation 1} \Rightarrow 7.9 = a + b + c + d \quad (5)$$

Put $x = 3$ $y = 13.2$

$$\text{Equation 1} \Rightarrow 13.2 = a + 3b + 9c + 27d \quad (6)$$

Solve Eq. 2 and 6

$$\begin{aligned} 2 &= a - 3b + 9c - 27d \\ 13.2 &= a + 3b + 9c + 27d \\ 15.2 &= 2a + 18c \end{aligned} \quad (7)$$

Solve Eq. 3 and 5

$$\begin{aligned} 4.4 &= a - b + c - d \\ 7.9 &= a + b + c + d \\ 12.3 &= 2a + 2c \end{aligned} \quad (8)$$

No. of terms given = 5

$$\begin{aligned} x &= \frac{20 - 35}{5} = \frac{-15}{5} = -3 \\ &= \frac{30 - 35}{5} = \frac{-5}{5} = -1 \\ &= \frac{35 - 35}{5} = \frac{0}{5} = 0 \\ &= \frac{40 - 35}{5} = \frac{5}{5} = 1 \\ &= \frac{50 - 35}{5} = \frac{15}{5} = 3 \end{aligned}$$

Solve Eq. 7 and 8

$$\text{Equation} \Rightarrow 2a + 18c = 15.2$$

$$\text{Equation} \Rightarrow 2a + 2c = 12.3$$

Multiply Eq. 8 by 9

$$\begin{aligned} 2a + 18c &= 15.2 \\ 18a + 18c &= 110.7 \\ -16a &= -95.5 \\ a &= \frac{95.5}{16} \\ a &= 5.96875 \\ &\approx 5.97 \end{aligned}$$

\therefore Pressure for a velocity of 35 miles = 5.97

Extrapolation

Illustration 13.36

Extrapolate the production for the year 2000 from the following data by Binomial expansion method.

Year	1995	1996	1997	1998	1999
Sales (Rs. in '000)	70	77	80	90	105

Solutions

x	x_0 1995	x_1 1996	x_2 1997	x_3 1998	x_4 1999
y	70	77	80	90	105
	y_0	y_1	y_2	y_3	y_4

∴ Five values are known; we use the formula $(a - b)^5$

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

$$y_5 - 5(105) + 10(90) - 10(80) + 5(77) - 70 = 0$$

$$y_5 - 525 + 900 - 800 + 385 - 70 = 0$$

$$y_5 - 110 = 0$$

$$\begin{aligned} y_5 &= 110 \\ &= 1,10,000 \end{aligned}$$

Illustration 13.37

Calculate the value of sales for 1995 from the following data.

Year	1993	1994	1995	1996	1997	1998
Sales (Rs in '000)	180	195	?	203	226	250

Solutions

x	1993	1994	1995	1996	1997	1998
y	180	195	—	203	226	250
	y_0	y_1	y_2	y_3	y_4	y_5

The equation for five known value we use $(a - b)^5$ formula

$$y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

$$250 - 5(226) + 10(203) - 10y_2 + 5(195) - 180 = 0$$

$$250 - 1130 + 2030 - 10y_2 + 975 - 180 = 0$$

$$1945 - 10y_2 = 0$$

$$1945 = 10y_2$$

$$y_2 = 1945/10$$

$$\begin{aligned} y_2 &= 194.5 \\ &= 19,45,000 \end{aligned}$$

Illustration 13.38

Estimate the production for 1997 and 2000 from the following data:

Year	1995	1996	1997	1998	1999	2000	2001
Production (in tonnes)	150	165	?	180	215	?	240

Solutions

x	x_0 1995	x_1 1996	x_2 1997	x_3 1998	x_4 1999	x_5 2000	x_6 2001
y	150	165	—	180	215	—	240
	y_0	y_1	y_2	y_3	y_4	y_5	y_6

$$\begin{aligned}
 & y_6 - 6y_5 + 15y_4 - 20y_3 + 15y_2 - 6y_1 + y_0 = 0 \\
 & 240 - 6(y_5) + 15(215) - 20(180) + 15y_2 - 6(165) + 150 = 0 \\
 & 240 - 6y_5 + 3225 - 3600 + 15y_2 - 990 + 150 = 0 \\
 & 15y_2 - 6y_5 - 975 = 0 \\
 & 15y_2 - 6y_5 = 975 \quad (1) \\
 & \Delta^5(f(x_0)) = 0 \\
 & (E - 1)^5(y_0) = 0 \\
 & y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0 \\
 & y_5 - 5(215) + 10(180) - 10y_2 + 5(165) - 150 = 0 \\
 & 5 - 1075 + 1800 - 10y_2 + 825 - 150 = 0 \\
 & y_5 - 10y_2 + 1400 = 0 \\
 & y_5 - 10y_2 = 1400 \quad (2)
 \end{aligned}$$

