

স্কুল অব বিজনেস
SCHOOL OF BUSINESS
বাংলাদেশ উন্মুক্ত বিশ্ববিদ্যালয়

MBA 2307
BUSINESS STATISTICS FOR DECISION MAKING

—Course Development Team—

Writer

Mohammad Ali Miyan

Professor

Department of Management

University of Dhaka

Dr. Nasirul Islam

Associate Professor

Open School

Bangladesh Open University

Dr. Qazi Md. Galib Ahsan

Professor

School of Business

Bangladesh Open University

Editor and Style Editor

Dr. Md. Mayenul Islam

Professor

School of Business, BOU

Coordinator

Dean

School of Business

Bangladesh Open University

This book has been published after being refereed for the students of
School of Business, Bangladesh Open University

MBA 2307
BUSINESS STATISTICS FOR DECISION MAKING

স্কুল অব বিজনেস
SCHOOL OF BUSINESS

Bangladesh Open University
বাংলাদেশ উন্মুক্ত বিশ্ববিদ্যালয়

MBA 2307

BUSINESS STATISTICS FOR DECISION MAKING

Published by: Publication, Printing and Distribution Department, Bangladesh Open University, Gazipur - 1705. © Bangladesh Open University. **Date of Publication :** August, 2025. **Computer Compose & Desk-Top Processing :** Salauddin Ahmed and Mahbubul Alam **Cover graphics:** Abdul Malek, **Cover Design:** Monirul Islam, **Printed by :** Mania Art Press, 53/1, North Brook Hall Road, Banglabazar, Dhaka-1100.

All rights reserved by the School of Business, Bangladesh Open University. No part of this book can be reproduced in any form without proper permission from the publisher.

Preface

Bangladesh Open University started its MBA program in "Distance Mode" more than thirty years ago. Distance learning is very different from traditional learning. Except the National University, all other universities, both public and private, in Bangladesh offer their undergraduate and post-graduate programs in Business Studies following exactly the U.S. semester system. In the U.S. semester system, a 3-credit semester course has 45 hours of class lectures. In Bangladesh, the number of credit hours a course has is used to measure how much work is done in it. The BOU MBA program requires students to take "MBA 2307: Business Statistics for Decision Making," which is a 3-credit course. This course is required for all students who are enrolled in the MBA program.

In "Distance Mode," students have to study on their own using the materials that the institution gives them. Students get 12 tutorial classes in a semester, just before the final exam, to talk to tutors about problems they are having with the study materials for each subject. This helps them be ready for the final exam and do their assignments correctly. When the university asked the three of us to write a study guide on Management Accounting, we chose to think about the following issues:

(i)	Mode of Imparting knowledge	Distance Mode
(ii)	The students	Students with different backgrounds can join the program
(iii)	The contents	The Curriculum Committee of the University has specified issues to be included, although there are certain very important issues left out.
(iv)	The existing literature	In the market innumerable text books are available on management accounting. At the same time there are some internationally recognized research based journals on Business Statistics available.

In a distance mode, students should begin their studies independently, following the study materials provided by the university. So, a "Study Book" is just a collection of courses for studying. Study courses must be structured to enable students from all disciplines to utilize clear English and simplistic, everyday examples to enhance reader engagement with the topics.

There are 13 units in the Business Statistics for Decision Making course. This study guide has 49 lessons, which is about 3 lessons for each discussion topic. Each lesson is set up such that a student can finish it in an hour. This means that a student usually has to spend 49 hours (1×49) to learn the topic. This will make the endeavor equal to 49 hours

of typical classroom lectures. The following are the study modules and lessons in the manual:

Unit 1	Introduction to Business Statistics	2 lessons
2	Collection of Statistical Data	4 lessons
3	representation of Statistical Data	2 lessons
4	Measures of Central Tendency	5 lessons
5	Measures of Variation	4 lessons
6	Correlation Analysis	4 lessons
7	Regression Analysis	2 lessons
8	Index Numbers	3 lessons
9	Probability and the Three Important Distributions	4 lessons
10	Test of Hypothesis	5 lessons
11	Chi Square Test	4 lessons
12	Sampling and Sampling Distribution	6 lessons
13	Business Forecasting and Time Series Analysis	4 lessons

We have looked at all the available texts on Business Statistics for Decision Making to find significant topics and methods that should be taught in the classes. At the end of the handbook, there is a list of the books that were evaluated and acknowledged. We are grateful to all the writers of the books and articles we used to write this book. We want to thank Prof. Dr. Md. Harun-Ar-Rashid of Chittagong University for reviewing the book. We don't mind saying how serious he was and how much time he spent going over the handbook. He went through every sentence, word and problems used in the manual.

We thanked the following people for their help:

- (i) Editor and Style Editor: Prof. Dr. Md. Mayenul Islam, SOB, BOU
- (ii) Coordinator: Dean, School of Business, BOU
- (iii) Former Dean of SOB: Prof. Dr. A. T. M. Tofazzel Hossain
- (iv) Computer Operatory: Md Salauddin Ahmed & Mahbubul Alam

We are grateful for the ideas and suggestions our friends and coworkers have offered us. Last but not least, we want to thank our family for letting us work on the manual instead of spending time with them.

If the readers, students, and tutors like the course material, we will feel like our work was worth it. It's concerning that it took more than ten years to put together the guidebook. We think that the delay is due to administrative carelessness and negligence. We are sorry for the pain that students and teachers have had to go through.

We welcome any and all suggestions for making the manual better.

Authors

Professor Mohammad Ali Miyan
Dr. Nasirul Islam and
Professor Dr. Qazi Mohammad Galib Ahsan

Contents

		Page No.
Unit –1	Introduction to Business Statistics	1
	Lesson#1: Origin, Growth and Definition of Statistics	3
	Lesson#2: Statistical Methods and Their Uses	11
Unit – 2	Collection of Statistical Data	21
	Lesson#1: Statistical Inquiry and Sources of Data	23
	Lesson#2: Framing a Questionnaire and Scrutinizing Statistical Data	31
	Lesson#3: Classification of Statistical Data	37
	Lesson#4: Lesson 4: Tabulation of Statistical Data	49
Unit – 3	Representation of Statistical Data	57
	Lesson#1: Graphs for Representation of Statistical Data	59
	Lesson#2: Diagrams for Presentation of Statistical Data	73
Unit – 4	Measures of Central Tendency	81
	Lesson#1: Measures of Central Tendency	83
	Lesson#2: The Median	93
	Lesson#3: The Mode	99
	Lesson#4: Geometric Mean, Harmonic Mean	105
	Lesson#5: Relationship between Different Measures of Central Tendency	111
Unit – 5	Measure of Variation	119
	Lesson#1: Variation or Spread or Dispersion	121
	Lesson#2: Measures of Dispersion (Distance Measures)	125
	Lesson#3: Measure of Dispersion (Average Deviation)	131
	Lesson#4: Relative Measures of Dispersion	143
Unit – 6	Correlation Analysis	151
	Lesson#1: Fundamentals Concept of Correlation	153
	Lesson#2: Methods of Studying Correlation	163
	Lesson#3: Rank Correlation	193
	Lesson#4: Concurrent Deviation and Least Squares Method	201
Unit – 7	Regression Analysis	219
	Lesson#1: Introduction to Regression Analysis	221
	Lesson#2: Methods of Obtaining Regression Lines	233
Unit – 8	Index Numbers	269
	Lesson#1: Introduction to Index Number	271
	Lesson#2: Methods of construction of Index Numbers	279
	Lesson#3: Quantity Index Number, Tests of Index Numbers and Other Index Methods	293

Unit – 9	Probability and the Three Important Distributions	323
	Lesson#1: The Concept of Probability	325
	Lesson#2: Random Variables, Probability Distribution, the Binomial Distribution	331
	Lesson#3: Normal Distribution	335
	Lesson#4: Poisson Distribution	339
Unit – 10	Test of Hypothesis	343
	Lesson#1: Hypothesis Testing	345
	Lesson#2: Types of Decisions and Error	351
	Lesson#3: Testing Hypothesis about Population Mean and Difference Between Two Population Mean	361
	Lesson#4: Testing Hypothesis about Population Proportion and Difference Between Two Population Proportion	375
	Lesson#5: Non-Parametric Test	381
Unit – 11	Chi-Square (χ^2) Test	389
	Lesson#1: Chi-Square Distribution	391
	Lesson#2: Condition for the Application of χ^2 Test and Uses of χ^2 Table	397
	Lesson#3: Test of Independence	401
	Lesson#4: Test of Goodness of Fit and Test of Homogeneity	407
Unit – 12	Sampling and Sampling Distribution	415
	Lesson#1: Sampling – Purposes & Types	417
	Lesson#2: Sampling and Non Sampling Error	425
	Lesson#3: Sampling Distribution	431
	Lesson#4: Statistical Quality Control	437
	Lesson#5: Acceptance Sampling by Attribute and Sampling Plan	443
	Lesson#6: Central-Limit Theorem	449
Unit – 13	Business Forecasting & Time Series Analysis	453
	Lesson#1: Time Series	455
	Lesson#2: Measurement of Trend	463
	Lesson#3: Seasonal Variations	475
	Lesson#4: Cyclical Variation	483
	References	489
	Appendix	491

INTRODUCTION TO BUSINESS STATISTICS



The study of statistics or statistical methods is playing an increasingly important role in nearly all phases a human endeavor. Formerly dealing only with political state; today the influence of statistics has spread the business, agriculture, economics, communications, education, biology, chemistry, physics, political science and other fields of science and engineering.

In this unit we have discussed introductory issues relating to historical development of statistics, definition of statistics, characteristics of statistics and statistical methods and their uses by dividing all these into two lesson.

Lesson 1: Origin, Growth and Definition of Statistics

Lesson Objectives:

After completing this lesson you will be able to:

- Describe the origin and growth of statistics;
- Define statistics;
- State the characteristics of statistics;
- Explain the scope and importance of statistics;
- Explain the limitations of statistics.

Introduction

The subject statistics in its modern sense is the outcome of the nineteenth century intellectual endeavor. The subject matter however, has got its origin in the ancient time. As people abandoned their nomadic way of life and started living in social groups. There arose the instinct of knowing each other's strength, wealth and position both at individual and group level. The group chiefs or rulers felt the need of collecting and assimilating information relating to population and wealth so as to enable them to impose and realize taxes, know each other's military strength as well as financial position. The formulation of administrative, fiscal and military policies was facilitated by the knowledge of manpower and material strength. This background has tempted some authors to identify the subject of statistics as the "science of kings."

Ancient Pharaohs and Hebrews used to collect information about population, land and wealth. Record of such collection of data is found in Egypt at about 3050 B.C. in connection with the construction of Pyramid. Subsequently, a census of all land in Egypt was conducted by Rameses II at about 1400 B.C. to redistribute the land among the inhabitants of Egypt. History also tells us that the Chinese, the Roman and the Greek rulers used to conduct similar census of population and wealth.

Historical records of the middle ages reveals that statistics were also collected at government level.

The purpose of such collection was more or less the same as before, viz., the formulation of administrative, fiscal and military policies of the government. But the scope and frequency of such collection was enlarged to some extent. Some evidences of such collection are found in the history of this period. Domesday survey was undertaken by William, the conqueror, during the middle ages. Investigations were also carried out by the great Muslim ruler, Al-Mamun of Baghdad, emperor Frederick II of Germany and Edward II of England. In Indian subcontinent, during the reign of Moghul Emperor Akbar, a population and land survey was carried out and the findings were compiled in the famous volume entitled 'Ain-i-Akbari'.

So far we have noted that collection of statistics was done only at the state level and its scope was confined to the need of the state viz. formulation of administrative and military policies. But from the mercantilistic period, the scope of the subject was enlarged bringing within its field the formulation of commercial and industrial policy of the state. This was due to the adoption of the politico-economic policy

Ancient Pharaohs and Hebrews used to collect information about population, land and wealth. Record of such collection of data is found in Egypt at about 3050 B.C. in connection with the construction of Pyramid.

Historical records of the middle ages reveals that statistics were also collected at government level.

known as Mercantilism by the governments of the East European countries. Until the mercantilistic period, state played an insignificant role in the economic field. The adoption of the mercantilism brought the state directly in the economic sphere. Mercantilism involved a policy of the government to subjugate the interest of the individual to the interest of the community, the maintenance of overall prosperity in the country, and favourable balance of trade. This called for the control of the individual's economic activity by the state. The formulation of economic policy under mercantilism necessitated collaborate and systematic collection of statistical data on trade, commerce and industry.

The beginning of the 17th century witnessed the new use and application of statistical data. In many English and German cities statistical records on birth, marriage and deaths were used by the Protestant churches to check on illegitimate births.

The beginning of the 17th century witnessed the new use and application of statistical data. In many English and German cities statistical records on birth, marriage and deaths were used by the Protestant churches to check on illegitimate births. In 1612, Professor George Obrecht of Strasbourg University initiated the idea of vital and criminal statistics. He demonstrated with practical illustrations the use of the statistics in developing methods for reforming the moral character of the people as well as to devise a system of life insurance and pension. The first analytical study in the field of vital statistics was made by Capt. John Graunt of London in 1661. From the analysis, he concluded that the population of a country could be reasonably estimated from the birth and the death records. In 1691, Caspar Neumann of Breslan collected figures of 5,869 deaths along with age and made some significant conclusions regarding life expectancy at certain age group. Edmund Halley, a great scientist and astronomer, made use of Neumann's works and prepared the first complete life table. This made possible the calculation of life expectancy at each age level and thus, laid the foundation of a scientific system of life insurance. Sir William Petty contributed to the development of life insurance by the preparation and analysis of mortality tables. The first insurance institution came up in England in 1698 and in 1699 "Society of Assurancy for Widows and Orphans" was established in England.

In the 17th century, there was also a significant development in the science of mathematical probability. The theory of probability was developed by notable scientists like Pascal, Formet, De Mere, Huygens and J. Bernoulli.

In the 17th century, there was also a significant development in the science of mathematical probability. The theory of probability was developed by notable scientists like Pascal, Formet, De Mere, Huygens and J. Bernoulli. This was developed gradually in connection with determining the chances of winning in gambling. J. Bernoulli (1654-1705), a Professor of Basel, enunciated the theory of probability in his great work "Ars Conjectandi". His work was published posthumously, in 1713, by his nephew Nicholas Bernoulli. N. Bernoulli also worked on probability and gave the theory of "moral expectation." The theory of probability subsequently formed the basis of modern statistics.

It was Gottfried Achenwall who, in 1749, first used the term 'Statistics' to refer to a complete and independent subject matter.

In the 18th century, need for improving the method of analyzing and interpreting the statistical data was much felt. The need for improvement became more important with the increase in the volume of statistical data and their application in a great variety of fields. J.P. Sussmilch (1707-1767) made an attempt in this direction and tried to demonstrate the doctrine of 'Natural Order' with the help of practical data. His work was published in 1761. Until the middle of the 18th century, most of the authorities like Graunt, Petty, Halley, Sussmilch carried out their work under the name of Political Arithmetic. It was Gottfried Achenwall who,

in 1749, first used the term ‘Statistics’ to refer to a complete and independent subject matter. He derived the term ‘statistics’ from the Italian word ‘statista’ “meaning states and he defined the subject as “the political science of several countries.”

G. Achenwall used the word statistics to refer to a complete subject matter. But he referred to the subject matter as political science and his idea was to compare political condition of different states. The foundation of the science of statistics, as is understood today, was laid by the famous Belgian astronomer and mathematician L.A.J. Quetelet (1796-1874). His initial field of research was meteorology which led him to study the different phenomena of vegetation, animal and mankind. He made an extensive study of the social, physical and moral characteristic of human being. He discovered that every phenomenon so studied produces more or less similar results. He found that in every phenomenon there is an average or norm and individuals show a tendency of variation from the average. This variation follows a mathematical regularity when the number of occurrences is large. He illustrated that all human actions like crimes, accidents, suicides are subject to the same law. Quetelet demonstrated with the help of empirical data that the variation of individuals from the norm follows the binomial law, which is contained in the mathematical law of probability. Thus he gave the interpretation of variability and the principle of ‘Constancy of Great Numbers’ upon which the theory of modern statistics rests.

The foundation of the science of statistics, as is understood today, was laid by the famous Belgian astronomer and mathematician L.A.J. Quetelet (1796-1874).

A significant contribution in the field of mathematical probability was made by Laplace (1749-1827). His work was published in 1812. His ideas, analytical methods and findings are regarded as outstanding contributions in the field of mathematical probability. Contribution made by him helped in the development of the science of statistics.

Sir Francis Galton (1822-1911), made use of statistical data to work on heredity of man and laws governing the transmission of physical and mental characteristics from one generation to another. Thus he gave the concept of ‘regression line’. His successor Karl Pearson (1857-1936) worked extensively with biological data and made a number of remarkable contributions in the field of statistical theory. Many of the statistical methods bear his name. The contemporary scene of statistical development was dominated by the original and remarkable contributions of R. A. Fisher.

Sir Francis Galton (1822-1911), gave the concept of ‘regression line’. His successor Karl Pearson (1857-1936) worked extensively with biological data and made a number of remarkable contributions in the field of statistical theory.

The development of the theory of probability in the 18th and the 19th centuries by great mathematicians like Bernoulli, Laplace and others; law of variability and ‘Constancy of Great Numbers’ by Gauss paved the way for the modern science of statistics. The notable contribution of Karl Pearson and R. A. Fisher led the subject to its present stage of development and established the subject as a separately identified branch of human knowledge. Improvements in theory of statistics were also effected by the works of August Meitzen, Pareto, Adam, Edgeworth, A. L. Bowley, G. Udny Yule, M.G. Kendall and a great many other authors.

The notable contribution of Karl Pearson and R. A. Fisher led the subject to its present stage of development and established the subject as a separately identified branch of human knowledge.

The application of statistics in the field of economics was slow upto the later part of the 19th century, although the beginning was made by persons like W. Petty and G. King in the later part of the 17th century. The 18th

century enlargement in the volume and variety of statistical data did not appreciably increase their application in the field of economics. The 19th century witnessed the growing application of statistics in the field of economics. W.S.Jevons in his work '*Theory of Political Economy*' published in 1871, emphasized the need of testing the validity of economic laws with the help of statistical laws and as a corollary emphasized the need for more complete and precise statistical information. He gave the concept of seasonal movements, secular trends, cycles in a time series and the concept of index numbers. In other words, he applied statistics into the analysis of economic variable. In the early 20th century the liaison between Statistics and Economics was to some extent established by the efforts of a good number of economists, noted among them are Alfred Marshall, Pareto, F.Y.Edgeworth. For handling of economic data certain new methods were also devised at this time. The improvements in statistical theories and their application also facilitated the application of statistical methods in Economics. However, the period after World War II witnessed the increased application of statistical theories in the formulation of economic policies of modern states.

Statistical methods and ideas were devised, practiced and enunciated by a great variety of individuals in different countries and at different times spreading over centuries.

From the above discussion it can be observed that statistical methods and ideas were devised, practiced and enunciated by a great variety of individuals in different countries and at different times spreading over centuries. In the 20th century the works of these individuals were systematically arranged and harmonized to form the science of Statistics. New methods as well as new application of existing methods are being devised gradually. Thus the chain of good statistical methods are expanding day by day.

Definition of Statistics

A study of the growth and evolution of the subject Statistics indicates that the term has been used by different authors at different times to indicate different aspects of knowledge. Thus, it is necessary to formulate a workable definition of the subject so as to permit commonality of thought and coherency in subject-matter.