Solution Eq. (1) and (2)

$$\begin{aligned}
 & 15y_2 - 6y_5 = 975 \\
 \text{For example } 2 \times 6 & \Rightarrow -60y_2 + 6y_5 = -8400 \\
 & -45y_2 = -7425 \\
 & y_2 = \frac{7425}{45} \\
 & y_2 = 165
 \end{aligned}$$

Substitute y_2 value in Eq. (1)

$$\begin{aligned}
 & 15y_2 - 6y_5 = 975 \\
 & 15(165) - 6y_5 = 975 \\
 & 2475 - 6y_5 = 975 \\
 & 2475 - 975 = 6y_5 \\
 & \frac{1500}{6} = y_5 \\
 & y_5 = 250
 \end{aligned}$$

Hence, the estimated production for 1997 is 165 tonnes and for 200 is 250 tonnes.

SUMMARY

Interpolation

Reading a value which lies between two extreme points.

Extrapolation

Reading a value that lies outside to two extreme points. Estimation of the value for future period.

Methods of Interpolation and Extrapolation

Graphic Method

Algebraic Method

- Binomial Expansion Method
- Lagrange's Method
- Parabolic curve Method

Newton's Method

- Newton's Gauss (forward) Method
- Newton's Gauss (Backward) Method
- Newton's Method of Backward differences
- Newton's Divided Difference Method

Newton-Gauss (Forward) Method

It can be applied only when the independent variable (x) increases by equal intervals and the variable to be interpolated lies in the middle of the series.

Newton's Gauss (Backward) Method

It can be applied only when the variable to be interpolated lies at the end of the series and the differences between all the two adjoining variables x are the same.

Newton's Divided differences Method

When the value of independent variable X advances by unequal intervals, this method can be adopted to find out the missing figures.

Newton's Method of Advancing Differences:

It is applied only when the differences in the independent variables are same.

Lagrange Method

It can be applied both for regular and irregular intervals of the independent variable X . It can also be applied both for interpolation in the beginning and in the end.

Parabolic Curve Method (or) Method of Simultaneous Equation

It can be applied for interpolating any value of a dependent variable (y) for a given value of an independent variable (x).

FORMULAE

1. Binomial Expansion Method Condition:

- (i) The x variable advanced by equal intervals, i.e. 5, 10, 15 etc.
- (ii) The value of X for which Y is to be interpolated is one of the class limits of X series. When the method is applied, the terminal $(Y-1)^n$ must be expanded and Y is made equal to zero.

$$(Y-1)^n = y^n - ny^{n-1} + \frac{n(n-1)}{2}y^{n-2} - \frac{n(n-1)(n-2)}{3}y^{n-3} + \dots = 0$$

n = the number of known values

In other words, we have the following result :

No. of equations for determining the Known values Unknown values

$$2 \text{ or } \Delta_0^2 \quad Y_2 - 2y_1 + y_0 = 0$$

$$3 \text{ or } \Delta_0^3 \quad Y_3 - 3y_2 + 3y_1 - y_0 = 0$$

$$4 \text{ or } \Delta_0^4 \quad Y_4 - 4y_3 + 6y_2 - 4y_1 + y_0 = 0$$

$$5 \text{ or } \Delta_0^5 \quad Y_5 - 5y_4 + 10y_3 - 10y_2 + 5y_1 - y_0 = 0$$

$$6 \text{ or } \Delta_0^6 \quad Y_6 - 6y_5 + 15y_4 - 20y_3 + 15y_2 - 6y_1 + y_0 = 0$$

$$7 \text{ or } \Delta_0^7 \quad Y_7 - 7y_6 + 21y_5 - 35y_4 + 35y_3 - 21y_2 + 7y_1 + y_0 = 0$$

2. Newton's Method Condition:

The x variable advanced by equal intervals : That is, 5, 10, 15 etc.

$$\begin{aligned} Y_x = Y_0 + x\Delta_0^1 \frac{x(x-1)}{2!} \Delta_0^2 + \frac{x(x-1)(x-2)}{3!} \Delta_0^3 \\ + \frac{x(x-1)(x-2)(x-3)}{4!} \Delta_0^4 \end{aligned}$$