The term 'statistics' has been derived from the Italian word 'statista' or the Latin word 'status' meaning political state

The term 'statistics' has been derived from the Italian word '*statista*' or the Latin word '*status*' meaning political state. Although initially the term 'statistics' was used to refer to the information relating to the activities of political state yet the term embraced much wider meaning later on. In simple word the word statistics is concerned with scientific methods relating, summarizing and presenting and analyzing data for drawing valid conclusions and making reasonable discussion on the basis of such analysis. In its modern form, the term is used in two different meanings. In the first place, it is used in plural sense to refer to numerical information or data. In the second place, it is used in singular sense to refer to the subject embracing the methods and techniques of dealing with numerical data. In other words, in plural sense one refers to the raw material itself while in the singular sense one refers to the methods of dealing with the raw material. In our everyday use the quantitative information regarding birth, prices and wages are termed as birth statistics, price statistics and wage statistics respectively. In all these cases, the word statistics is used in plural sense to refer to numerical data relating to the specific field. If we

look to these quantitative information from a different angle it will be clear that certain methods, procedures and techniques are involved in their collection, presentation, analysis and interpretation. These methods, procedures, and techniques form an independent subject of study known as 'Statistics'. Here the term 'statistics' is used in the singular sense. The distinction between the two implications of the term is not of much significance for practical usage. However, from an academic point of view the distinction is important to keep in view.

Statistics in Plural Sense

Statistics in its plural sense should be termed as statistical data rather than statistics. In common usage, the word statistics means statistical data. A good many renowned authors have defined statistics in plural sense. A few of these definitions are quoted here for elucidation of views on this.

A good many renowned authors have defined statistics in plural sense

Webster defined statistics, as "statistics are classified facts respecting the condition of the people, in a state, specially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement."

Dr.A.L.Bowley defined statistics as "statistics are numerical statements of facts in any department of inquiry placed in relation to each other."

According to Yule and Kendall, "By statistics we mean quantitative data affected to a marked extent by a multiplicity of causes."

Connor defined the same as "statistics are measurements, enumerations or estimates of natural or social phenomena, systematically arranged so as to exhibit their inter-relations."

Webster's definition refers only to the facts relating to the condition of the people of a state and does not include the data relating to other fields of knowledge. But at present the scope of statistics is much more comprehensive and wider to include all aspects of human activity.

Webster's definition refers only to the facts relating to the condition of the people of a state and does not include the data relating to other fields of knowledge. But at present the scope of statistics is much more comprehensive and wider to include all aspects of human activity.

According to Bowley's definition statistics are numerical data concerning any field of enquiry, arranged in relation to each other. But it does not make any particular reference to the method of collection, accuracy of data and the factors influencing the data. Yule and Kendall state the quantitateness of the statistics and the multiplicity of causes affecting them but did not particularly take into account other elements. In his definition, Connor only stated that statistics are numerical statements of social and natural phenomena arranged in systematic manner to reveal the relationship among them.

The most exhaustive definition of statistics has been given by Prof. H.Secrist. According to him, "By statistics we mean aggregate of facts, affected to a marked extent by a 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."

Characteristics of Statistics

An analysis of the above definitions indicates certain basic characteristics of statistical data or statistics (in plural sense). These basic characteristics are :

1. Statistics are aggregate or population of facts: As is evident from the term itself, statistics must be composed of more than one single figure. A single figure cannot be called statistics but a group or aggregate of figures are. A single sale or purchase may be of interest to a particular seller or buyer but is not important from a statistical point of view. However, the aggregate of all the sales transactions in a particular town at a particular date by category of merchandise are statistics.

2. Statistics must be numerically expressed: Qualitative expressions like good, fair, bad are not statistics. To be statistics facts must be expressed in numerical terms. Facts can be expressed more precisely and accurately in figures rather than in qualitative terms. A description of the size of firms in an industry in qualitative terms like large, medium, small can hardly give one a precise idea about the size of the firms. In the industry the size of the firm can be better expressed in terms of their capital such as firms of Tk. 1,00,000 capital stock, Tk. 3,000 capital stock and Tk. 5,000 capital stock. These quantitative expressions of the size of the firms can be termed as statistics. However, simple mathematical expressions must not be confused with statistics. Mere numerical figures like 5, 7, 9, 11 are not statistics. To be statistics they must be expressed in specific units of measurement and placed in context.

3. Statistics must relate to a department of enquiry: These should be the results of an enquiry relating to a pre-determined field. The purpose of enquiry must be unambiguous and clear.

4. Statistics must be comparable and homogeneous: One of the important purposes of statistics is to permit comparison between two or more phenomena. The collection of statistics should be made in a way so as to permit this comparison. An essential pre-requisite of comparability is the homogeneity and comparability of data. Unconnected numerical figures collected without any purpose is not statistics. Height of a man, income of a family and sale proceeds of a shop, even if expressed in numerical figures, combinedly will not constitute statistics. Because they are neither comparable, nor homogeneous, nor capable of being placed in relation to each other.

5. Statistics are affected to a considerable extent by multiplicity of causes: Statistics are not the effect of a single factor. They are affected by a large number of causes and it is difficult to study one cause separately. The saving of a person is affected by his income level, saving habit, expenditure pattern, etc. The sale of a retail shop is affected by its size, location, sales efforts, price policies, etc. In both the above cases, a study of saving and sale can only be made into consideration the different elements affecting these variables.

6. Statistics are to be enumerated with reasonable accuracy: Usefulness of statistical data depends upon their accuracy. These are not likely to be absolutely accurate, but attempt should be made to minimize the error and maintain a fair degree of accuracy in the enumeration. However, the degree of accuracy to a major extent depends upon the type of information to be collected and the purpose for which these are to be collected.

7. Statistics should be collected in a systematic manner: The collection of statistical data should be systematic and not haphazard. Haphazardly collected data might lead to erroneous conclusions.

Statistical methods
are used to
elucidate the varied
and complex
statistical data

Thus, statistics are aggregate of facts, quantitatively expressed, placed in a comparable and homogeneous manner enumerated with reasonable accuracy, collected in a systematic manner, and related to any department of enquiry.

Statistics in Singular Sense

Many renowned authors have also defined statistics in singular sense. An analysis of some of these definitions will make the implication of the term more perceptible.

Bowley defined statistics as “the science of measurement of social organism, regarded as a whole in all its manifestations.” By this definition he has limited the scope of the science of statistics to the study of social organism, that is, of human being and his activities. But the science of statistics has got much wider field of application. So this definition is not exhaustive enough to cover the scope of the subject as it is today. Even the author himself in his works emphasised the varied application of statistics. The same author has again narrated statistics as “the science of counting.” This definition also narrows down the field of statistics to only one aspect. In addition, the definition is inadequate from two other points of view. Statistics does not involve mere counting but also the process of estimating and forecasting. In practice, it is hardly possible and in some cases physically impossible to have a complete and accurate count. The population of intermediate period is only estimated and not counted. Thus the size of population of a particular year can be estimated with fair degree of accuracy although no actual counting is made. On the other hand, the above definition only refers to the collection of data and not the further process of analysis and interpretation of collected data. But the aspects of collection, analysis and interpretation are of equal importance in statistical study.

Bowley also termed statistics as “the science of averages.” Although average is an important tool of analysing statistical data, yet it is only one of the many statistical tools and cannot represent the science of statistics wholly. Other techniques like the study of variation, correlation and regression analysis, sampling and the like have got wide application in statistical work. Hence this expression can hardly give us a complete picture.

W. I. King defined 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 collection of estimates.” His definition places more emphasis on the interpretation rather than collection and analysis of data.

Any definition which does not cover the present meaning and scope of the subject statistics is obviously inadequate for describing the subject. An appropriate definition should be exhaustive enough to include all those aspects which are associated with the present day statistics and to exclude those concepts which are no more regarded as statistics. Thus, the science of statistics or simply statistics is the study of theories, methods and techniques developed and applied for collection, classification, presentation, analysis and interpretation of statistical data relating to any sphere of enquiry.

The science of statistics or simply statistics is the study of theories, methods and techniques developed and applied for collection, classification, presentation, analysis and interpretation of statistical data relating to any sphere of enquiry.

Summary

Statistics is the quantitative information of any inquiry. It is the scientific technique of collection, analysis interpretation and explanation for future development after any of data. In business field, collection of data on cost and benefit of an industry and interpretation for future development after analyzing the collected data.

Self-Assessment Question:

Short Question:

1. What do you mean by origin of statistics?
2. Define statistics?
3. Define statistics in plural sense?
4. Write down the characteristics of statistics?
5. Define the statistics in singuter sense?
6. Write the limitation of statistics.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Ancient pharaohs and Hebrews used to collect information about
 - (a) Population, land and wealth
 - (b) Mean, Median and Mode
 - (c) Regression, correlation and tune series
 - (d) Real, imaginary and cardinal
- (ii) Record of collection of data by ancient pharaohs and Hebrews were found in Egypt at:
 - (a) About 3150 B. C
 - (b) About 3000 B. C
 - (c) About 3500 B. C
 - (d) About 3050 B. C
- (iii) "Statistics is quantitative data affected to a marked extent by a multiplicity of causes" is defined by
 - (a) Sir Francis Galton
 - (b) Nicholas Bernoulli
 - (c) Yule and Kendall
 - (d) Prof. H. Secrist
- (iv) Who narrated statistics as "The science of counting"
 - (a) Prof. L A. J. Quetelet
 - (b) A. L. Bewley
 - (c) Sir Francis Galton
 - (d) R. A. Fisher
- (v) Who termed statistics as "The science of averages"
 - (a) A. L. Bewley
 - (b) Nicholas Bernoulli
 - (b) W. I. King
 - (d) Prof. George Obrecht.

2. Write "T" if the statement is true and "F" if the statement is false:

- (i) During the reign of Moghul Emperor Akbar, a population and land survey findings were compiled in "Ain-i-Akbari"
- (ii) Sir Francis Galton (1822-1911), made use of statistical data to work on heredity of man.
- (iii) The term Statistics has been derived from the Italian work "Status"
- (iv) Professor George Obrecht initiated the idea of vital and criminal statistics in 1621.
- (v) Goltfried Achenwall in 1749 first used the term "Statistics" to refer independent subject matter.

Answer:

Multiple-Choice Question: 1. (i) a (ii) d (iii) c (iv) b (v) a

True/False: 2. (i) T (ii) T (iii) F (iv) F (v) T

Lesson 2: Statistical Methods and Their Uses

Lesson Objectives:

After completing this lesson you will be able to:

- Describe statistical methods and their uses;
- Understand division of statistics;
- Explain scope and importance of statistics;
- Explain limitations of statistics;
- Narrate the distrust of statistics.

Introduction

Statistical methods are the devices and techniques applied in the collection, classification, analysis and presentation of statistical data and to make estimates, predictions and decisions therefrom. In short, statistical methods are nothing but the scientific tools devised to deal with statistical data. Yule and Kendall defined statistical methods, “By statistical methods we mean methods specially adapted to the elucidation of statistical data affected by a multiplicity of causes.”

Statistical methods are nothing but the scientific tools devised to deal with statistical data.

Statistical data to be comprehensible and useful must go through certain scientific processes. The purpose of processing is to simplify the mass of data, to make them more useful and to make the same easily understandable to common mind. Statistical methods are used to elucidate the varied and complex statistical data. They have been devised to study those phenomena which are not susceptible to study by experimental methods owing to the multiplicity of causes affecting the phenomena. Experimental methods are extensively used in physical sciences like Physics and Chemistry where experiments can be conducted under controlled condition and the data are more or less accurate and the causes affecting the data are few and simple. It is possible for the physicists and chemists to isolate a particular cause and determine the direct effect of that cause separately by experimental methods. This does not hold good in sciences and social sciences like Biology, Sociology and Economics. As a result, experimental method was found unsatisfactory for the purpose of elucidating data regarding social, economical and biological studies. Data in these fields are affected by complex, multifarious and heterogeneous causes many of which are difficult to isolate and control. Like physicists and chemists, the sociologists must study the phenomena as they exist and the external forces affecting the phenomena cannot be controlled as is the case in laboratory experiments. So they apply some special methods, other than experimental methods to deal with such situation. These special methods have been developed and applied to deal with economic, biological and social variables and although they are not as precise as experimental methods yet these are capable of producing results with fair degree of accuracy, if properly applied. The special methods so devised are called statistical methods. Statistical methods are mostly concerned with large numbers. Casual disturbances in the data cancel out each other in the long run leaving residue which is taken as the effect of more important and persistent factors.

Statistical methods are used to elucidate the varied and complex statistical data.

Application of statistical methods to collection and analysis of statistical data aids in the interpretation, prediction and decision with a reasonable degree of accuracy.

Application of statistical methods to collection and analysis of statistical data aids in the interpretation, prediction and decision with a reasonable degree of accuracy. Exclusion of statistical methods from the collection and analysis of information might lead to fallacious and misleading conclusions.

The application of statistical methods is not limited to social and economic phenomena. They have got wide application in physical sciences also. Experiments in the fields of physical science like Physics and Chemistry are also subject to some extent of inaccuracy resulting from human error and other variable causes which cannot be fully controlled in a laboratory. This element of inaccuracy again makes the application of statistical methods useful in those fields where experimental methods are generally applicable. Thus statistical methods have got wide application in natural sciences as well as in social sciences. However, they are more useful in the study of social sciences.

With the increase in the volume and variety of statistical data and with the opening of the new fields of investigation, statistical methods became more and more useful. Statistical methods like tabulation, averages, standard deviation, correlation and regression analysis, time series analysis, interpolation, index numbers, sampling and the like are finding their growing application. The study of these and other statistical methods is the theme of the book.

Exclusion of statistical methods from the collection and analysis of information might lead to fallacious and misleading conclusions.

Statistical methods are based upon general mathematical theories like the theory of probability, law of large numbers, theory of error of observations. Statistical theories are really the 'exposition' of statistical methods.

It is true that common people are only interested in the finished product and they do not bother much for the underlying techniques. But the finished product is always the out come of technique. Passengers travelling by air have very little interest in aero-dynamics involved in flying the aircraft; but the slightest carelessness of the technician or a deviation from the techniques by the pilot may produce ruinous results. Likewise, statistical results to be useful and informative must be the product of proper techniques. Deviation from the techniques or their improper application might produce misleading results.

Statistical methods like tabulation, averages, standard deviation, correlation and regression analysis, time series analysis, interpolation, index numbers, sampling and the like are finding their growing application.

Division of Statistics

The science of statistics can be broadly divided into the following four divisions :-

Statistical Theory

Statistical Methods

Descriptive Statistics

Inductive Statistics.

Descriptive Statistics and Inductive Statistics refer to the application of statistical methods and theories

Statistical theory includes general mathematical theories like the law of large numbers, theory of probability, theory of errors of observations, etc., which form the basis of statistical methods. Methods, techniques, and procedures involved in the collection, analysis and interpretation of statistical data constitute the Statistical Methods. These two divisions are mostly mathematical in nature and deals with general theories, formulae, equations and their derivation.

Descriptive Statistics and Inductive Statistics refer to the application of statistical methods and theories. Descriptive Statistics deals with collection, presentation, analysis, and interpretation of statistical data without involving generalisations. On the other hand, generalisation, predictions, estimations, and decisions made from collected and analysed data using statistical methods and theories come within the fold of Inductive Statistics. When we compute average with the sole purpose of describing a given phenomenon, we are using the mean as a tool of descriptive statistics, whereas if we make use of the average for making prediction or general conclusion, we are using it in inductive sense. Average salary of the employees of a factory may be described by calculating the mean salary of all the employees of the factory. This involves only description of mean salary. But if this mean averages of a particular factory is used as an estimate of average salary of the employees of all the factories in an industry instead of computing the average for all the employees in the industry, then we use the mean in inductive sense. In other words, inductive statistics involves induction (inference) about the population parameters by observing a representative part of that population. Thus, the distinction between descriptive and inductive statistics is based upon the purpose for which the data are used and not on the method employed.

Descriptive Statistics deals with collection, presentation, analysis, and interpretation of statistical data without involving generalisations. On the other hand

Generalization, predictions, estimations, and decisions made from collected and analysed data using statistical methods and theories come within the fold of Inductive Statistics

Scope and Importance of Statistics

Science of statistics brings within its fold all quantitative study relating to any field of enquiry. Whatever might be the field of enquiry, statistics as a tool of intelligent judgment is indispensable. Statistics, therefore, embraces all those fields of human knowledge where realisation of the significance of large numbers and of collective judgment is involved. The scope of intelligent judgment turns to be extremely limited and confusing where numerous and complex data influenced by widely varying causes are involved. In such a situation statistics has got a special role to play. Statistical methods provide means of handling such heterogeneous data and thereby assist in arriving at an intelligent judgment of a collective phenomenon. This is more particularly true in case of social sciences only. They are also useful to natural sciences. The field of statistics is all pervading, the limitation being its applicability to quantitative study. It not only serves the present but also helps in analysing the past events and thereby throws light on the future course of events. Statistics has now been recognised as a separate branch of human knowledge in its own right. The scope of statistics is ever increasing. With the passage of time the scope of statistics is embracing more and more new fields and the chain of good statistical methods is increasing day by day. The fundamental concepts and ideas may not change much, but there may be a shift in emphasis from one aspect to the other with the passage of time.

Statistics is indispensable in business and commerce. Every modern business manager must make elaborate use of statistics for making business decisions. The domain of business is enshrouded; by risks and uncertainties. A successful business manager is one who can reduce the risk and uncertainty to a minimum possible level. To accomplish this the

Statistics is indispensable in business and commerce. Every modern business manager must make elaborate use of statistics for making business decisions.

The statistical methods enable the estimation or prediction of future course of action by analysing the past and present available records

business manager is constantly required to draw upon past experience and knowledge as well as the detailed knowledge of the present condition. Statistics provides the record of the past experience as well as facts relating to present condition. On the other hand, the statistical methods enable the estimation or prediction of future course of action by analysing the past and present available records. A business manager's decision in the face of uncertainty must be based upon statistical information and techniques to guard against disastrous consequences. Under the large-scale system of production, production is done in anticipation of probable demand. So, a manufacturer of a certain product must make proper estimate of demand for his products and take into consideration other factors like seasonal variation, business cycles, changes in tastes and fashions, the purchasing power of consumers, etc., affecting the stipulated demand and then proceed to adjust his production schedule both in quantity and time schedule accordingly. The extent of precision with which the manufacturer can infer the probable demand and the effect of other allied factors has a definite effect upon his business success. Any inaccuracy in the estimate may result in overproduction or under production, both being undesirable from the point of view of a business manager. Overproduction may result in loss whereas underproduction may keep demand unsatisfied resulting in the loss of attainable profit. This is also true of a trader. A trader is required to adjust his stock on the basis of anticipated demand. Thus statistics both in terms of raw materials and tools turns to be a valuable guide to business managers.

Financial institutions like banks-industrial or commercial, are required to rely mostly on past records for their efficient operations. A banker must have clear knowledge of the variation involved in the calls of money at different periods of time maintain monetary reserve accordingly. Without this knowledge the bank cannot venture upon a systematic basis of lending. This the banker can have by analysing past variation on the calls for money upon the bank and proper knowledge of the economic condition of the people of the region in which the bank is operating.