Y_x = The figure to be interpolated

Y_0 = the value of Y at origin

$X = \frac{\text{Year of interpolation} - \text{The value at origin}}{\text{Difference between two adjoining years}}$

Δ = are the differences between various values

3. Lagrange's method conditions

When X series advanced by unequal intervals

$$\begin{aligned} y_x &= \frac{(x-x_1)(x-x_2)(x-x_3)\dots(x-x_n)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)\dots(x_0-x_n)} \\ + y_1 &= \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_n)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_n)} \\ + y_2 &= \frac{(x-x_0)(x-x_1)(x-x_3)(x-x_n)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)(x_2-x_n)} + y_3 \end{aligned}$$

$$+ \frac{y_n(x - x_0)(x - x_1)(x - x_2) \dots (x - x_{n-1})}{(x_n - x_0)(x_n - x_1)(x_n - x_2) \dots (x - x_{n-1})}$$

Y_1 = The figure to be interpolated

X_0, X_1, X_2 = The given values

Y_0, Y_1, Y_2 = Corresponding values of y variable

Y_n = The tongue to be interpolated

4. Parabolic Method :

$$Y = a + bx + cx^2 + dx^3 + \dots + nx^n$$

a, b, c, d are constants

EXERCISES

(a) Choose the best option:

1. Binomial expansion method is applicable when:
 - (a) The x variable advances by unequal intervals.
 - (b) The value of X for which Y is to be interpolated is one of the limits of X series.
 - (c) X variable advances by equal intervals and value of X for which Y is to be interpolated is one of the class limits of X series.
 - (d) X variable advances by unequal intervals and the value of X for which Y is to be interpolated is one of the class limits of X series.
 - (e) None of these.
2. The Newton's method is applicable when:
 - (a) The X variable advances by unequal intervals.
 - (b) The values of X for which Y is to be interpolated is one of the class limits of X series.
 - (c) It does not really matter what happens to X and Y variables.
 - (d) X Variable increases by equal intervals and it is not necessary that the value of X for which Y is to be interpolated is one of the class limits of X series.
3. Where X series advance by unequal intervals, the appropriate formula of Interpolation to be used is
 - (a) Binomial expansion method
 - (b) Newton's method
 - (c) Lagrange's method
 - (d) all of these
 - (e) none of these.

4. The extrapolated value of a certain time period shall be
 - (a) equal to actual value
 - (b) less than the actual value
 - (c) more than the actual value
 - (d) most likely estimate under certain assumptions
 - (e) none of these.
5. Interpolation relates to
 - (a) Future
 - (b) Present
 - (c) Past

Answer

1. (c) 2. (d) 3. (c) 4. (d) 5. (c)

(b) Fill in the blanks:

1. The technique of estimating a past figure is termed as _____.
2. _____ give us forecast for the future.
3. The binomial expansion method is appropriate when the independent variable X advances by _____.
4. $(y - 1)^3$ _____.
5. When the figure to be interpolated falls at the end of the series appropriate formula of interpolation is _____.
6. The Lagrange's method is appropriate in those cases when x variable advances by _____.
7. _____ method is also known as method of simultaneous equations.
8. The simplest method of interpolation is _____.

Answers

- | | |
|-------------------------------------|------------------------------|
| 1. Interpolation | 2. Extrapolation |
| 3. Equal intervals | 4. $y_2 - 3y_2 + 3y_2 - y_0$ |
| 5. Newton's Gauss (Backward)formula | |
| 6. Unequal interval | 7. Parabolic curve method |
| 8. Graphic method. | |

(c) Theoretical Problems

1. Define interpolation state the benefits of interpolation.
2. What is interpolation? State briefly the different methods of interpolation. **(B. Com., BDU, MKU, MSU)**
3. Distinguish between interpolation and extrapolation.

4. Define interpolation? State the procedure for binomial expansion method of interpolation. (B. Com., CHU, BDU, MKU, MSU)
5. What do you understand by a divided difference?
6. Define interpolation and state the procedure for Lagrange's method of interpolation.
7. Explain Newton's method of interpolation and its procedure. (B. Com., CHU, BU, MKU)
8. Explain the significance and utility of interpolation and extrapolation.
9. Define interpolation and explain the procedure for the parabolic curve method of interpolation. (B. Com., BDU, MSU, BU)

(d) Practical Problems

10. Find out the sales for 2002 from the following data.

Year	2000	2001	2003	2004	2005
Sales	36	47	59	68	85
Answer 53.4					(B. Com., BDU, MKU, MSU)

11. Extrapolate the profit for 2006 from the following data:

Year	2001	2002	2003	2004	2005
Profit (Rs in '000)	8	15	25	40	56
Answer 63					(B. Com., BU, CHU)

12. From the following data interpolate the number of students who secured more than 59 marks by using Lagrange's method.

Marks (more than):	25	36	45	55	70
No. of Students :	65	63	40	18	7

(B. Com., BDU, MKU, MSU)

13. Using the Lagrange's formula of interpolation, find from the data the number of workers earning between Rs 3000 and Rs 4000.

Earning in ('00):	15–20	20–30	30–45	45–55	55–70
No. of workers :	73	97	110	180	140
Answer 53					(B. Com., MKU, MSU)

14. Estimate the value of y when $x = 125$.

X	50	100	150	200	250
Y	120	105	80	60	30

Answer 92.11

15. The value of x and y are given below. Find out the value of y when $x = 10$.

<i>X</i>	5	6	9	11
<i>Y</i>	12	13	14	16

Find out the value of *y* when *x* = 10. (B. Com., BDU, BU, CHU)

16. From the following data, estimate the number of employees in a firm who earn more than Rs 1,200 but not more than Rs 2,400.

Income more than Rs:	500	1,000	1,500	2,000	2,500	3000
No. of persons:	600	550	425	275	100	25

17. From the following data of the population of a city in lakhs, find out population for 1995:

Year:	1970	1975	1980	1985	1990	1995	2000
Population:	200	220	260	180	350	?	430

(B. Com., BDU, MKU, MSU)

18. Extrapolate the business done in 1987 from the following data:

Year:	1998	1989	1990	1991	1992
Business done (Rsm. lakhs):	150	235	364	525	780

19. From the following data, obtain the value of *Y* when *x* = 9 using Newton's Forward Difference Interpolation formula:

<i>X</i> :	3	7	11	15	19
<i>Y</i> :	42	43	47	53	60

(B. Com., BDU, MKU, CHU)

20. Calculate *Y* when *X* is 12 from the following data by interpolation:

<i>X</i> :	10	20	30	40	50
<i>Y</i> :	2.3	3.0	3.4	3.7	3.9

21. The wages of workers in a town are stated below:

Wages less than (Rs)	30	40	50	60	70
No. of workers	45	66	85	100	120

Estimate the no. of workers earning Rs 45 and Rs 65.

Answer 72 and 108

22. Estimate the value of *Y* when *x* = 125.

<i>X</i>	50	100	150	200	250
<i>Y</i>	120	105	80	60	30

Answer 92.11

23. Calculate the missing value from the following:

606 Business Statistics

Year	1996	1997	1998	1999	2000
Profit (Rs in '000)	24	31	?	47	57
Answer 38.5					(B. Com., BDU, MKU, MSU)
24. Interpolate the number of students who scored below 65 marks through Lagrange method from the following data.					
Marks less than	20	40	60	80	100
No. of students	14	39	48	62	70
Answer 51					(B. Com., BDU, MKU, MSU, CHU)

Appendix

- (i) Logarithms Tables
- (ii) Antilogarithms Tables

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9		1	2	3	4	5	6	7	8	9	Mean Differences
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37		
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34		
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31		
13	1139	1173	1203	1239	1271	1303	1335	1367	1399	1430	3	6	10	18	16	19	23	26	29		
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27		
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25		
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24		
17	2304	2330	2355	2380	2405	2430	2455	2480	2604	2629	2	5	7	10	12	16	17	20	22		
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21		
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20		
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19		

Contd.

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3484	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3892	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4189	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5085	5079	5092	6105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5186	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	6729	5740	5762	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6084	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10

Contd.

	Mean Differences									
	0	1	2	3	4	5	6	7	8	9
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6245	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6593	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	5575	6684	6693	6702	6712
47	6721	5730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7651
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846

Contd.

	Mean Differences									
0	1	2	3	4	5	6	7	8	9	
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8489	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079

Contd.

	Mean Differences									
0	1	2	3	4	5	6	7	8	9	
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9142	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9340	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9638
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9680	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9803	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9949	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996

ANTI-LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	Mean Differences
00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0	0	1	1	1	1	1	2	2	2
01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0	0	1	1	1	1	1	2	2	2
02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0	0	1	1	1	1	1	2	2	2
03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0	0	1	1	1	1	1	2	2	2
04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0	1	1	1	1	1	1	2	2	2
05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0	1	1	1	1	1	1	2	2	2
06	1148	1153	1156	1159	1161	1164	1167	1169	1172	1175	0	1	1	1	1	1	1	2	2	2
07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0	1	1	1	1	1	1	2	2	2
08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0	1	1	1	1	1	1	2	2	3
09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	0	1	1	1	1	1	1	2	2	3
10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1286	0	1	1	1	1	1	1	2	2	3
11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0	1	1	1	1	1	1	2	2	3
12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	1	1	1	2	2	3
13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	1	1	1	2	2	3
14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	1	1	1	2	2	3
15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0	1	1	1	1	1	1	2	2	3

Contd.