The whole structure of insurance business is based upon the Law of Large Numbers. It would not have been possible to carry on insurance business without the help of mortality table which is prepared on the basis of statistical records on age at death as well as the rate of mortality. Expectation of life is estimated by using mortality tables and the amount of premium is determined accordingly. Other types of insurance like fire, marine, accident, unemployment, sickness, etc. are also based upon estimate of probable happenings or occurrences of such phenomena.

Every modern manager must have a clear grasp of the trends of the business activities as well as clear understanding of the economic framework within which he is to operate. He is to have previous knowledge of business cycles-booms and depressions, inflationary condition and so on, to equip himself accordingly. Statistics by indicating trends forewarns a business manager of the advent of such condition. They constitute valuable guide to speculators, underwriters, sharebrokers, investors and traders in future.

Statistics also aids transportation agencies like railway, tramway company, motor transport and shipping organisations. For example the railway company operating in a wide area should be able to predict the possible passenger and goods load in different areas at different times and there upon adjust its rolling stock and operational arrangements. A wrong estimate would result in higher operational cost. For making estimates the railway authority has to depend upon past knowledge provided by statistical records. In this way statistical records enable transportation agencies to operate with economy and efficiency.

Statistics helps in the formulation of organisational policy and the managerial planning. Formulation of policy in business matters is to be based upon adequate information of the situation and their proper analysis and interpretation. Starting from the general overall organizational policy to the sectional policy, like sales policy, procurement policy, financial policy, personnel policy-all need to be based upon proper diagnosis of the situation. In this respect, statistics is a valuable tool.

With the gradual increase in business complexity the need of managerial planning for achieving optimum working efficiency has increased enormously. The increased competition calls for maximum production with minimum cost. This is being done by increasing working efficiency and by minimising wastes. Scientific management as a tool of increasing efficiency has got wide acceptance. For scientific management the efficiency of the contributing factors in production is to be adjudged separately. This calls for extensive use of statistics. Modern concepts like Time Study, Motion Study, Job Evaluation, and Merit Rating are based on statistical techniques. Budgeting, which is an important devices for allocation and control of expenditure within an organisation, is based upon statistical information. Modern cost accounting makes use of statistical methods for analysing the various elements of cost as well as for the purpose of cost control. In this respect accounting records are found to be inadequate and need to be supplemented by statistical records. Cost control is vital in modern industry as most of the manufacturers compete on cost. In construction business, proper determination of cost before hand is very important since the quotations are to be submitted before the actual cost has been incurred. Cost quotations must be reasonable. High quotation may result in the loss of contract whereas a low quotation might result in loss in execution. Modern management has to make extensive use of statistical tools in controlling the quality of the product. Statistical tools like control chart, sampling techniques, normal curve, are used in maintaining the quality standard of product. Present day consumers are quality conscious and they want quality products. The management must endeavour to offer quality product. With the enormous increase in the scale of business operation and size of business organisation, it has become simply impossible on the part of the management to appreciate and understand a problem by studying the large mass of detracted numerical figures. Statistical methods provide the tools for classifying, concising, analysing and presenting the masses of numerical information so as to enable the management to have a dispassionate view of any problem.

Modern concepts like time study, motion study, job evaluation, and merit rating are based on statistical techniques.

Statistical methods provide the tools for classifying, concising, analysing and presenting the masses of numerical information so as to enable the management to have a dispassionate view of any problem

The kinship between statistics and business has led to the coining of the term “Business statistics”, dealing with collection, analysis, interpretation, and presentation of business facts.

Statistics cannot study qualitative phenomena like civilization, friendship, skill and intelligence unless these are reduced to precise quantitative forms.

Statistics has got an important place in business research. Statistical methods are applied for improving and testing technicalities involved in production. They are also extensively used in product development, market research and analysis and in many other spheres.

Thus it seems that statistics is lifeblood of successful business and commerce. Statistical methods and statistical data have turned to be a vital element in the decision making process. The kinship between statistics and business has led to the coining of the term “Business statistics”, dealing with collection, analysis, interpretation, and presentation of business facts. The importance of statistics in business is so much felt that many business organisations maintain their own statistical section. Trade associations and chambers of commerce have made collection and dissemination of statistical information an important function. However, in emphasising the importance of statistics in business, the importance of other elements must not be overlooked. Statistics only provides the tool and their proper utilisation depends upon the managers themselves. They are to be supplemented by practical knowledge and experience.

Limitations

Statistics in spite of its varied use and applications, has got its limitations too. In studying the subject, the limitations are worth keeping a mind. These are :-

- 1) Statistics studies mass phenomena and throws light on the results of collective actions. For example, per capita saving in Bangladesh can only give us an idea of average saving made by the people in general. This does not reveal the saving made by each individual or does not preclude the existence of indebtedness by any individual. Statistics cannot reveal this kind of situation. Thus a situation where a study of the individuals constituting the group is required, statistics is an unsatisfactory tool of study. It can only provide the group average without revealing the individual characteristics.
- 2) Laws of statistics are true only in the long run. Laws in natural sciences are invariant in nature but statistical laws are variant in nature. Statistical expressions are in terms of averages, approximations and probabilities. The nature of the subject does not allow it to be exact. The laws hold good when a fairly large number of cases are involved.
- 3) Statistics is applicable only in quantitative study. Statistics cannot study qualitative phenomena like civilization, friendship, skill and intelligence unless these are reduced to precise quantitative forms. In such situations statistical methods are found to be less appropriate. Therefore, the application of statistical methods turns to be limited only to studies of quantitative nature or to phenomena which can be expressed in exact terms.
- 4) Statistics may produce faulty conclusions either due to deliberate manipulation or due to inappropriate use. Deliberate manipulation may be made both at the time of collection as well as at subsequent interpretation. An advertiser of product might quote faulty figures to

serve his own end. Fallacious conclusions may also result from the use of statistics without their proper context. Data collected for one purpose, if used for another purpose, will lead to faulty conclusions.

5) Statistical methods provide only one approach to the study of a phenomenon. There are other methods or ways of looking to a phenomenon; statistics is only one of the many ways. Statistical evidences give only approximate idea of a situation. In general, the statistical evidence, to be valid, should be supplemented by other evidences.

Statistical evidences give only approximate idea of a situation. In general, the statistical evidence, to be valid, should be supplemented by other evidences.

6) Like other sciences, Statistics has the chance of being misused. Statistical methods need to be carefully and prudently used, otherwise, their application will result in misleading conclusions. Non-experts might make hell out of statistics.

7) Statistics only provides the raw material and tool for making judgment and inferences but they do not constitute inferences for any study. They are only the means to an end, not the end in itself.

In the above paragraphs the main limitations of statistics have been outlined. An user of statistics should take cognizance of these limitations before making any tangible conclusion. In spite of these limitations statistics has got wide utility and importance in many sphere of human activity. In fact, the use and importance of statistics much outweighs the limitations.

In fact, the use and importance of statistics much outweighs the limitations.

Distrust of Statistics

Notwithstanding the wide application of statistics in different branches of human knowledge, some amount of popular distrust towards statistics is observed. Common attitude towards statistics is sharply divided. While one section believes that figures can prove anything, the other section believes that figures can prove nothing. The attitude of the extremists in both these respects are either due to over reliance or due to ignorance of statistics methods leading to failure in distinguishing between truism and falsehood. For neither of these views statistics can be blamed. As have been said, statistics in themselves are not inferences; they only prepare the ground for making inferences. Sometimes inferences derived from statistical analysis are taken as guaranteed and too much reliance is placed on the inference due to over-enthusiasm. While this is not desirable on the one hand, on the other hand it is not true that statistics cannot prove anything. Statistical methods provide useful tools for any inductive type of study and inferences derived from the proper application of statistical methods hold good to a large extent. So fault lies with the user, not with statistics. It has been rightly observed, 'statistics are like clay of which you can make a god or a devil, as you please'.

statistics are like clay of which you can make a god or a devil, as you please'

Fallacious conclusions and false arguments may result from the ignorance of the methods or due to deliberate manipulation of the methods. One may jump to a conclusion from a set of figures being ignorant of their context or being ignorant of proper methods of analysis and interpretation. Again, the unscientific method of collection may also result in faulty conclusion. One may also deliberately manipulate the

figures to serve his own purpose. He may quote one part of the data leaving the other part to prove his pet conclusions. Diametrically opposite conclusions may be drawn from the same set of data to serve the user's purpose. As a tool statistics can equally support true as well as false conclusions. Statistics only describes a quantitative phenomenon, classify, analyse and condense the facts to lay the ground for arriving at a well thought-out conclusion. As have been said, one may deliberately tamper with the data, having full knowledge, to having little knowledge of the application of statistical methods and respect to statistics. But for this statistics cannot be blamed. Users are to be blamed. Truly speaking, unrepresentative or incomplete figures compiled without any regard to statistical methods are not statistics. So long figures are derived with adherence to the principles underlying statistical methods and are used for the purpose for which they are meant, they cannot support false conclusions.

Figures are derived with adherence to the principles underlying statistical methods and are used for the purpose for which they are meant, they cannot support false conclusions.

One of the main shortcomings of statistics is that they do not always indicate their quality on face. An unrepresentative and crude table prepared without any regard to principles may appear to be equally informative like the one prepared with a great deal of labour and strict adherence to statistical principles to a casual observer. The same may not be true of a careful observer who may be able to discover apparent anomaly in the table. To properly evaluate a table the reliability of the source of information should be kept in mind. Another problem arises owing to the nature of expression. Statistics expresses facts quantitatively in definite forms and as such looks precise and the common people has a psychological attachment to accept them as true. But the reliability of an expression does not depend upon preciseness; it depends upon the method of their compilation.

Summary

In conclusion it can be said that statistical methods are very delicate and sensitive tools likely to be misused in the hands of an inapt user. They need to be used with care and restraint. Distrust arises owing to inapt handling and improper use. The limitations do not make the subject valueless. The subject itself cannot be blamed for the fault of the users. The improper use or inappropriate application of science is not peculiar to statistics alone. The same may arise in the case of other natural or social sciences. If due to the limitations one decides to do away with statistics it will be something like killing the goose which lays the golden eggs. In spite of the limitations, the science of statistics is rendering and will continue to render valuable services to mankind. With the gradual advancement of the science of statistics and greater amount of understanding of its intricacies, the limitations are fading away. The growing consciousness of statistical methods both on the part of users and commoners, diminishes the chance of distrust. However, the students in the field of statistics would do well to keep in mind these limitations to guard against pitfalls.

Self-Assessment Question:

Short Questions

1. Define descriptive statistics?
2. What do you mean Inductive statistics?
3. Write the limitation of statistics?
4. Define the scope of statistics.
5. Write the different types nature of statistics.
6. Explain the scope and importance of statistics?

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Experimental Methods are extensively used:
(a) Physical Science (b) Social Science
(c) Economical Studies (d) Biological Studies
- (ii) An administration prepares a series of charts and graphs pertaining to the patient that have stayed at the hospital during the past month; he/she is using which general category and statistical analysis?
(a) Quantitative Analysis (b) Inferential Analysis
(c) Descriptive Analysis (d) None of the above
- (iii) When a marketing manager surveys a few of the customers for the purpose of drawing a conclusion about the entire list of customer, the manager is applying:
(a) Inferential statistics (b) Descriptive Statistics
(c) Quantitative Statistics (d) None of the above
- (iv) Which of the following is not true of statistics?
(a) Statistics organizes and analyzes information
(b) Statistics allows conclusions about the data to be drawn
(c) Statistics answers questions with 100% certainty.
(d) Statistics collects and summarizes data.
- (v) The average age of the students in a statistics class is 23 years. Does this statement describe?
(a) Inferential statistics (b) Descriptive statistics
(c) Qualitative statistics (d) None of the above

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) Statistical methods are nothing but the scientific tools devices to deal with statistical data.
- (ii) The application of statistical method is limited to social and economical phenomena.
- (iii) Descriptive statistics deals with collection, presentation, analysis and interpretation of statistical data involving generalizations.
- (iv) The distinction between descriptive and inductive statistics is based upon the purpose for which data are used and not on method employed.
- (v) Statistics are not inferences; they only prepare the ground for making inferences.

Answer:

Multiple-Choice Question:

1. (i)- a. (ii)- c (iii)-a (iv)- c (v)-b

True/False

2. (i)- T (ii)- F (iii)- F (iv)- T (v)- T

Exercise

1. Define Statistics. Discuss its importance, scope and limitations.
2. What are the deferent types of statistics? Discuss the difference in descriptive and deferential statistics.
3. Explain statistical methods. How statistical methods are used to solve the problems related to business aspects of a country?
4. Explain clearly what do you understand by the business of statistics. Discuss its scope and limitations.
5. (a) Mention the characteristics of statistics.
(b) How statistical methods help in taking decision in respect of business?
6. (a) Discuss the characteristics of statistics.
(b) Discuss the importance of statistics in taking decision in Business.
7. Discuss the scope and nature of the business statistics. Explain how the problems related to business are solved using statistical methods.

COLLECTION OF STATISTICAL DATA



An essential preliminary step in statistical work is the collection of data. The collected numerical data form the base on which all subsequent works depend. They are the raw materials upon which statisticians base their analysis. Value of the analysis and interpretation depends on the quality of raw data. If the original data are inaccurate and are collected without using proper statistical techniques the conclusions drawn from them will also be misleading. Therefore, the collection should be done through a statistical enquiry following proper statistical principles.

In this Unit an attempt has been made to discuss statistical inquiry and sources of data framing a questionnaire and scrutinizing statistical data, classification of statistical data and tabulation of statistical data dividing this unit into four lessons.

Lesson 1: Statistical Inquiry and Sources of Data

Lesson Objectives:

After completing this lesson you will be able to:

- Understand the meaning of statistical inquiry;
- Define the problem;
- Explain purpose and scope of inquiry;
- Understand what is statistical unit and its characteristics;
- Understand the importance of degree of accuracy in planning statistical inquiry;
- Describe methods of field investigation for collecting primary data;
- Lesson properties of good classification.

Introduction

Data constitutes the foundation of statistical analysis and interpretation. R. A. Fisher mentioned that, “it is the applied mathematics to aggregate the collected data.” Data are collected from persons, places, industries, units available at a particular time and these are collected by statistical inquiry. Thus we need to discuss from whom, by whom and how the data are collected.

Statistical Inquiry

Statistical inquiry means an inquiry or search for information that is capable of numerical expression. So the term implies search for knowledge with the help of statistical methods of collection, tabulation, analysis and interpretation. Statistical inquiry involves several stages from the initial planning to the final interpretation. Planning the inquiry is an essential first step in the assembling of statistical data. The first step in planning is the definition of the problem and ascertaining the purpose and scope of the inquiry. Then comes the question of choosing the type of inquiry, fixation of unit of measurement and to some extent, determination of the degree of perfection or accuracy desired.

Defining the Problem

All human efforts are directed towards the solution of some problems. Problems are rooted in the need of making sound decisions. No effort to a mass evidence would be required had there been no problem. Statistical work is directed towards the solution of an existent or dormant problem. So the first requisite is to clearly identify and define the problem. A clear and concise definition of the problems will greatly help the inquiry so undertaken. Any deviation from the definition may call for a complete or partial change in the method of inquiry. When the problem is well defined, the chances of confusion and inaccuracy at the subsequent stages are minimized. So any statistical inquiry directed towards the solution of a problem must start with a clear definition of the problem.

Statistical inquiry directed towards the solution of a problem must start with a clear definition of the problem.

Purpose and Scope of Inquiry

A statistical inquiry should have certain purpose and scope. The purpose may be specific or general. A specific purpose inquiry is directed towards the solution of a specific problem. General purpose inquiry may

The purpose of the inquiry may be general or specific, A clear determination of the purpose and scope of the inquiry is essential before the actual collection of data starts.

obtain data which may be used for several specific purposes, e.g., population census. The purpose of the inquiry - general or specific, will determine the scope of the inquiry. A clear determination of the purpose and scope of the inquiry is essential before the actual collection of data starts. This enables the investigator to resolve the various problems involved in collection of data such as what information is to be collected, from whom they are to be collected, what frequency and periodicity of collection is to be followed and so on. The actual collection work may create difficulties and confusions unless the scope and purpose of the inquiry is pre-determined. Any ambiguity in the purpose and scope of the inquiry might lead to the collection of undesired information to the exclusion of essential information. This results in the wastage of time, energy and money. All these can be avoided by pre-determination of the purpose and scope of the inquiry. A clear understanding of the purpose of the inquiry on the part of the field operators will ensure better collection of data and uniformity in the process of collection. In determining the scope of an inquiry the cost involved in the inquiry must be compared with the expected utility to be derived from the inquiry. In other words, statistical inquiry should be a paying proposition.

Statistical Unit

Statistical unit forms the basis of recording statistical data.

The collection of statistical information involves the task of determining the unit in which the desired information is to be collected. The collection of data involves measurement, observation or counting of information to be expressed numerically. In order to avoid any ambiguity in the data the unit in terms of which the same is to be measured, observed or counted should be very precisely and clearly stated. Such a unit, some time referred to as a statistical unit, forms the basis of recording statistical data. Any ambiguity or inadequacy in the definition of the statistical unit will result in fallacious inferences. So it is evident that the units need to be very clearly defined and understood by those who will actually carry out the field investigation. The definition must be rigid and passed on to the field enumerators with clear instruction to adhere to the same. Any deviation will result in the lack of uniformity in the collected information rendering them unsuitable for comparison. Even if the fieldwork is carried out with all fairness and sincerity, still such information cannot be the basis for making valid conclusion. The definition of the unit is not only required for field operation but also for aiding the subsequent analysis and interpretation.

For the purpose of statistical inquiry it is necessary to have a restricted and well-formulated definition of the problem

The definition of the unit is not always an easy task. A task of counting things of the same type may sound to be very simple, the unit being a person or an accident or a thing. But if we take up the question of annual income of a section of population, it will bring with it the ideas like direct income, indirect income, individual income, family income, the treatment of overtime payment and bonus payment and the like. All these ideas about income are to be so synchronized that everyone in the investigation team refers to a particular connotation of income throughout the study. Similar is the case in the study of many other phenomena. For the purpose of statistical inquiry it is necessary to have a restricted and well-formulated definition of the problem. After clearly

establishing the definition, it is necessary to lay down the unit in which it is to be expressed, that is, whether it is to be expressed in value like taka, paisa or in simple number or it is to be expressed in quantity like maunds, seers, kg, grams or lbs. or it is to be expressed in hours like kilowatt-hour of consumption or in tons, miles and so on. The nature of study will determine the unit in which the measurement is to be expressed. The overall definition of statistical unit is also affected by the purpose and scope of inquiry. The mere establishment of a definition of a statistical unit and the means of its expression is not sufficient. The same must also be clearly understood by all members participating in the work and there must be strict compliance with the same throughout the survey. A clear understanding of any data in published form needs understanding of the units used in their compilation.

The definition of statistical unit to be unambiguous and precise should have some common characteristics that are:

a) The definition must be specific, self-explanatory and simple. The definition should be expressed in unambiguous and easily ascertainable terms so that little difficulty is encountered in following it. Simplicity is also a must because the enumerators who will be working in the field are expected to be of average intelligence only. They should be told in definite terms what to be included and what to be excluded.

b) The units must be homogeneous and uniform. Homogeneity is the first requirement in choosing the unit since the lack of homogeneity in the data might render the same unsuitable for comparison. The unit must imply the same meaning at different times in the course of the same inquiry. However, when owing to complex nature of the phenomena the chosen unit cannot be applied to all cases under observation, then it may be necessary to break the unit into sections, classes or groups in order to achieve homogeneity. In studying the income of villages, it may be necessary to divide income into agricultural and non-agricultural types to make the intragroup comparison more accurate. In defining the statistical unit the aspect of homogeneity must be kept in view.

Homogeneity is the first requirement in choosing the unit since the lack of homogeneity in the data might render the same unsuitable for comparison.

c) The unit must be stable. Once a unit has been chosen all the workers must follow it at different times and places during the same inquiry. If fluctuating unit is used for measurement then a definite mode of converting the same to standard unit must be laid down. In a survey spread over a long period of time complete stability in unit is difficult to maintain. However, in such cases some sort of stability can be achieved by using the method of conversion.

The Statistical unit must be stable and suitable and appropriate to the purpose of inquiry

d) The unit must be suitable and appropriate to the purpose of inquiry and the same should be capable of clear and correct ascertainment. Individual's notion towards a problem or towards a term varies widely and it is but natural to define a unit in terms of the objective of an inquiry. A unit suitable for one inquiry may not suit another inquiry. In each study it is necessary to set out the unit in terms of its objective and scope. For example, if the purpose is to ascertain the rate of wage in an industry, the wage of a worker is taken as a unit. On the other hand, if the purpose is to study the rate of production in the industry, each firm in the industry may be taken as a unit.

A unit suitable for one inquiry may not suit another inquiry. In each study it is necessary to set out the unit in terms of its objective and scope

Degree of Accuracy

The degree of accuracy maintained throughout the investigation and analysis should be reported upon in the final compilation.

The next step in planning a statistical inquiry is to lay down the degree of accuracy desired. In certain cases high degree of accuracy may be required and the plan should be formulated accordingly, while in most cases a high degree of accuracy may not be required and a reasonable standard of accuracy may serve the purpose. The scope and purpose of the inquiry affect the degree of accuracy to be maintained. The time and cost factor have also some definite effect upon the level of accuracy that can be maintained. Complete accuracy may not be worth attaining. A prompt and timely report with a tolerable accuracy level may be more useful than a delayed but more accurate report. Having regard to these three elements, namely, objective, time and cost, a decision on the level of accuracy to be attained is to be made and the collection of data should be planned according to the level of accuracy decided upon. At the interpretation stage of the data the degree of precision followed is to be kept in mind. The degree of accuracy maintained throughout the investigation and analysis should be reported upon in the final compilation.

Sources of Data

Primary data arise out of primary or original inquiry and involve direct field investigation.

After the preliminary plan of an inquiry has been decided upon, it is necessary to look for the sources of data, method of collection and, as a corollary to the method of collection, the choice of material to be collected and the management of the field force. The source of data can be classified into two viz., primary source and secondary source. The data procured from primary source are termed primary data and the data procured from secondary source are termed secondary data. Primary data arise out of primary or original inquiry and involve direct field investigation. Secondary data are those which are collected and published by various agencies for their own purpose but can be used by others also. The difference between primary data and secondary data is only in terms of their respective use. The same data are primary in the hands of the original collector but secondary in the hands of others. Price statistics collected and published by the Bangladesh Bureau of Statistics are primary data in the hands of the Bangladesh Bureau but those are secondary data to other agencies. The source from which the data are to be obtained has got a direct bearing upon the method through which they are to be collected. Accordingly, the method of collecting primary data differs widely from that of collecting secondary data.

Secondary data are those which are collected and published by various agencies for their own purpose but can be used by others also

Methods of Field Investigation for Collecting Primary Data

A good many methods of collecting primary data are found in use. The method chosen should be appropriate to the inquiry. In choosing a particular method of collecting primary data the objective of the survey as well as the time and cost involved should be considered. The important methods are:-

1) Interview by enumerators with a prepared schedule or questionnaire: Under this method the enumerator is provided with a prepared questionnaire and he puts the questions to the informant and records the answers. The informant does not fill in the schedule himself

but the enumerator fills it up. The enumerator needs to have clear understanding of the implication of each question and the way in which the information is to be sought and the mode of filling up the schedule. Obviously, this method needs qualified and trained enumerators. Much of the success under this scheme depends upon the standardization of the questions and the skill and tactfulness of the enumerators. If the problem of having qualified investigators can be overcome, this method provides quite a good result. Under this method, exhaustive type of questions can be included in the schedule or questionnaire, the scope of the survey can be enlarged and extensive investigation can be undertaken. Most of the research organizations undertake this method of investigation. In an extensive type of inquiry this is found to be the more suitable method of collecting information. In population census this method is inevitably used because of the vast size of the population and the nature of its composition.

2) Schedules to be filled in by the informants themselves: This is also called 'Mailing Method'. Under this method the questionnaires are sent to the individual respondents through mail with a request to fill them up and send them back to the researcher. Usually, stamped addressed covers are supplied to the respondents along with the questionnaire. Under this method, framing the questionnaire is very important. Questions included in the questionnaires should be simple, easy and self-explanatory. The nature of questions should be such that man with average intelligence can easily answer them. Usually, the answers to questions turn to be yes or no type and possible alternative answers are quoted in the schedule. This method is relatively less expensive. Information covering a population spread over a large area can be collected within a fairly short period of time and at a lesser amount of cost. This type of inquiry is undertaken by private agencies and sometimes by the government agencies too. But this method of collection suffers from a number of drawbacks. The success of the method depends upon the efficient preparation of the questionnaire as well as the responsiveness of the informants. Experience shows that a large number of informants do not care to return the schedule. Even if the questionnaire is returned, there is a chance of its being filled up incompletely and in a haphazard and cumbersome way. The possibility of misunderstanding of a question and wrong answering, purposively or ignorantly, cannot also be ruled out. Low rate of literacy prevailing in our country is also a barrier in adopting this method of investigation. Owing to these limitations this method has got limited use and is used mostly in the survey of opinion.

Questions included in the questionnaires should be simple, easy and self-explanatory.

3) Direct observation by enumerators: Under this method the enumerator is provided with a schedule incorporating information required and he goes to the field of observation and records the required information from his personal observation. He directly observes the phenomena and records the same. No one needs to be interviewed or questioned. The method is relatively simple but the reliability of the information collected depends upon the sincerity and diligence of the enumerator. The number of cars passing through a road or the number of

The method is relatively simple but the reliability of the information collected depends upon the sincerity and diligence of the enumerator.

shoppers visiting a shop in a particular day may be observed by posting an enumerator besides the road or the shop entrance.

4) Direct personal observation: Under this method the person interested in studying a problem undertakes the task of personally observing the phenomenon. In fact, the investigator might himself get to the spot, mix up with the people, may even live with them and from such personal knowledge he may make careful observations about the condition of the people in all aspects or the aspect in which the investigator is interested. This kind of first hand information is of immense value but this is possible only when the field of inquiry is not extensive. The chance of personal bias also cannot be ruled out altogether.

This method is used where information to be collected are of complex type and a direct approach for information may not produce the desired result

5) Indirect oral investigation: This method is used where information to be collected are of complex type and a direct approach for information may not produce the desired result. The investigator may have a standard list of questions and he interviews the persons well informed of the phenomena about which the information is sought. Their views are then taken as evidences for making inferences. Where the field of inquiry is very vast, for example, a whole nation, a good number of informed individuals may be called upon to give their views on the problem. This method is mostly undertaken by committees and commissions appointed by the government to investigate a particular problem. The scope of this type of inquiry is very limited. Its success greatly depends upon the capability and intellectual attainment of the interviewer. Only persons having sufficient knowledge of the problem and free from bias should be interviewed. A considerable care is needed in evaluating the evidences given by the interviewers. This method is not of much use in socio-economic surveys.

Under this method local correspondents are employed to give the estimate of a particular phenomenon

6) Information through local correspondents or sources: Sometimes local correspondents are employed to give the estimate of a particular phenomenon. They provide their own estimates in their indigenous way and these estimates form the basis of compilation. This method is mostly employed in crop estimates. This method has very little systematic basis and the information obtained is also not very reliable. However, when only an approximate idea is required, this method promptly provides information with a little cost.

Self-Assessment Questions:

Short Question:

1. Define Statistical inquiry
2. What do you mean by statistical unit?
3. Define source of data?

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Parking at a shopping centre has become a very big problem. Shop Administration are interested in determining the average parking time (e.g., the time it takes a customer to find a parking spot) of its customers. An administrator inconspicuously followed 290 customers and carefully recorded their parking time. Identify the data collection method used by the administration in this study.
 - (a) Data from a survey
 - (b) Data collection observationally
 - (c) Data from a designed experiment
 - (d) Data from a published source
- (ii) What method of data collection would you use to collect data for a study where political supporter wished determine if his candidate is leading in the polls?
 - (a) Use a survey
 - (b) Use a published source
 - (c) Take a census
 - (d) A designed experiment
- (iii) Which of the following data collection methods is most likely to generate the largest non-response?
 - (a) Mail survey
 - (b) Direct observation
 - (c) Telephone surveys
 - (d) Personal interviews
- (iv) Which of the following data collection method is most likely to be used to determine numbers of cars passing over the flyover in a day:
 - (a) Direct observation by enumerators
 - (b) Direct personal observation
 - (c) Information through local correspondents
 - (d) None of the above.
- (v) In developing and conducting a written survey, what is the purpose of the pre-test phase?
 - (a) To make sure that cost of developing the survey instrument is not too great.
 - (b) To generate initial data for analysis
 - (c) To catch any problems with the questionnaire before it is finally administered.
 - (d) To make sure that the respondents like the issues being addressed by the survey

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) Primary data are data that are collected for specific use.
- (ii) Typically, it is possible to include a large number of questions in a phone survey than in a mail survey since it takes less time to complete the survey over the phone.
- (iii) PQS Marketing Research Department recently conducted a survey of 9500 customers asking questions about such things as how satisfied the customer is with PQS’s service. The data collected in this survey would constitute primary data for the market research department.
- (iv) A class project calls for the students to access the data compiled from Bangladesh 2000 census. The data that the students would be using are considered primary data by the students.
- (v) Mail questionnaires typically generate poor response rate.

Answer:

Multiple-Choice Question:

1. (i)- b (ii)- a (iii)-a (iv)- a (v)-c

True/False

2. (i)- T (ii)- F (iii)- T (iv)- F (v)- T

Lesson 2: Framing a Questionnaire and Scrutinizing Statistical Data

Lesson Objectives:

After completing this lesson you will be able to:

- Describe the process of framing a questionnaire;
- Understand selection of investigations and their training;
- Understand scrutinizing the primary data;
- Explain collection of secondary data.

Introduction

The adoption of the first and the second methods of collecting data, namely, questionnaire filled in by enumerators and questionnaire filled in by respondents, inevitably requires the framing of a questionnaire. As has already been mentioned the framing of a suitable questionnaire is an essential requisite for successful application of both the methods. The framing of a questionnaire is an art and requires great care and skill. Extracting information from respondents is a difficult task as human beings have got their own value judgment. Success of investigation depends upon the skilled and tactful framing of the questionnaire. The framer should have a clear conception of the purpose and scope of inquiry as well as the detailed knowledge of the field to be covered by the inquiry. Each question is required to be carefully analyzed and its purpose clearly understood. Sometimes dummy tables are prepared to test the utility of each question and to help subsequent analysis. The suitability of the questionnaire to the respondents is mostly prejudged by undertaking pre-testing of the questionnaire before it is finalized. In the process, any misunderstanding in the questionnaire is removed and some difficult questions may be dropped altogether. In addition the points mentioned above the following points are to be kept in mind in framing a questionnaire:

- (i) The questions asked should be as few as possible. Unnecessary and irrelevant questions should be avoided. The number of questions should not be too many to cause irritation on the part of the respondent. It is to be kept in mind that people do not like answering too many questions or filling up a lengthy questionnaire.
- (ii) Questions should be such that they can be understood and answered by the least educated and intelligent respondents. This will, of course, depend upon the composition of the population. Difficult and intricate questions should normally be avoided.
- (iii) The questions should be carefully worded. They should be precise and unambiguous. Simple and appropriate wording should be used in the question. Any technical word included in a question must be subject to clear explanatory note. Words of colloquial origin in a region may be included in a question for better understanding. Questions should be precise, clear and devoid of any duplicity in their meaning.

The framing of a questionnaire is an art and requires great care and skill

The suitability of the questionnaire to the respondents is mostly prejudged by undertaking pre-testing of the questionnaire before it is finalized

Questions should be precise, clear and devoid of any duplicity in their meaning.

Questions should be such that they can be answered without any bias

Questions should be so framed that they will exactly cover the information required in the study

Questions should be arranged in logical sequence.

The success of investigation and the subsequent classification and analysis of information arising out of investigation, to a considerable extent, depends upon the intelligent framing of the questionnaire.

- (iv) Questions should be capable of being answered readily in a simple way. Answers to questions should be usually of 'yes' or 'no' type. Questions entailing the need of elaborate answering may irritate the respondent. To avoid the problem, sometimes-possible alternative answers are given with the questionnaire and the respondent is required to give his opinion in favour of either of the quoted alternative answers.
- (v) Questions should be such that they can be answered without any bias. The respondents might be biased in answering a question, which might vitiate the result. If a mother is asked about the misdeeds of her son, she, being affectionate to him, may not disclose the same. Unmarried girls are hesitant to state their actual age and might state a lower figure than the actual. All these are the cases of biased answering. Questions involving bias or prejudice in answering should be avoided as far as possible.
- (vi) In framing questions due regard should be paid to the religious, communal and political beliefs of the respondents. Too much private things should not be asked unless specifically required. Questions should not be unnecessarily inquisitorial.
- (vii) Questions should be so framed that they will exactly cover the information required in the study. Each question should have an objective and be clear in meaning. Questions are not to be too many but sufficient to cover the information required for the study. Simplicity must not be at the cost of coverage.
- (viii) There must not be anything in the question, which might give rise to suspicion in the mind of the respondents. Questions should be such that they are answered spontaneously and truthfully. If the respondent has got the slightest suspicion that the information might be used against his interest, he will be reluctant to provide the same.
- (ix) The arrangement of the questions in the questionnaire is also important. Questions should be arranged in logical sequence. One question should be corroboratory to the other. Easier questions should be asked initially. All questions relating to one aspect should be grouped under one heading. Tricky questions should not be put up simultaneously. Questions involving check should be put intermittently. Arrangement of questions should maintain the flow of thought in the mind of the respondent.
- (x) Overall design of the questionnaire should be nice and attractive. Sufficient space is to be provided for answers. Detailed instruction showing the way of filling up the questionnaire should accompany the questionnaire. When the questionnaire is to be filled up by the respondent, the way in which it is to be done should be included in their training. Many of the factors, connected with the framing of the questionnaire, although may sound to be of common sense type at the first look, yet have got profound significance. The success of investigation and the subsequent classification and analysis of information arising out of investigation, to a considerable extent, depends upon the intelligent framing of the questionnaire.

Selection of Investigators and their Training

When the investigators are to be employed for securing information, there is the need of selecting proper persons and their training. Selection of qualified investigators is very important for conducting the survey. It is needless to say that the quality of field force will largely determine the reliability of the information. They need to be diligent, intelligent, painstaking and truthful. Intelligence must be accompanied by integrity. Irresponsible and evasive investigator may not go to the field and may fill up the schedules sitting in his home with fictitious answers. This might completely distract findings of the survey. Therefore, in the selection of investigators care should be taken to avoid irresponsible elements. In certain type of investigations persons directly interested in the phenomena need to be carefully avoided from selection as they may indulge in biased reporting. Another requirement is that the investigators should have a polite and courteous manner.

Selection of qualified investigators is very important for conducting the survey

The selection of enumerators is to be followed by their intensive training. They must be given proper training involving all aspects of field operation before they are actually placed in the field. Whatever might be the background of the investigators selected, they are to be given knowledge of the new venture they are going to undertake. The purpose and scope of the survey should be explained to them. The purpose of putting each question should be explained to them. The purpose of putting each question should be very clear to them; otherwise proper information may not be secured. The manner in which the questions are to be asked and the manner in which the answers are to be recorded should also be clear. The units of measurement to be used and the manner of expression must also be clear to them. In short, all aspects of field operation should be covered by the training programme. After the theoretical part of the instruction is over, they should be sent for practical field training before they actually start the work. Sometimes they are attached to experienced field workers for training. This aspect of training is of much importance as the investigators get a chance of pointing out any aspect not covered by the theoretical training and of clearing up any misunderstanding regarding any question. In practice, prepared and detailed instructions illustrating the mode of recording answers to questions are supplied to the field investigators.

They must be given proper training involving all aspects of field operation before they are actually placed in the field.

Scrutinizing the Primary Data

The data collected through field investigation are to be subjected to a close, continuous and rigorous checking. This aspect is also regarded as the editing of collected information. A close check will reveal inconsistencies, possible errors and omissions in the schedule. Rigorous checking might reveal some internal inconsistency in the data. However, through editing it is difficult to verify the authenticity of the data unless there is apparent inconsistency. The plan for editing should be uniform and standardized to ensure uniformity in all schedules. Some mistakes may be corrected if there is any countercheck in the questionnaire. The investigator may also correct some anomaly if he is sure of it. Any deliberate correction must be done carefully. If required, defective and unsatisfactory schedules may be sent back to the field for correction. If

The data collected through field investigation are to be subjected to a close, continuous and rigorous checking.

Scrutiny of data is important before starting their classification.

the schedule is totally incomplete and thoroughly erroneous it may have to be rejected outright and another attempt may be made to find the information afresh. This may not be possible always. When such is the case, incomplete schedules are to be rejected. Scrutiny of data is important before starting their classification.

At the time of editing attempt is made to make all the schedules uniform in terms of units of measurement and mode of expression so that little difficulty arises in their classification. One investigator might express the quantity of the same commodity in tons, another in kilograms and still another in grams. All these expressions are to be standardized. If ton is used, then all should be expressed in ton or in part thereof.

Collection of Secondary Data

Data collected by one person or agency are used by another person or agency, these are termed secondary data.

When the collection of primary data is not possible, usual course is to make use of the data collected by others. When the data collected by one person or agency are used by another person or agency, these are termed secondary data. A careful review of the secondary data is needed before they are used. It is necessary to know the methods utilized in the original collection, the purpose for which the original collection was made, units used in the collection, reliability and representativeness of the data and degree of accuracy followed in the collection as well as the element of accuracy found in the presented figures. Each of these factors will have to be carefully evaluated before making a decision regarding the use of the same. Profitable use of secondary data can be made if they can be made to suit the purpose.

The collection of secondary data is relatively easier than collecting primary data. No elaborate arrangement for collection is needed.

The collection of secondary data is relatively easier than collecting primary data. No elaborate arrangement for collection is needed. The only important thing is the knowledge of their existence. As has already been told, secondary data are mostly available in published form and the publishers usually put up the various information relating to method of collection, purpose of collection, the composition of the population, level of accuracy maintained and so on in the publication itself. The users only collect the publication or publications and utilize the data contained in them for his own purpose. The sources of secondary data are mostly the publications of government and semi-government organizations, private agencies or individuals, international bodies and the various newspapers, journals, periodicals, etc.

Government and semi-government organizations, municipalities, local bodies and universities publish a lot of official statistics in report form.

Private agencies like business houses, trade associations, chambers of commerce, private research organization as well as individual scholars engaged in research work provide statistical information in published form.

International organizations like UNO, FAO, ECAFE, IBRD, IMF, UNESCO etc., have got their official publications containing valuable statistical information.