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	Mean Differences
16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0	1	1	1	1	2	2	2	3	3
17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0	1	1	1	1	2	2	2	3	3
18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0	1	1	1	1	2	2	2	3	3
19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0	1	1	1	1	2	2	3	3	3
20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0	1	1	1	1	2	2	3	3	3
21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0	1	1	1	1	2	2	2	3	3
22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0	1	1	1	1	2	2	2	3	3
23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0	1	1	1	1	2	2	2	3	4
24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0	1	1	1	1	2	2	2	3	4
25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	1	1	2	2	2	3	4
26	1820	1824	1825	1832	1837	1841	1845	1849	1854	1858	0	1	1	1	1	2	2	3	3	4
27	1862	1866	1871	1875	1870	1884	1888	1892	1897	1901	0	1	1	1	1	2	2	3	3	4
28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0	1	1	1	1	2	2	3	3	4
29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0	1	1	1	1	2	2	3	3	4
30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0	1	1	1	1	2	2	3	3	4
31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0	1	1	1	1	2	2	3	3	4
32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0	1	1	1	1	2	2	3	3	4
33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0	1	1	1	1	2	2	3	3	4
34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1	1	1	1	1	2	2	3	4	5
35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1	1	1	1	1	2	2	3	4	5

Contd.

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	Mean Differences
	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1	1	2	2	2	3	3	4	4	5
37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1	1	2	2	2	3	3	4	4	5
38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1	1	2	2	2	3	3	4	4	5
39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1	1	2	2	2	3	3	4	5	5
40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1	1	2	2	2	3	3	4	4	5
41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1	1	2	2	2	3	3	4	4	5
42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2585	1	1	2	2	2	3	3	4	4	5
43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1	1	2	2	2	3	3	4	4	5
44	2754	2761	2767	2773	2780	2786	2793	2799	2806	2812	1	1	2	2	2	3	3	4	4	5
45	2813	2826	2831	2832	2844	2851	2858	2864	2871	2877	1	1	2	2	2	3	3	4	5	6
46	2884	2891	2897	2904	2911	2917	2924	2931	2936	2944	1	1	2	2	2	3	3	4	5	6
47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1	1	2	2	2	3	3	4	5	6
48	3020	3027	3034	3041	3048	3059	3062	3069	3076	3033	1	1	2	2	2	3	3	4	5	6
49	3090	3097	3105	3112	3119	3120	3133	3141	3148	3155	1	1	2	2	2	3	3	4	5	6
50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	2	2	3	3	4	5	6
51	3236	3243	3251	3258	3266	3272	3281	3289	3296	3304	1	2	2	2	2	3	3	4	5	6
52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	2	2	3	3	4	5	6
53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	2	2	3	3	4	5	6
54	3467	3457	3483	3491	3499	3598	3516	3524	3532	3540	1	2	2	2	2	3	3	4	5	6
55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	2	2	3	3	4	5	6

Contd.

	Mean Differences									
	0	1	2	3	4	5	6	7	8	9
56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707
57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793
58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882
59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972
60	3982	3990	3999	4009	4018	4027	4036	4046	4055	4064
61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159
62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256
63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355
64	4366	4375	4385	4395	4406	4416	4426	4436	4446	4457
65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560
66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667
67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775
68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887
69	4899	4909	4920	4932	4943	4955	4966	4977	4999	5000
70	5012	5028	5035	5047	5058	5070	5082	5093	5105	5117
71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236
72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358
73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483
74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610
75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741

Contd.

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	Mean Differences
76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12	
77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6012	1	3	4	5	7	8	10	11	12	
78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13	
79	6166	6180	6194	6209	6228	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13	
80	6310	6324	6330	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13	
81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14	
82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14	
83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14	
84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15	
85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15	
86	7244	7261	7278	7293	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15	
87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16	
88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16	
89	7702	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16	
90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17	
91	8129	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17	
92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	9	11	13	15	17	
93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18	
94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	13	17	19	
95	8910	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	13	17	19	
96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19	
97	9318	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20	
98	9556	9572	9594	9618	9638	9661	9688	9705	9727	9750	2	4	7	9	11	13	16	18	20	
99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20	