Trade journals, technical and financial journals as well as newspapers contain a good deal of statistical information.

The decision as to whether primary data or secondary data are to be used in an investigation is largely determined by the object and scope of the investigation and the availability of suitable secondary data. Time and cost have also got a determining effect upon the choice. Before starting the primary inquiry one should be sure that no original work has been done in this field which might serve his purpose. There is no point in undertaking primary investigation, which is costly as well as time consuming when suitable secondary data are already available. It may happen that secondary data source can only partly provide the desired information. In such a case the use of secondary data should not be ruled out; rather secondary data should be used as far as these can meet information requirement and for the rest of the information primary investigation should be conducted. In this way, in the same study, both primary and secondary methods of collection can be profitably used.

Before starting the primary inquiry one should be sure that no original work has been done in this field which might serve his purpose.

Self-Assessment Questions:

Short Question

1. What is degree of Accuracy?
2. Write one important character for training a questionnaire?
3. Define primary sources of data
4. Define secondary source of data
5. Write two example for secondary/primary source of data

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) If inaccuracies exist in the values of the data recorded, what is indicated?

(a) Nonresponse bias	(b) Unethical statistical practice
(c) Selection bias	(d) Measurement error
- (ii) A Company conducted survey of its employees to determine their level of satisfaction with various company policies. The data collected from this survey are:

(a) Primary data	(b) Secondary data
(c) Experimental data	(d) None of the above
- (iii) For which data collection method is it most important to have a polished looking survey form?

(a) Telephone survey	(b) Written questionnaire
(c) Experimental design	(d) Personal interviews
- (iv) Data that are collected from the entire population are referred to as:

(a) Primary data	(b) Secondary data
(c) A census	(d) A sample
- (v) The GMG Airlines Internet site provides a questionnaire instrument that can be answered electronically. Which of the four methods of data collection is involved when people complete the questionnaire?

(a) Published success	(b) Experimentation
(c) Surveying	(d) Observation

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) A Mobile Company recently met with a group of its customers to ask questions about the service and products provided by the company. The data collected in this process would be an example of data collected through direct observation.
- (ii) Analysis performed using secondary data is typically considered inferences or for the purpose of preparing business reports.
- (iii) On a survey, the questions pertaining to the background of the respondent (age, gender etc) are referred to as demographic questions.
- (iv) The method of data collection called direct observation is always associated with gathering data from people.
- (v) Recently, an analyst in a company’s marketing department surveyed customers regarding how often they buy a particular product. One customer indicated that she purchased the product 17 times in the last six months, but the analyst recorded the response as 71 times. This is an example of observed bias.

Answer:

Multiple-Choice Question:

1. (i)- d. (ii)- a (iii)-b (iv)- c (v)-c

True/False

2. (i)- F (ii)- F (iii)- T (iv)- F (v)- F

Lesson 3: Classification of Statistical Data

Lesson Objectives:

After completing this lesson you will be able to:

- Understand classification of statistical data;
- Explain the properties of good classification;
- Describe the basis of classification;
- Understand array formation;
- Define frequency distribution;
- Define variable and attribute;
- Construct frequency distribution of variables and attributes.

Introduction

Statistical data must undergo the process of classification and tabulation to form the basis of valid inferences. Once stages relating to a statistical inquiry have been satisfactorily arranged and sources of data is set, the next stages is the actual collection of the data. Collection of data is only an important first step in the attainment of the objective of an inquiry. This needs to be accompanied by the grouping of the data into classes and their condensation in tabular form. Even in case of secondary data it becomes necessary to rearrange and regroup them to suit the purpose for which they are to be used. If the collected information is huge covering a wide range of material as happens in case of population census, mechanised devices may be employed for processing the information. Whether the work is done manually or mechanically, the essential principles of classification and tabulation are the same, although these may vary in detail.

Classification is the process of arranging the data according to some common characteristics. The characteristics are inherent in the data. Through classification the collected data are arranged into groups or classes according to their different characteristics. An example will make the matter more clear. In population census information is collected on various characteristics of the population like age, sex, literacy, occupation, place of residence, religion, etc. These voluminous data so collected are to be properly classified before they are put to use. Grouping will be made in such a way that the data in the same group should be as homogeneous as possible. Large mass of the collected statistical information would never have been graspable to common mind had there been no procedure of arranging them in some logical order. Through the process of classification heterogeneous data are so divided that within the division or class they possess the same attribute or characteristics. L. R. Connor has defined classification as the “process of arranging data in groups according to resemblances and affinities”. The object of classification is to bring out clearly the diverse characteristics of data.

The main purpose of classification is to enable the investigator to have a clear picture of a phenomenon with less mental effort. The grouping permits the elimination of unnecessary detail. Classification brings into light similar and dissimilar points of view and clarify many confusing

Classification is the process of arranging the data according to some common characteristics.

The main purpose of classification is to enable the investigator to have a clear picture of a phenomenon with less mental effort.

situations. Through the process of logical arrangement of data, classification permits easy comparison between related facts. It also helps one to make correct observations and inferences. Another important purpose of classification is to make the data suitable for further processing and tabulation. Classification is, in fact, an important step towards tabulation. Heterogeneous data are hardly capable of tabular presentation. Data must be classed before starting tabulation.

Properties of Good Classification

In making classification care should be taken to see that the classification is exhaustive enough to include all the items in some class. Classification should not be made on a piecemeal basis. Such a situation will create confusion and ambiguity. Every item of information must be placed in some class. The classification should not be overlapping. The classes must be mutually exclusive. The basis of classification should be clearly understood by those making classification. The basis of classification should be sufficiently stable throughout the work without being unnecessarily rigid. Classification should be flexible enough to accommodate new situations but this should not be done at the cost of stability.

Care should be taken to see that the classification is exhaustive enough to include all the items in some class.

Basis of Classification

As has already been said classification is done on the basis of characteristics possessed by the individual units. Classification of the data can be done in numerous ways and it is difficult to draw precise rules for classification. However, the following four types of classification can be distinguished:

Classification is done on the basis of characteristics possessed by the individual units

Classification is done on the basis of the occurrences of the events within a certain period of time

[1.] Time may be logical basis of classification. It may be necessary to study the phenomena in terms of the time involved in their occurrences. For example, it may be desirable to classify the foreign trade data into different shipping periods. Internal trade statistics may be classified into weekly, monthly or quarterly figures. In other words, the classification is done on the basis of the occurrences of the events within a certain period of time. Cases involving comparison over time requires classification on time basis. Data classified and arranged on the basis of time form a time series. An example of time series is given below

Table 2.1 Time series showing the production of tea in Bangladesh during 1990-98

Year (July to June)	Production of tea (in thousand lbs.)	Year (July to June)	Production of tea (in thousand lbs.)
1990-91	49,001	1994-95	70,921
1991-92	46,470	1995-96	64,790
1992-93	53,000	1996-97	73,380
1993-94	60,000	1997-98	81,620

[2.] Space or geographical location may also be an important basis of classification. This refers to the distribution of items in physical space subject to artificial boundaries. Population census figures can be classified on regional basis like divisions, districts, sub-divisions, urban and rural, etc. This is an example of spatial classification. When an enquiry is made on international level, the data need to be classified on the basis of national boundaries. Trade statistics can also be classified on geographical basis. Data classified and arranged on the basis of space form a series known as *spatial series*. An example of spatial series is given below

Data classified and arranged on the basis of space form a series known as spatial series.

Table 2.2 Spatial series showing the net cropped area in different districts of Bangladesh in 1995-96

District	Net cropped area in thousand acres	District	Net cropped area in thousand acres
Dhaka	1,210	Dinajpur	1,105
Mymensingh	2,399	Rangpur	1,698
Tangail	599	Bogra	716
Faridpur	1,197	Pabna	883
Chattogram	736	Khulna	1,024
Chittagong hill tracts	175	Bakerganj	1,163
Noakhali	916	Patuakhali	650
Cumilla	1,230	Jessore	1,205
Sylhet	1,816	Kushtia	568
Rajshahi	1,678	Total	20,968

[3.] The data may also be classified on the basis of qualitative characteristics. Possessions of common characteristics or attributes constitute the basis of such classification. Items having common attribute in quality or condition are grouped together. The principle is that the like item should go with the like and the unlike with the unlike. The qualitative characteristics which are not directly measurable in numerical terms are called attribute. Attribute is the descriptive characteristic possessed by an individual that is not quantitatively measurable. Classification of data according to attribute is determined by the presence or absence of an attribute in the individual. Consideration of an attribute gives us two classes automatically, one possessing the attribute and the other devoid of the attribute. The population of Bangladesh may be classed on the basis of their literacy, so there will be two groups – literate and illiterate. Similarly, population can be classified into male and female, blind and non-blind, employed and unemployed, farmer and non-farmer, etc. In such a classification people having the common attribute are classed under one group. Each class turns to be exclusive of the other. This type of classification on the basis of one attribute constituting only two classes is called *simple or two-fold or dichotomous classification*.

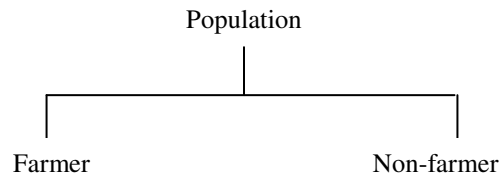
The qualitative characteristics which is not directly measurable in numerical terms are called attribute.

Classification on the basis of one attribute constituting only two classes is called simple or two-fold or dichotomous classification

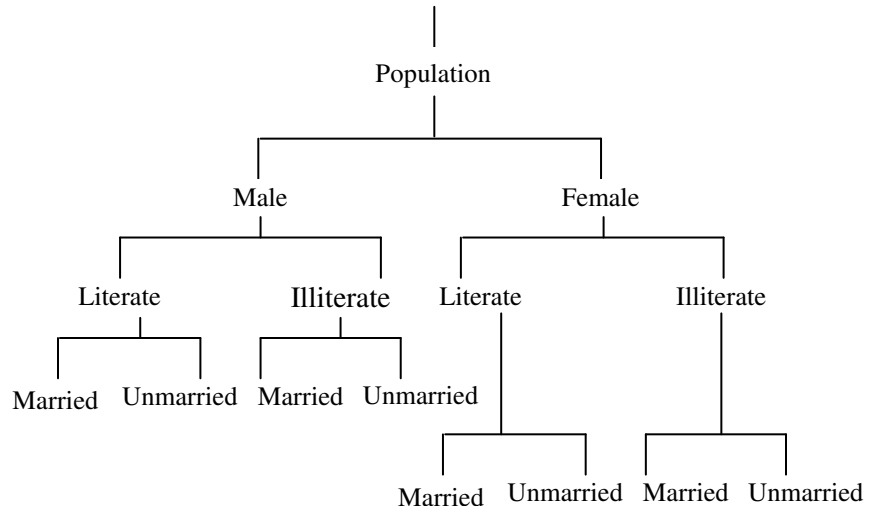
When classification is made on the basis of more than one attribute, several classes will come up. In our above example, if the population is divided not only on the basis of literacy but also on the basis of sex at the same time, then it will give rise to several classes such as male literate, male illiterate, female literate and female illiterate. The population can be further subdivided in terms of their occupation and religion. In this way the same population can be subjected to various modes of classification on the basis of different attributes. When the classification gives rise to more than two classes it is termed as *Manifold Classification*. In Manifold Classification the same population is divided and subdivided into various classes in terms of different attributes possessed by the population. The data classified according to attributes form a frequency distribution. Examples of simple and manifold classifications are given below.

When the classification gives rise to more than two classes it is termed as *Manifold Classification*.

Simple Classification



Manifold Classification



Statistics relating to height, weight, income, expenditure, production, etc. can be classified quantitatively. Classification according to quantitative characteristics is also called classification by variable.

[4.] Classification may also be made on quantitative basis where the phenomenon is quantitatively measurable. Non-descriptive characteristics come under this type of classification. This kind of classification rests upon some unit of measurement expressed in weight, volume, length, value, etc. The nature of the data will determine the type of measurement to be used in classifying the same. The distinguishing character between the different classes in this type of classification is not the attribute but the quantitative unit involved. For example, the export and import figures of Bangladesh can be studied or classified in terms of value or volume involved. In other words, the quantity of import and export forms the basis of classification. Statistics relating to height, weight, income, expenditure, production, etc. can be classified quantitatively. Classification according

to quantitative characteristics is also called classification by variable. The characteristic which can be measured numerically and the magnitude of which vary from individual to individual is called a variable. To be precise, the quantitative characteristics are termed a variable and when the data are classified on the basis of numerical measurements of the characteristics the classification is termed classification by variables. Here the classes are separated by numerical boundaries and therefore, they tend to be more definite than the classes formed by attributes. For example, we can divide the population on the basis of their age into two classes one having persons of age upto 18 years and another consisting of persons of more than 18 years of age. These two classes are more definite than the classes formed by attributes, e.g., boy and adult. In the same way, attributes like poor and rich, tall and short can be expressed in definite numerical terms by measuring income and height. It should be noted here that all characteristics are not amenable to quantitative measurement and that is why classification by attribute finds its place. Data classified either by variable or by attribute form the frequency distribution.

Classification according to quantitative characteristics is also called classification by variable.

The characteristic which can be measured numerically and the magnitude of which vary from individual to individual is called a variable

The types of classification dealt with above are by no means exhaustive. These four are the distinctive methods of classification. It is very difficult to lay down precise rules for classification. The mode of classification is, to a large extent, determined by the object of the inquiry and the nature of information collected. So the point to be emphasized here is that the mode and the method of classification should be decided upon before the actual tabulation work starts.

Array Formation

If the collected information is relatively few, a simple process is to arrange them in some logical order without involving the process of condensation. When the collected figures are put in an orderly manner by their magnitude we have an array. The data involved may be arranged either in ascending or in descending order of magnitude.

Illustration 2:1

Arrangement of the data in ascending as well as in descending order of magnitude.

493, 1847, 588, 1539, 955, 1004, 628, 1530, 457, 1640, 809, 1947, 918, 567, 1043.

The data arranged in *ascending* order of magnitude.

457, 493, 567, 588, 628, 809, 918, 955, 1004, 1043, 1530, 1539, 1640, 1847, 1947.

The data arranged in *descending* order of magnitude.

1947, 1847, 1640, 1539, 1530, 1043, 1004, 955, 918, 809, 628, 588, 567, 493, 457.

The mode of classification is, to a large extent, determined by the object of the inquiry and the nature of information collected.

In this way the data are to be brought to an understandable form although no assimilation is done. Only they are arranged in an orderly manner. This works quite well so long the data contain a small number of items. But when large mass of data is involved as happens in most statistical work, this cannot work. In such a case the array will be a huge one and no significant grasping of the data would be possible. So, there turns to be the need of condensing the data in some way.

Frequency Distribution

The limits or boundaries are called class limits.

The width of each class is called the class interval.

The number of items falling within the limits of a class interval is known as frequency of that class and is called class frequency.

Continuous variables are those which assume any numerical value within certain range.

The variable which cannot be expressed in every fractional value but is to be shown only in integral number is called a discontinuous variable

The variable which cannot be expressed in every fractional value but is to be shown only in integral number is called a discontinuous variable

The process of condensation starts with the grouping of the data in order of magnitude. The data are grouped by assigning some arbitrary limits or boundaries and putting the items falling within the range of the limits into the group. The limits or boundaries are called **class limits**. These are the highest and the lowest values of the class. These two limits are called upper limit and lower limit of the class. The lower limit indicates the lowest value that can be included in the class and the upper limit indicates the highest value that can come under the class. The width of each class is called the **class interval**. The number of items falling within the limits of a class interval is known as frequency of that class and is called class frequency. Frequency is, in general, the number of occurrences of the items. The arrangement of the data into class intervals showing the frequency of each class is known as frequency distribution.

Types of Variables

The variables are of two types- continuous and discontinuous or discrete. All variables are not subject to the same precision of measurement. Again certain phenomena are indivisible in nature and as such they are to be measured in terms of their number. Continuous variables are those which assume any numerical value within certain range. For example, income, age, production, birth rate, etc. can only be measured within a definite range and exact precision is difficult to attain. Continuous variable has an element of continuity, which are the individual values of the variable flow from one to the other continuously. Continuous variable or series takes the form of approximations and are shown within the range of certain limit. The variable which cannot be expressed in every fractional value but is to be shown only in integral number is called a discontinuous variable. As the item turns to be indivisible or discrete it is also called discrete variable. Discontinuous data are not subject to direct measurement but are to be derived by counting. Discontinuous variable or series is capable of exact measurement unlike the continuous series. Examples of discontinuous series are the number of persons in the family, number of employees in the factory, number of shops in the market and so on. In a discrete series there is no continuity in the flow of items. They constitute definite breaks between various items totally exclusive of each other. The example of continuous and discontinuous series is given below.

Continuous Series:

Table 2.3 Distribution of families according to the value of their dwelling houses.

Money value of dwelling house (in taka)	Number of families
Below 1,000	11
1,000 to below 2,000	127
2,000 to below 4,000	25
4,000 to below 6,000	3
6,000 to below 8,000	3
8,000 to below 10,000	2
10,000 to below 12,000	1
12,000 and above	3

Discontinuous Series:**Table 2.4 Frequency distribution of retail stores according to the number of salaried staff.**

Number of salaried staff	Number of retail shops
0	60
1	55
2	28
3	15
4	14
5	4
6 or above	13

As the item turns to be indivisible or discrete it is also called discrete variable.

Source: Retailing of Consumers' Goods in East Pakistan, Bureau of Economic Research, Dhaka University, 1965.

Construction of Frequency Distribution of Variables

Construction of frequency distribution involves certain steps like decision on the number of classes in which the data are to be divided, size or magnitude of class intervals, fixing up the class limits and so on. The step of construction of frequency distribution of variable are discoursed below:

1. **Number of classes:** An intelligent determination of the number of classes or groups into which the data are to be divided is an important task. The number of classes chosen should not be too many or too few. Too many classes will involve too much of detail working and simplicity of the grouping would be lost. Too few classes will be insufficient to reveal the characteristics of the data, as much of the information would be lost in the process. So the number of classes should not be too many but sufficient enough to unfold the characteristics of the data. A number of things need to be considered in determining the number of classes. It is necessary to know the number of items or units that are to be classified. The distribution of items has also got some affect upon the choice of the number of classes. The lowest and the highest values of the series show the range of the distribution. If the items show a tendency of concentration, then a small number of classes may be sufficient. In choosing the number of classes care should be taken to see that items with too wide gaps should not be included within the same class. Items with wide gaps, if included in the same class, will result in an unrepresentative mid-value. The number of classes into which the data are to be classified is again influenced by the objective of the study as well as the level of accuracy desired. Any distinguishing feature revealed by the data should also be considered in classifying them. No precise rule for determining the number of classes can be laid down. Numbers of classes are chosen by the statistician in each case keeping in view the points discussed above.

The lowest and the highest values of the series show the range of the distribution

2. **Class interval:** The size or magnitude of class interval is determined by the number of classes into which the data are to be divided and the range of the items constituting the data. The width of the class interval is its size or magnitude. That is, the difference between the lower limit and upper limit of the class is the magnitude of that class. If the class is 25 to 50 then the magnitude of the class is 25. What size the class will assume is determined by dividing the total range of the data (i.e., the difference between the lowest value and the highest value in the series) by the

Magnitude of the class is technically called class interval.

number of classes to be formed. For example, if the range of the data is 51 to 100 and if we decide to have 5 classes of equal size then we shall have 10 as the magnitude of each class. Magnitude of the class is technically called class interval.

3. **Class limits** - After deciding the class interval we are to fix up the limits of each class. In deciding class limits it is neither necessary nor convenient to start with the lowest figure found in the ungrouped data. Usually to equalize the classes some convenient value is taken as the lower limit of the first class. Similarly, some convenient value is taken as the upper limit of the last class. The values so chosen may not be found in the ungrouped data but are taken to secure computational advantages. The class limits should be so devised that the classes turn to be mutually exclusive. There must not be any confusion and doubt about the location of a particular item. The limits should not be overlapping and should clearly indicate the lower and upper ends of each class. If the data reveal any tendency of concentration at any particular value, the class limits should be designed in such a way that this characteristic is revealed. But it is difficult to observe such tendency in the ungrouped data unless there is some clear indication of the same.

The usual methods of assigning class limits are two – exclusive and inclusive.

The usual methods of assigning class limits are two – exclusive and inclusive. When the upper limit of one class constitutes the lower limit of the next class it is called exclusive method of assigning class limits. In this, upper limit of the first class is actually excluded from that class and included in the next class. In the same way upper limit of each class is excluded from the class and included in the next class. When this type of class limit is followed it is necessary to remove ambiguity by either putting a note below the table or by writing the class intervals in such a way that it is clear on the face. In contrast to this, the inclusive method of assigning class limits removes the ambiguity by actually placing the items. Here the classes turn to be inclusive of both the lower and the upper limits of the class. The value assigned to the upper limit of the first class is not taken as the lower limit of the next class. The lower limit of the next class starts with the figure next to the upper limit of the first class. In other words, both the upper limit and the lower limit of a class turn to be the property of the class. The examples of exclusive and inclusive methods of assigning the class limits are given below.

When the upper limit of one class constitutes the lower limit of the next class it is called exclusive method of assigning class.

The Inclusive method of assigning class limit the value assigned to the upper limit of the first class is not taken as the lower limit of the next class.

Table 2:5- Class limit (exclusive and inclusive methods)

Inclusive Method Frequency distribution of families according to size		Exclusive Method Frequency distribution of shops according to their monthly sale	
Family size	Number of families	Monthly sale in Tk.	Number of shops
1-2	23	0 and under 1,000	58
3-4	71	1,000 and under 2,000	49
5-6	111	2,000 and under 3,000	33
7-8	97	3,000 and under 4,000	10
9-10	58	4,000 and under 5,000	9
11-12	26	5,000 and under 6,000	9
13-14	17	6,000 and under 10,000	11
15-16	7	10,000 and under 20,000	6
17-18	9	20,000 and over	1

Illustration 2:2

The individual output of 60 female workers of an industrial firm in one week are given below:

501	520	534	549	557	542	555	548	542
535	523	524	536	536	559	561	552	543
526	511	525	536	542	548	507	523	538
540	555	579	575	553	547	539	533	519
519	532	571	539	537	545	551	580	588
509	521	548	527	522	548	538	522	552
549	544	537	545	526	508			

Present the above data in the form of a frequency table using class interval of ten units of output.

Solution:

Construction of frequency table showing the distribution of output of 60 female workers using tally marks.

Class Interval of output	Tally Marks	Frequency
501-510	III	4
511-520	III	4
521-530	III III	10
531-540	III III III	15
541-550	III III III	13
551-560	III III	8
561-570	I	1
571-580	III	4
581-590	I	1
Total		60

Frequency Distribution by Attributes

Distribution of individual items according to some attribute or category is called frequency distribution by attribute and the number of items conforming to a particular attribute is called the frequency of the attribute or class which means the number of occurrences of the attribute. For example, if 100 students are classified into male and female students and the number of male and female students found to be 70 and 30 respectively, then 70 is the frequency of male students and 30 is the frequency of female students. This distribution of students according to their sex into male and female groups is known as frequency distribution by attributes. An example of classifying raw data by attributes is given below:

Distribution of individual items according to some attribute or category is called frequency distribution by attribute.

Illustration 2:3

The records of occupation of 50 families are given below:

Service, business, profession, business, labourer, labourer, profession, service, service, labourer, labourer, profession, service, business, service, labourer, service service business, labourer, labourer, business, labourer, service, labourer, service, business, labourer, labourer, profession, service, labourer, business, service, labourer, business, labourer, labourer, business, labourer, profession, labourer, service, business, labourer, service, labourer, business, profession, labourer,

Construction of frequency table showing the distribution of 50 families according to their occupation

Occupation	Tally Marks	Frequency
Service	III III III	13
Business	III III I	11
Profession	III I	6
Labourer	III III III III	20
Total		50

Self-Assessment Questions:

Short Question

1. Define title
2. Write down about foot note of the table
3. What do you mean about tabulation
4. Define frequency distribution table
5. Define continuous variable
6. What do you mean by attributes.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) A postal workers counts the number of complaint letter received by the general post office service in a given day. Identity the type of data collected?
(a) Qualitative (b) Quantitative (c) None of the above
- (ii) Classify the color of automobiles on a used car lot as:
(a) Quantitative (b) Qualitative (c) None of the above
- (iii) Which of the following is a continuous quantitative variable?
(a) The color of a student's eyes
(b) The number of employees of a university
(c) The amount of milk produced by a cow in one 24 hour period
(d) The number of gallons milk sold at the local grocery store yesterday
- (iv) Quantitative variables classify individuals in a sample according to:
(a) Numerical measure (b) Physical attribute
(c) Exhibited trait (d) Personality characteristics
- (v) A student is asked to rate an instructor on a scale of 1-10 on the instructor's ability to teach. The student is to fill in a corresponding circle on a evaluation sheet. This is an example of collection what type of data?
(a) Qualitative (b) Inrightful
(c) Discrete (d) Continuous

2. Write "T" if the statement is true and "F" if the statement is false:

- (i) The sales data from a company measured weekly for the past year would be considered cross-sectional data since the sales values are computed from the entire company.
- (ii) The only absolute criteria that must be satisfied when constructing a frequency distribution where the variable is being grouped into classes is that the classes must be mutually exclusive.
- (iii) The upper and lower limits of each class in a frequency distribution are also referred to as the data arrange.
- (iv) Classification of data on the basis of one or more attribute constituting only two classes is called simple or two-fold or dichotomous classification.
- (v) Classification of data according to characteristics which can be measured numerically and the magnitude of which vary from individual to individual is called a variable.

Answer:

Multiple-Choice Question: 1. (i) b (ii) b (iii) c (iv) a (v) a

True/False: 2. (i) F (ii) F (iii) F (iv) F (v) T

Lesson 4: Tabulation of Statistical Data

Lesson Objectives:

After completing this lesson you will be able to:

- Understand the meaning of tabulation;
- Learn the different parts of a table;
- Learn the different types of table;
- Understand main consideration in the construction of a table;
- Describe need and importance of tabulation;
- Understand practical steps in tabulation;
- Understand forms of tables.

Introduction

Tabulation is the process of condensation. It is the systematic and orderly presentation of classified data in a definite form so as to elucidate the characteristics of the data.

In statistical tables the numerical information is presented in such a form that the information so presented turns to be readily understandable. Tables are designed to summarize facts revealed by enquiry and to present them in such a way that all the important factors contained in the data under review are displayed. Tables tend to simplify the presentation and facilitate comparison between related facts. Tabular presentation takes the form of arranging statistical data in columns and rows. The idea of a table will be clear if we look to the different parts of a table.

Parts of a Table

Tables contain certain parts. These are: -

a) **Title** – Each table has its title describing the contents. The title should be as precise and simple as possible. Title is the summary description of the contents of the table to guide the reader. It should not be too detailed but should be able to give idea of the contents of the table in a condensed form. Sometimes sub titles are also given under the table-head. This is something like a further clarification of the title. Sometimes explanatory notes are also given in the table-head to clarify any confusion likely to arise in the mind of the reader. As explanatory note is likely to make the table-heading cumbersome, it should be avoided as far as possible.

Title is the summary description of the contents of the table to guide the reader.

b) **Stub**- The stub contains two things – the stub head and the stub entries. The stub head actually describes the characteristics of the stub entries. The stub entries are the actual delimitation into which the data are classified. This is normally placed in the first column of the table starting from the left hand side.

The caption describes the data placed in each column of the table.

c) **Head or Caption** - The caption describes the data placed in each column of the table. A table may have many columns and each column is required to have a heading describing the things entered into the same. These are the column heads and the caption contains all the column heads that are found in the table. When sub-columns are used sub-heads are given to each.

Body of the table contains the statistical data in classified form

The source-note is important as this might permit the reader, if desired, to gather more information by going through the original data.

d) **Body**- Body of the table contains the statistical data in classified form. Enough space is to be provided for neat and tidy presentation of the numerical information. Arrangement of the items in the body of the table should be in some orderly manner.

e) **Footnote** - A footnote might be used to disclose anything which is not easily perceptible on the face of the table. The way precaution in the use of the table, any limitation or omission in the table, etc., are given in the footnote at the bottom of the table.

f) **Source-note** - If it is desired to disclose the sources of the information along with the table the usual method is to put source-note after the footnote disclosing the source of the information. The source-note is important as this might permit the reader, if desired, to gather more information by going through the original data. Source-note should be clear as to the actual source of the information.

g) **Number** - In practical work a large number of tables are prepared in the course of a study and to facilitate the identification of the tables it is usual to give a number to each table. The number may be put up above the title or below the source-note as is found convenient. In true sense the number does not constitute a part of the table.

The following illustration will make the various parts of a table discussed above more perceptible: -

Illustration of the different parts of the table

TITLE

Sub-Title

Explanatory Notes

Stub Head	Caption			
	Column Head		Column head	
	Sub Column Head	Sub Column Head	Sub Column Head	Sub Column Head
Stub Entries	B	O	D	Y

Footnotes

Source-note

Types of Tables

Tables are classified according to the purpose for which they are employed. Basically tables are prepared for two purposes. Tables may be prepared for reference or general use and accordingly they are named as reference tables. They do not endeavour to focus specific points but reveal the information in general. As they are primarily used as the source of information to others they should be constructed in such a way that the information can be extracted from them without much efforts. Reference tables are usually lengthy and put up in the appendix of a publication.

Tables can also be constructed to focus on a particular problem concerning the investigation. These tables are called text tables. Text tables are simpler than the reference tables and they are more analytical. Unlike the reference tables, text tables reveal only a specific prominent point or are designed to answer a specific question. Text tables may also present more than one fact but only a limited number of facts at a time. Text tables can be derived from the reference tables to suit the purpose of the individual researcher. This is done by picking up the needed information as well as by condensing and rearranging the data through averages, percentages and other devices. Studies involving specific problems mostly uses textual type of tabulation. Text tables mostly accompany the text of the report. The line of distinction between these two types of tables is only in terms of use and is not of much significance.

Text tables are simpler than the reference tables and they are more analytical

Main Considerations in the Construction of a Table

The construction of neat and informative tables is of particular need in every study. To form the basis for making inferences and valid conclusions tables must be well drawn up. It is difficult to give precise rules for construction of tables. Recognizing that the tabulation is an art requiring considerable skill and intelligence on the part of the tabulator the following general rules or considerations may be provided for guidance:

- 1) The table should be as simple as possible. Too much complexity might vitiate the purpose of tabulation. Too much of information should not be jumbled up in one table. The information should be spread up into several simple tables as far as the cost permits.
- 2) Before the actual preparation of the table it is advisable to have a rough draft of the table. This will give a clear idea as to the shape and size of the table as well as its overall appearance.
- 3) The table should be well balanced in appearance. The size of the different columns should be so adjusted that the overall length of the table is proportionate to its breadth. Columns should not be of varying length and breadth.
- 4) The units of measurements to be used in the table should be clearly stated and be uniform throughout a column.
- 5) The title of the table should be precise and be placed on the top of the table. The title should preferably be on the centre top of the table.
- 6) Directly comparable figures in different columns should be placed side by side to facilitate comparison.
- 7) Different types of letters or different types of ruling may be used to distinguish between figures as well as columns. In writing column totals different types of letters may be used other than the types used in writing the individual items. A column may be distinguished from sub-column by using heavy rulings in the case of former and light rulings in the case of latter.
- 8) Wherever possible figures may be rounded up convenient whole in order to avoid complications. This should be accompanied by a footnote.
- 9) The size of the table should not be too large. This makes the table unmanageable in one hand and difficult for the reader on the other. The size should be such that the reader can have a clear grasp of the whole thing at a glance.

The title of the table should be precise and be placed on the top of the table.

There should not be any ambiguity in the entry of the items in the table.

10) There should not be any ambiguity in the entry of the items in the table. The expressions should be clear. Indications like 'etc.' 'so on' should not be used. Abbreviation of words should be avoided as far as possible. Missing items should be clearly indicated as 'missing' rather than indicating them by zero.

The rules for tabulation are to be decided in each case in terms of the purpose of the inquiry and suitability of the data.

11) The stub entries are to be arranged in terms of the characteristics possessed by the data. The stub entries follow the classification of data in terms of space, time, quantitative or qualitative characteristics. The arrangement may be made in chronological, historical, conventional, progressive, alphabetical as well as in ascending or descending order. Any type of arrangement may be followed keeping the overall objective of the inquiry and the nature of information in view.

The guidelines outlined above may not have universal application nor these are all exhaustive. The rules for tabulation are to be decided in each case in terms of the purpose of the inquiry and suitability of the data. However, a tabulator's work would be much facilitated if he keeps the above guides in view while proceeding with tabulation work.

Need and Importance of Tabulation

Tabulation enables the numerical facts to be presented in such a way that their analysis, interpretation and subsequent computation becomes easier.

The need and importance of tabulation cannot be overemphasised. Tabulation enables the numerical facts to be presented in such a way that their analysis, interpretation and subsequent computation becomes easier. Decision-makers neither have the opportunity nor have enough time to go through bulky data. They want the information in a precise form so that, conclusions can be drawn from them without much wastage of time and energy. Tabulation is a useful tool in this respect. The condensed facts presented in table can be easily visualized and the needed information can be easily sorted out. The comparability of the data increases significantly when they are placed side by side in a table. This also helps the establishment of relationship between different phenomena. Tabulation paves the way for further condensation of the data by presenting them in suitable forms for mathematical treatment. Statistics is the study of large numbers. The study of a large number of cases is difficult unless some process of condensing the information is available. Tabulation provides a mechanism of condensation and thereby vitally contributes to the study of large numbers. Tabulation plays a crucial role in making the figures appealing and perceptible to the common mind.

Practical Steps in Tabulation

When information is collected through schedules three major steps in the construction of tables can be distinguished. These are:

Immediately after classification each item of information is extracted from the schedule as per the classification and placed in work sheets under appropriate class headings. This is the simple process of getting the information transferred from the schedules to the work-sheets to facilitate handling and proper itemization.

The next step starts with the summarization of the entries in the work-sheet. After summarization, totals are transferred to new sheets. These new sheets form the basis of preparing final tables. Grand totals for all the items are obtained in these summary sheets.

The last step is the preparation of final tables containing the results of the summary sheets. At this stage many of the unnecessary details are eliminated and only relevant figures are kept and presented in tabular form.

Forms of Tables

Forms of tables may be single, double, triple or manifold, according to the number of characteristics covered by the table. Practical illustrations will make the idea more clear.

Simple table shows only one characteristic. The data are presented only in terms of one of their characteristics. In two-fold table two characteristics are included. Similarly, manifold tables show many characteristics. Examples follow:

Forms of tables may be single, double, triple or manifold

Table 2.6 Simple table showing imports of Bangladesh during 1997-05

Year	Total Imports (Million Taka)
1997-98	1133
1998-99	1185
1999-2000	1181
2000-2001	1158
2001-2002	1732
2002-2003	2042
2003-20004	2470
2004-20005	2104

Table 2.7: Two-fold table showing the distribution of consumers according to their education and occupation

Education	Occupation					Total
	Fixed salary job	Busin ess	Profes sion	Wage earner	Others	
Illiterate	4	11	1	25	2	43
Upto class IV	6	22	2	13	2	45
Above class IV but not matriculate	14	30	2	2	2	50
Matriculate but not graduate	24	12	3	-	1	40
Graduate & above	8	3	7	-	1	19
Total	56	78	15	40	8	197

Source: *Retailing of Consumer Goods in East Pakistan*, published by Bureau of Economic Research, Dhaka University, 1965.

Table 2.8: Three-fold table showing the population by sex in Urban and Rural areas of Bangladesh in 1961 and 1974

Locality	1961			1974		
	Male (000)	Female (000)	Total (000)	Male (000)	Female (000)	Total (000)
Urban	1150	1090	2640	3539	2734	6273
Rural	24799	23400	48199	33533	31672	65205
Total	26349	24490	50839	37072	34406	71478

Source: *Bangladesh Population Census, 1974*.

Self-Assessment Question:

Short Question:

1. Define tabulation
2. Explain classification
3. What is body of the table?
4. Define footnote
5. Define caption

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) A table may have many
(a) Picture (b) Diagram (c) Column (d) Row
- (ii) The stub head usually describes
(a) Contents by the table (b) Characteristics of table
(c) Heading of table (d) None of the above
- (iii) The title of the table should be preferably placed
(a) On the centre top (b) On the top right
(c) On the top left (d) None of the above
- (iv) Forms of table may be
(a) Single (b) Double (c) Triple (d) All the above
- (v) Generally tables may be prepared to focus:
(a) Specific information (b) General information
(c) Detail information (d) All the above

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The caption describes the data placed in each column of the table
- (ii) Reference tables are not usually lengthy and not put up in the appendix of a publication.
- (iii) The line of distribution between reference table and text tables is only in terms of their use.
- (iv) Simple table shows more than are characteristics.
- (v) Text tables are simpler than the reference tables and they are more analytical.

Answer:

Multiple-Choice Question:

1. (i)- c. (ii)- b (iii)-a (iv)- d (v)-b

True/False

2. (i)- T (ii)- F (iii)- T (iv)- F (v)- T

Exercise

1. (a) What do you mean by statistical investigation? Discuss the different steps of statistical investigation.
(b) Write down the different steps of collection of primary data.
2. (a) What is the difference between primary and secondary data?
(b) Explain the terms related to primary data.
3. (a) Explain the method of data collection. Write down the steps used in collecting primary data.
(b) Define population, sample, sampling unit, questionnaire with examples.
4. (a) What are the different types of inquiry?
(b) Explain census method and sample survey method of data collection.
5. (a) Explain classification and tabulation. Write down the uses of classification and tabulation.
5. (a) What is the difference between classification and tabulation?
(b) Write down the advantages of classification and tabulation in statistical analysis.
7. (a) What do you mean by frequency distribution? Write down the procedure in preparing frequency table.
(b) The following are the systolic blood pressure (in mm Hg) of some patients visited in an out door of a hospital.

90	92	98	85	80	85	84	110	120	118	95	105	100	102	104	110	112
115	105	100	98	95	90	85	80	86	70	75	80	85	88	90	95	98
110	104	103	102	112	115	118	120	119	116	101	104	100	105	90	98	100
115	116	92	90	88	92	94	96	77	75	85	84	74	77	85	90	92
94	96	110	108	104	111	118	116	114	100	110	111	113	114	110	118	100

Prepare a frequency table using above data.

8. (a) Explain the principle in deciding number of classes, class limits, class boundaries, mid-value of a class.
(b) Write down the disadvantages in constructing frequency table.
9. (a) Explain the methods in constructing frequency table.
(b) The following data represent the temperature (in $^{\circ}\text{C}$) and humidity (in %) in different days of the year:

Temperature:	33.0	33.5	32.6	32.4	32.8	32.2	33.4	33.4	32.2	33.7	33.8
Humidity:	82	81	85	84	81	78	81	82	84	80	78
Temperature:	25.2	27.9	30.2	31.9	33.8	31.3	31.2	32.9	33.8	32.5	29.0
Humidity:	81	76	71	81	82	83	89	89	84	82	82
Temperature:	21.2	27.6	30.7	34.0	34.9	35.7	32.8	32.8	32.6	29.8	26.7
Humidity:	84	75	69	74	74	76	82	90	89	88	86
Temperature:	22.5	27.3	28.8	30.9	32.2	32.7	30.5	30.8	31.6	32.4	30.7
Humidity:	78	71	72	81	82	86	90	90	86	85	80

Construct a frequency table using above data.

REPRESENTATION OF STATISTICAL DATA



Row data does not provide any comprehensive idea about the population. However, preliminary inference can be drawn from classified data when it is presented in tabular form. Further comprehensive idea about the data and population can be obtained when the data are presented in graphs and diagrams. The graphical representation of data rendered comprehensive idea to these who are layman in statistical data analysis.

Lesson 1: Graphs for Representation of Statistical Data

Lesson Objectives:

After completing this lesson you will be able to:

- Understand graphical methods of representation of statistical data;
- Describe the advantages of graphical representation over tabular presentation;
- Describe limitations of graphical methods;
- Understand the procedures of construction of graphs.

Introduction:

Statisticians may find a table easy to understand and explain. But to a large number of people who use the data for making decisions, a table may only present mass of figures. Persons who are less apt with numerical figures find it difficult to appreciate the trends and significance brought by the tabulated data. This is true for most of the managers who are not expected to be trained in mathematics. All these points imply that there is the need of further condensation of statistical data and their consequent representation in a compact and understandable form. To accomplish this objective, graphic methods of representation of statistical data have been devised.

Graphical methods are the devices for representation of classified and tabulated data in a simple but effective form. Graphical methods do not add anything to the tabulated data but they simply illustrate them in a more compact and comprehensible form. Figures may be boring and when placed in complicated tables they are likely to be overlooked. Graphic representation makes the figures more appealing and attractive. Graphs permit presentation of a few facts at a time without much complication. Graphs not only appeal to the common men and decision-makers but also have got extensive appeal to the researchers. Researchers take resort to graphical methods to effectively portray their research findings. They find graphs useful for simultaneous comparison of several facts. The graphic methods aid researchers in making logical inferences from the data. In modern time, the use of graphs in socio-economic surveys, budget statements, and market research has increased manifold. Graphs present the facts in simple manner and convincing form. A complicated table, which may not be interesting to many turns to be lively when, presented in graphic forms. They also render comparison between related data easier by placing them side-by-side. Graphs, however, result in the loss of much detail, as it is the process of further condensation. Too much loss of detail may not be desirable always. Care should be taken to see that this condensation in graphic form is not carried too far.

Graphical methods are the devices for representation of classified and tabulated data in a simple but effective form.

Graphs present the facts in simple manner and convincing form.

Advantages of Graphic Representation over Tabular Presentation

Graphic representation has got certain distinct advantages over tabular form of presentation. The advantages are:

- a) Graphic presentation is more simple, appealing and attractive than tabular presentation. The significance and trend of the data are more easily revealed by graphs.

Graphs are important tools in visual analysis.

Graphs also help in forecasting without much mathematical computation.

b) Graphs are important tools in visual analysis. A glance to the graph makes the data understandable. Tables need thorough investigation before any valid conclusion can be made from them.

c) Decision-makers find the graphical presentation more suitable as they save considerable time and energy. Tables are laborious and time consuming for analysis. Moreover, interpretation of figures presented in tabular form presupposes some acquaintance with mathematics. But this is not absolutely necessary when the figures are presented in graphic form.

d) Graphs permit easy comparison between related facts than tables. When different facts are depicted on graph and placed side by side the facts turn to be easily comparable. But such facts in tables may not permit direct comparison and even if it is possible, it may turn to be arduous.

e) Certain statistical methods are more amenable to graphic representation. Interpolation, extrapolation, study of mass like median and mode, regression analysis are some of the examples of such statistical methods. Graphs also help in forecasting without much mathematical computation.

Limitation of Graphic Representation

Graphic methods have got their limitations too. Graphic representation is the process of further summarisation and as such a good deal of detail is lost in the process. It is difficult to achieve full accuracy in graphs. Graphs give a 'bird eye view' of the phenomena and as such depict only trends and fluctuations. Actual values are lost in graphic representation. Unfamiliarity with the graphic method on the part of the users may make the graphic method less attractive to many and may also make misleading study of the same. This defect cannot be ascribed to graphic method. Another such kind of defect, which may arise in graphical representation, is the manipulation – both intentional and unintentional. A manipulation in the base line or scale may give two different reading of the same phenomenon. As graphs result in the loss of detail it is possible for biased statisticians to present figures in the graph in such a way that they would give false impression. So the presentation of data in graphic form needs considerable care and caution. Presentation of data through graphs is an art and needs considerable skill and judgement on the part of the person constructing the graph. The benefit of simple presentation and clear understanding is to be carefully weighed with the loss of detail and possibility of false impression.

Graphs give a 'bird eye view' of the phenomena and as such depict only trends and fluctuations.

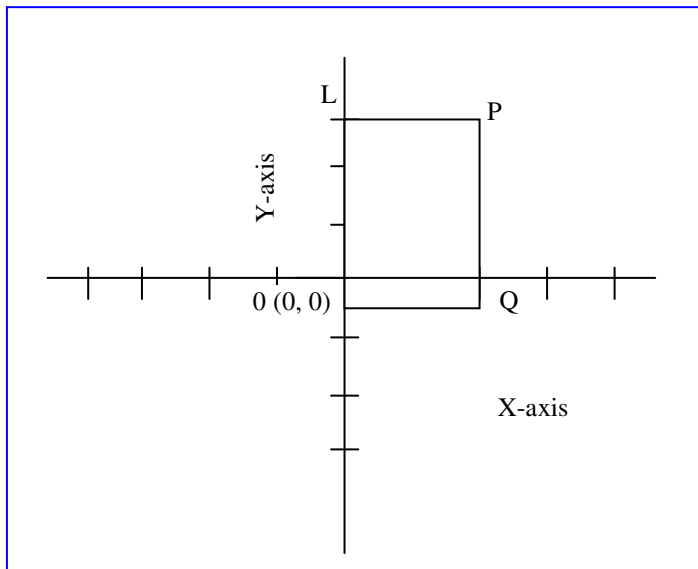
Concluding it can be said that skilled and unbiased graphical representation of data brings out the cardinal points underlying the data and makes them simpler and easier to grasp. There is an increasing tendency on the part of the statisticians to present numerical figures through graphs. Graphs have got wide acceptance in business and economic fields as a supplement to tabular presentation. Most of the business and economic reports, whether private or government, contain a good deal of graphs, charts, diagrams, pictograms etc., to attract the attention of readers, who may not like to get much into the tabulated data.

Graphs have got wide acceptance in business and economic fields as a supplement to tabular presentation.

Construction of Graphs

Graphs are usually drawn on two-dimensional plane. The structure for drawing graphs consists of two straight lines intersecting each other at right angles. These two straight lines are called axes – the horizontal line is termed as X- axis or axis of abscissa and the vertical line is termed as Y- axis or axis of ordinate. The point at which the two lines (X-axis and Y-axis) intersect each other is called the point of origin or zero point of the graph. The X-axis is taken as the line of origin for measurements along vertical direction, i.e., ordinate and Y-axis is the line of origin for measurements along horizontal direction, i.e., abscissa. The distance of any point to the right hand side from the Y-axis is taken as positive and the distance of a point to the left-hand side of Y-axis is taken as negative. Similarly, the distance of any point above the X-axis from the same is taken as positive and distance of a point below the X-axis is taken as negative. The axes divide the plane into four parts known as quadrants. A point in any of the four quadrants may be located with reference to two co-ordinates of the point drawn parallel to the axes of reference.

Figure 3:1



In fig. 1.0 XX' is the X-axis, YY' is the Y-axis and O is the point of origin. P is the point in the quadrant number I. PQ and PL are the two straight lines from the point P drawn parallel to X-axis and Y-axis respectively. The distance PQ is called x – co-ordinate or abscissa and the distance PL is called y co-ordinate or ordinate of the point P. According to the scale shown on the graph the x co-ordinate of the point P is 2 and the y co-ordinate of P is 3.

Usually a graph is drawn on a squared paper. The paper is ruled with horizontal and vertical lines intersecting each other perpendicularly. The scale is the unit of measurement, i.e., how many units are to be represented by a certain distance on the graph. Each square on the squared paper can be assigned with a scale value, e.g., one square may represent 5 units of a variable. A certain scale of measurement is to be decided upon by taking into consideration the size of the graph, the number of items or

When a graph is drawn on the assumption of equal distances meaning equal magnitude it is called arithmetic or natural scale.

Data classified in terms of space and attribute are more suitable for diagrammatic representation.

Data classified in terms of time and variable have been found to be most suitable for graphic representation.

Frequency graphs turn to be more attractive than frequency tables.

observations to be covered, the range of the variable and the object of presentation. In arithmetic line graphs it is desirable that the origin of the Y-scale should begin at zero and the X-axis should pass through the zero origin. Where it is not possible to start with zero as the origin, Y-scale may be broken. In such a case it must be clearly indicated by showing a break in the Y-axis. This will make the reader aware of this fact.

In graphical representation certain formalities are to be observed. The requirements regarding title, explanatory note, foot-note, source-note in graphs are almost the same as applicable to the preparation of tables. The title of a graph is required to be self-explanatory. Foot-notes are to be given where necessary. Since a graph is the illustration of tabulated data, the source of data is to be shown in the source-note. The factors represented by the X-axis and Y-axis are to be clearly indicated. The scale of measurement used along both the axes are to be similarly indicated. It is advisable to distinguish the axes from the rest of the graph by drawing them in bold lines. The resultant graph should also be distinctly visible. Too many things should not be put up in the same graph, as this is likely to make the graph cumbersome.

Plotting of an Arithmetic Scale Graph

When a graph is drawn on the assumption of equal distances meaning equal magnitude it is called arithmetic or natural scale graph. In an arithmetic scale graph, unlike logarithmic scale graph, equal magnitudes are represented by equal distance. This holds good both for ordinate and abscissa. It is advisable to use only one vertical scale for all curves presented in the same graph so as to make the comparison of fluctuations easier. The base line must pass through the zero origin of the Y-scale and the base line form the basis of measuring distances along the Y-axis. Like any other type of graph in arithmetic line graph the two axes represent two different variables. When two variables are involved it is customary to use X-axis for representing independent variable.

Having this preliminary knowledge of constructing graphs we can now proceed with the actual plotting of statistical data. It has already been observed that statistical data can be classified in terms of time, space, attribute and variable. Data classified in terms of space and attribute are more suitable for diagrammatic representation. Data classified in terms of time and variable have been found to be most suitable for graphic representation. Data relating to time form a time or historical series.

Frequency Distribution Graphs

The frequency distribution can be presented graphically and thereby the advantages of graphic representation can be secured. Frequency graphs turn to be more attractive than frequency tables. They reveal the distribution neatly and also indicate the points of concentration and deviations in the distribution. Frequency graphs bring out the shape and pattern of the distribution more clearly than the frequency tables.

The frequency distribution can be graphically represented in absolute, percentage as well as cumulative forms. The first two types of presentation, i.e., absolute and percentage forms give rise to histogram, frequency polygon and frequency curve. Plotting frequencies in cumulative form gives us cumulative frequency curve, which is also known as 'Ogive'.

Types of Graphs:

The following graphs which are used for presentation of statistical data

- (i) Histogram (ii) Frequency polygon
(iii) Frequency curve (iv) Ogive

(i) Histogram

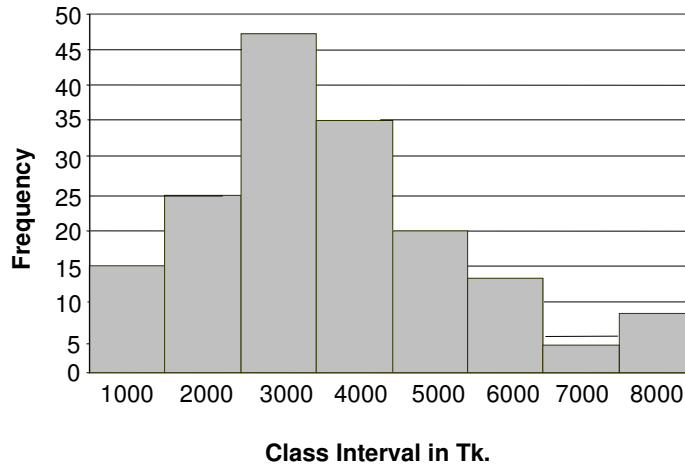
Histogram represents the frequencies corresponding to each class in a frequency distribution by vertical rectangles. The X-axis represents the class intervals of the variable and the frequencies of the class intervals are represented along the Y-axis. The scale in the X-axis is divided into as many columns as there are the class intervals in the frequency distribution. The breadth of each column shows the magnitude of class interval. These columns form the vertical rectangles. If the class intervals are of equal size the breadth of the rectangles will be of equal size. The varying class interval will result in varying breadth of rectangles. The Y-axis represents the class frequencies and the height of the rectangle will be determined by the frequency of the class represented by the rectangle. The Y-axis must start with the zero origin and unlike the temporal graph, is not amenable to any break of scale in the Y-axis. The scale chosen along the Y-axis must be able to accommodate the highest class frequency in the given frequency distribution. The X-scale need not start with the zero origin but it is convenient to leave some space between the first rectangle and the Y-axis. This means that the scale on the X-axis should be so ascertained that some gap is left before the first rectangle is plotted in order to distinguish it from the Y-axis. If we plot the data in this way we shall get a number of rectangles the breadth of which will show the magnitude of class interval and the height representing the class frequencies. The rectangles are attached to each other to give a continuous picture. The combination of all the rectangles constitutes a histogram. The total area of a histogram represents the sum of frequencies spread over different classes; to be more precise, area of each rectangle represents the frequency of the corresponding class. In other words, the area of each rectangle will be proportional to the frequency of the class represented by the rectangle. Example of a histogram representing the frequency distribution having continuous and equal class intervals is shown in the table 3.1 and Figure 3.2 respectively below:

Histogram represents the frequencies corresponding to each class in a frequency distribution by vertical rectangles.

Table 3.1 Frequency distribution of farm families according to the value of per acre farm production

Value of per acre farm production in taka	Number of farm families
0 to below 1000	17
1000 to below 2000	25
2000 to below 3000	47
3000 to below 4000	35
4000 to below 5000	20
5000 to below 6000	13
6000 to below 7000	4
7000 to below 8000	8

Fig. 3.2 Histogram representing the frequency distribution of farmers according to per acre farm production



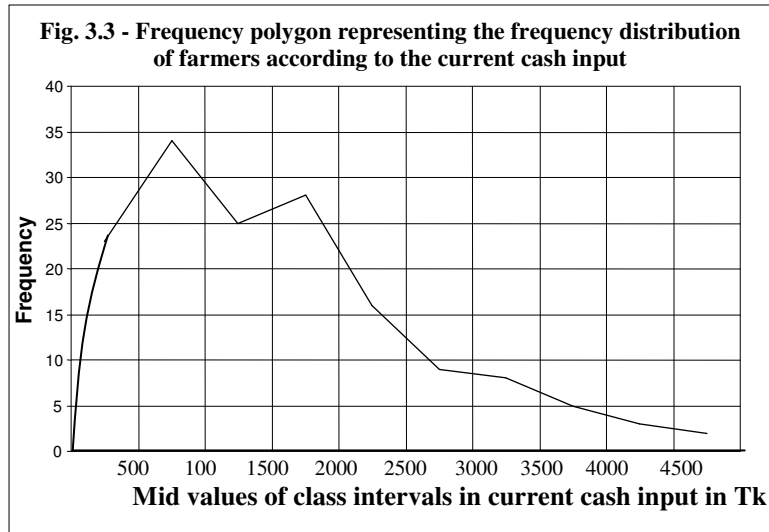
(ii) Frequency Polygon

A frequency polygon is a graph representing frequency distribution

Frequency distribution by variable can be represented by frequency polygon. The Y-axis represents the frequency. The mid-points of class intervals are taken along the X-axis. The class frequencies are plotted against the mid-values of the respective class intervals and then the plotted consecutive points are joined by straight lines. The figure so drawn is the frequency polygon. The height of each point along the vertical axis represents the frequency of the corresponding class. It is sometimes convenient to add one class at each end of the graph to make the graph complete. Both of these classes have zero frequency and as a result both ends of the polygon will be joined with mid-points of the extended classes on the X-axis. An example of frequency polygon is shown in figure 3.3 which is drawn from table 3.2 below:

Table 3.2 The frequency distribution of farms according to current cash input in Taka

Current cash input (Tk.)	Mid-value	Number of farms
0-500	250	23
500-1000	750	34
1000-1500	1250	25
1500-2000	1750	28
2000-2500	2250	16
2500-3000	2750	9
3000-3500	3250	8
3500-4000	3750	5
4000-4500	4250	3
4500-5000	4750	2

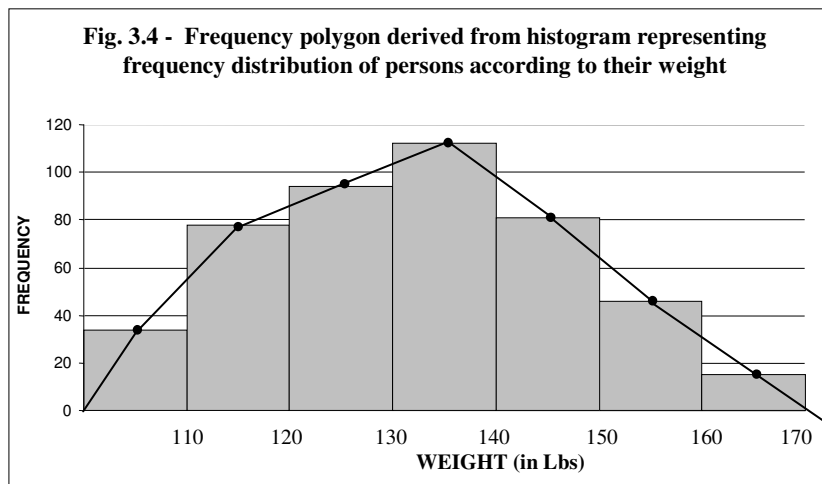


Derivation of Polygon from Histogram

A frequency polygon can also be derived from the histogram. Under this method mid-points of the rectangles at the upper end are determined and then these mid-points of the adjacent classes are connected by straight lines. The figure so drawn gives us a frequency polygon derived from histogram. In this case also both the ends of the polygon are extended to the mid-points of the extra classes taken on both sides to complete the figure. An example of a frequency polygon derived from histogram is given in Fig. 3.4.

Table 3.4 The frequency distribution of adult people according to their weight

Weight in lbs.	Number of persons
100-110	34
110-120	78
120-130	94
130-140	112
140-150	81
150-160	46
160-170	15



(iii) Frequency Curve

We have seen that both the histogram and frequency polygon produce graphs which are not smooth. In a frequency distribution if the number of observations is large then the number of classes can be increased so as to make the magnitude of class intervals smaller and smaller. And in such cases the graph representing the distribution will approach a smooth curve. In a histogram the area of each rectangle is proportional to the frequency of its class. If we decrease the size of the class intervals then the breadth of the rectangles will also reduce gradually. If the availability of data permits the continuation of this process then we find that the width of the rectangles is greatly diminished and the tops of the rectangles in the histogram tend to form a smooth curve. The same is true in case of a frequency polygon also. Such a curve is called frequency curve. The ordinate of the curve at any point on the X-axis is proportional to the frequency of the value at that point. Area under the curve between any two ordinates is proportional to the frequency of observations between the two values corresponding to the ordinates. Thus the ordinates as well as the area under the curve represent the frequency of the distribution.

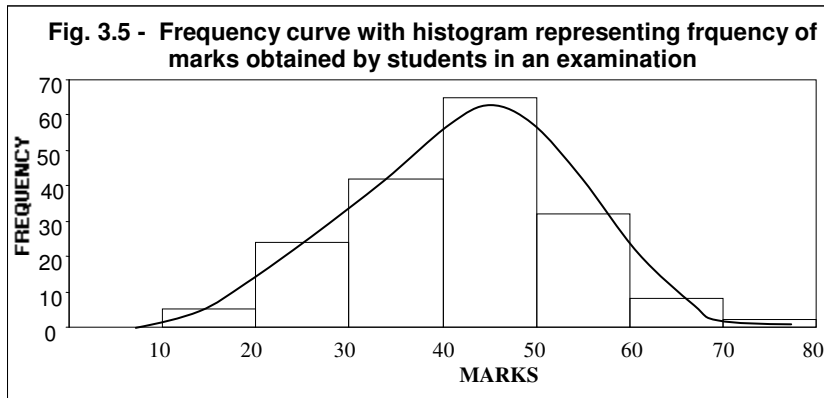
The concept of frequency curve is of particular importance in statistical theory. In practice, an accurate frequency curve is very difficult to draw. In real life data there is not likely to be enough observations to permit increase in the number of classes by reducing the magnitude of the class intervals. Theoretically, it is possible to think of a smooth distribution with the help of mathematical equation. But for practical data we endeavour to achieve smooth pattern of distribution by what is called free hand method.

In case of frequency polygon similarly a free hand curve is drawn through the plotted points in such a way that the curve drawn turns to be smooth.

Initially, histogram or frequency polygon is drawn to represent given frequency distribution. In a histogram or frequency polygon it can be observed that the upper ends of the rectangles of the histogram and the plotted points in a frequency polygon show irregularities which do not permit them to form smooth curve. To smooth out the irregularities a free hand curve is drawn through the upper ends of the rectangles of the histogram. In case of frequency polygon similarly a free hand curve is drawn through the plotted points in such a way that the curve drawn turns to be smooth. In both the cases the assumption is that the area under the frequency curve will be equal to the area of the histogram or the polygon. The drawing of the free hand curve requires great skill since any irregularity in the smoothing will give erroneous presentation. The curve should be as smooth as possible removing all turns. The extent to which the distribution will be smooth will depend on the nature of the data. The natural phenomena may be quite amenable to smooth distribution while economic phenomena are not likely to be so. No set rule can be given for drawing a free hand curve. The only check is that the area under the curve must represent the total frequency in this distribution. An example of frequency curve is given in Fig. 3.5.

Table 3.5 The frequency distribution of students according to marks obtained in an examination

Marks	Frequency
10-20	5
20-30	24
30-40	42
40-50	65
50-60	32
60-70	8
70-80	2



Shape of Frequency Curve

Frequency distribution may take different forms and shapes. Accordingly frequency curves also take different forms and shapes. Four of them are easily distinguishable. A distribution may be symmetrical curve. In this type of distribution highest frequency is shown in the central class and the frequency decreases to the left and the right of the central class in an identical manner and both the extreme ends represent zero frequency. An ideal symmetrical curve is rare to find in business and economic data. A particular symmetrical curve is called 'normal curve'. The distribution may be moderately asymmetrical and the curve representing this type of distribution will give us moderately asymmetrical curve. This is also called moderately skewed curve. In this type of curve the peak does not represent the central class of the distribution. Frequency concentrates either at the right or the left of the central class and accordingly peak of the curve will be either to the right or the left of the centre and one of the tails of the curve will be extended more than the other. Economic and social phenomena are usually found to take the form of moderately asymmetrical curve. The distribution may be extremely asymmetrical distribution. This type of distribution produces J-shaped curve. This is found particularly in wealth distribution. A distribution may have high frequency at both the ends and low frequency at the centre of the distribution. This type of distribution will give us U-shaped or V-shaped curve. It is rare to find this in practice. Examples of the above types of frequency are given in Fig. 3.6 below.

A particular symmetrical curve is called 'normal curve'

Asymmetrical distribution produces J. shaped curve.

Figure : 3.6 Different types of frequency curve

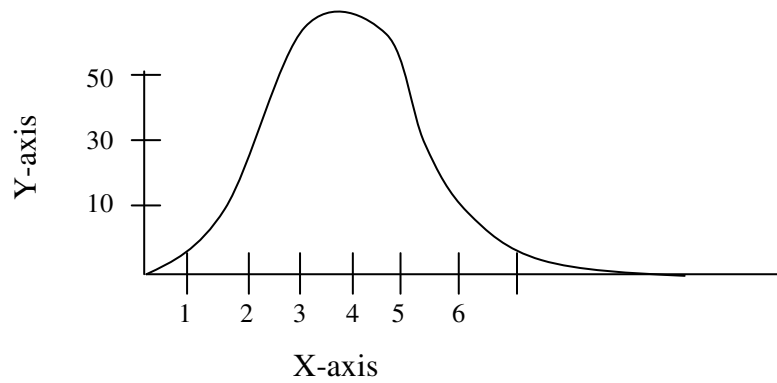


Fig 3.6 (a): Positively skewed curve

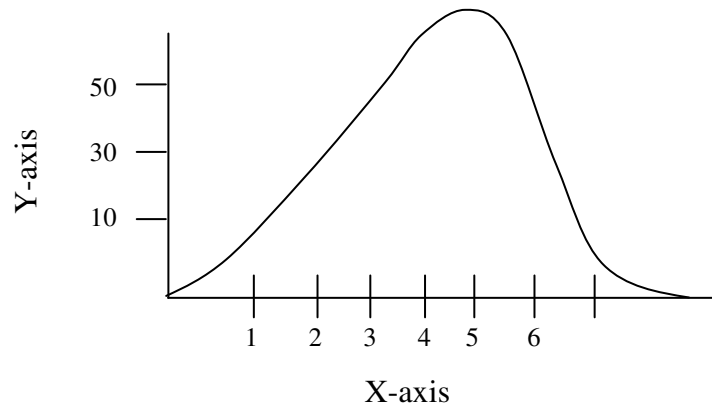


Fig 3.6 (b): Symmetric curve

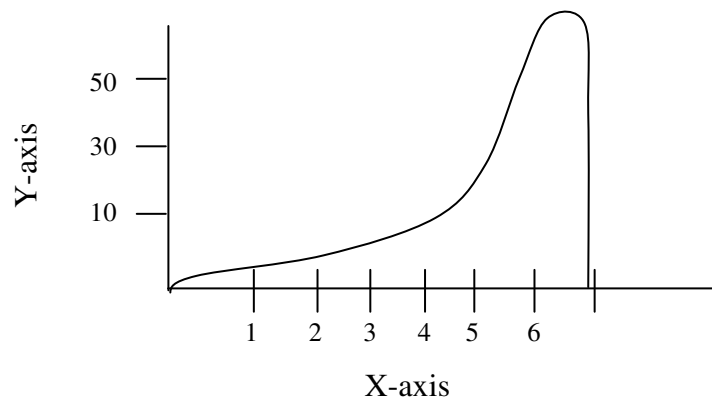


Fig 3.6 (c): Negatively skewed curve

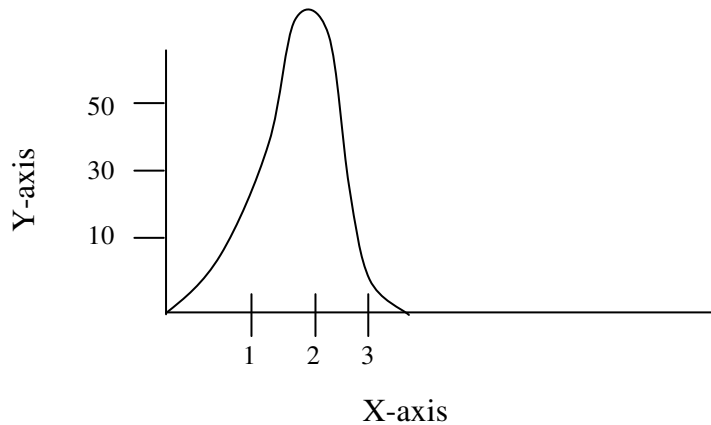


Fig 3.6 (d) Bell skewed curve

(iv) Ogive

In cases of histogram, frequency polygon and frequency curve we have used class frequencies as they appear. In case of ogive the class frequencies are not plotted as it is but are cumulated before they are plotted. The frequency of one class is cumulated with the frequency of the next class either in ascending or in descending order and then the accumulated class frequencies of that class and all the previous classes. The graphical representation of the cumulative frequency distribution is known as ogive or cumulative frequency curve. The method of construction is more or less the same as in the case of other types of frequency graphs. The upper or lower limits of class intervals are taken along X-axis. The Y-axis represents the cumulative frequency. The cumulative frequencies in ascending order are plotted against the upper limits of the class intervals and the cumulative frequencies in descending order are plotted against the lower limits of the class intervals. The plotted points are then connected by drawing lines through them. The curve so drawn is called ogive. The graph representing ascending cumulative frequency is called 'less than' ogive and the graph representing descending cumulative frequency is known as 'more than' ogive. Example is given in Fig. 3.7 below.

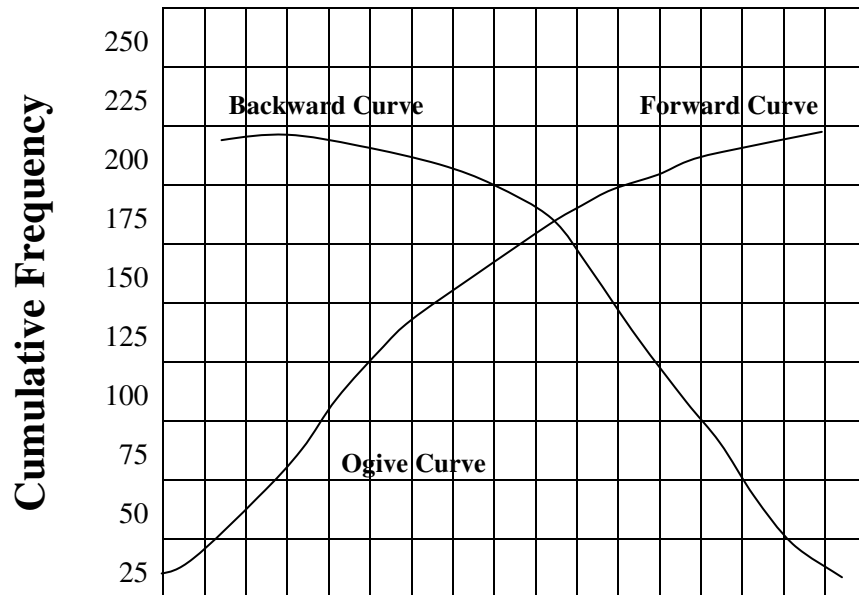
Table 3.6 - The distribution of employees of a factory according to their age and cumulative frequency

Age group in years	Number of employees (frequency)	Forward cumulative frequency (More than method)	Backward cumulative frequency (More than method)
15-20	7	7	200
20-25	20	27	193
25-30	55	82	173
30-35	41	123	118
35-40	27	150	77
40-45	21	171	50
45-50	18	189	29
50-55	10	199	11
55-60	1	200	1

The graphical representation of the cumulative frequency distribution is known as ogive or cumulative frequency curve.

The graph representing ascending cumulative frequency is called 'less than' ogive and the graph representing descending cumulative frequency is known as 'more than' ogive.

Fig. 3.7 - Ogive representing cumulative frequency of employees in different age groups. (Less than ogive representing forward cumulative frequency and more than ogive representing backward cumulative frequency)



Ogive is most suitable for comparison between the two related distributions. Ogive is drawn to locate certain measures such as median, quartiles, percentiles, etc., graphically. It enables us to find out how many observations are within or outside certain limits. Cumulative frequency distribution having unequal class intervals can also be represented by ogive. If the distribution has open end class intervals then it is not advisable to draw ogive.

Self-Assessment Question:

Short Question:

1. Define Histogram?
2. Write down about frequency polygon?
3. What do you mean by ogive
4. Distinguish between Graph and Diagram.
5. Define frequency curve?

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) A histogram is most commonly used to analyze which of the following?

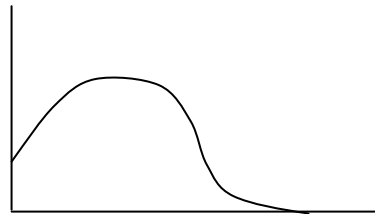
(a) Nominal level data	(b) Quantitative data
(c) Time series data	(d) Ordinal data
- (ii) Frequency distribution can be formed which of the following types of data?

(a) Both discrete and continuous	(b) Discrete only
(c) Continuous only	(d) Only numerical data

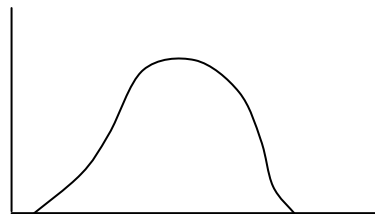
- (iii) The frequency distribution can be presented:
- (a) Pictorially
 - (b) Graphically
 - (c) Ordinally
 - (d) Sequentially
- (iv) A histogram used to display which of the following characteristics for a quantitative variable?
- (a) The approximately centre of data
 - (b) The spread in the data
 - (c) The shape of distribution
 - (d) All of the above
- (v) When polygon and histograms are constructed, which axis must show the true zero or “origin”?
- (a) The horizontal axis
 - (b) The vertical axis
 - (c) Both the horizontal & vertical axes
 - (d) Neither the horizontal nor the vertical axis
- (vi) The class width is the difference between
- (a) The upper class limit and lower class limit of a class
 - (b) Two regressive lower class limit
 - (c) The largest frequency and the smallest frequency
 - (d) The high and the low data values.
- (vii) For the data below, construct a frequency distribution using five classes. Describe the shape of the distribution:

3	6	7	6	0	6	1	7	8	4
1	5	7	5	9	1	5	3	9	9
2	2	3	0	8	8	4	0	2	4

- (a) Uniform
 - (b) Skewed to the right
 - (c) Symmetric
 - (d) Skewed to the lift
- (viii) Describe the shape of the distribution.



- (ix) Describe the shape of the distribution



- (a) Skewed to the left
- (b) Uniform
- (c) Skewed to the right
- (d) Symmetric

- (x) An ogive is a graph that represents cumulative frequencies or cumulative relative frequencies. The points labeled on the horizontal axis are the:

- (a) Lesson class limits (b) Midpoints
(c) Frequencies (d) Upper Class Limits

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The relative frequency is the frequency in each class divided by the total number of observations
(ii) When constructing a frequency distribution, classes should be selected in such a way that they are of equal width
(iii) If the values of the seventh and eighth class in a cumulative frequency distribution are the same, we know that there are no observations in the eight class.
(iv) Determining the class boundaries of frequency distribution in highly subjective.
(v) Frequency distribution by variable can be represented by ogive.
(vi) A frequency polygon can also be derived from the histogram.
(vii) Ogive are plotted at the mid point of the class groupings.
(viii) Frequency distribution are specifically for analyzing discrete data.
(ix) A histogram can be constructed for data that are either quantitative on qualitative.
(x) After developing a frequency distribution for a quantitative variable, a histogram can be developed with the histogram axis representing the values of the variable and the vertical axis representing the frequency of occurrence in each class on group.

Answer:

Multiple-Choice Question:

1. (i)- b. (ii)- a (iii)-b (iv)- d (v)-b (vi)- b (vii)- a (viii)- d
(ix)- a (x)- d

True/False:

2. (i)- T (ii)- T (iii)- T (iv)- T (v)- F (vi)- T (vii)- F (viii)- F
(ix)- F (x)- T

Lesson 2: Diagrams for Presentation of Statistical Data

Lesson Objectives:

After completing this lesson you will be able to:

- Understand about diagrams for presentation of statistical data;
- Learn about the consideration in the preparation of diagrams;
- Describe different types of diagrams.

Introduction

Diagrams are used to illustrate statistical data. They make the data more attractive and give a lasting impression in the minds of the viewers. One may not like to get into a table to bring out the significance of the data but he may be quite interested to look to a diagrammatic representation of the same data. Diagrams also leave much more effective and lasting impression in the mind of the reader than the numerical figures. They also make the data more readily intelligible to the common mind. When it takes pictorial form it turns to be eligible even to an illiterate person or to a person with little education.

Diagrams permit easy comparison between the comparable facts. Not only that they make the comparison easier, but also they help in analysis and interpretation of the data. Diagrammatic representation might bring out some inner significance of the data which may not be perceptible from the tabular presentation of the same data. They save time and make a huge mass of data perceptible at a glance. In brief, it can be said in general that the diagrammatic representation enjoys many of the advantages enjoyed by graphical representation which has been discussed in the first part of this unit.

Diagrams permit easy comparison between the comparable facts.

Diagrammatic representation also suffers from certain disadvantages. Biased people may use diagrams to prove their pet conclusions. If the diagrams are not carefully studied they may produce misleading impressions. Diagrams can only represent limited amount of information at a time. Too many things cannot be placed in the same diagram as it gives rise to complexity in understanding. Diagrams represent facts approximately and not in exact term. Diagram cannot be thought of as substitute of tabular presentation but as a complement to the same.

Diagrams represent facts approximately and not in exact term.

Many of the points discussed here cannot be said to be the incurable limitations of diagrammatic representation. Many of the limitations arise due to the improper use of diagrams. A neatly drawn diagram can overcome many of the limitations. The advantages of diagrammatic representation far outweigh the disadvantages and the method can be safely recommended as a supplement to the tabular presentation in the study of different socio-economic phenomena.

Considerations in the Preparation of Diagrams

A diagram must be neat as well as scientific. To make the diagram scientific it must be drawn in strict compliance with scale. Choosing a proper scale is very important. The nature and the size of the data will

In the presentation of diagram various kinds of colours, dots and lines should be used to make the various component parts.

mostly influence the choice of scale. The scale should be suitable to make the diagram reasonable in size, neat and distinct in appearance as well as sufficient enough to accommodate all the observations. The scale should be such that all the significant characteristics are incorporated in the diagram. A diagram is to be drawn with the help of drawing instruments and in actual drawing no tempering with the scale is to be allowed. In the presentation of diagram various kinds of colours, dots and lines should be used to make the various component parts distinctive. When possible the actual numerical figures from which diagram are drawn should accompany the diagram. All formal requirements like title, footnote, scale, label, etc., in case of diagram are the same as those for the graphic representation.

Different Types of Diagrams

Diagrammatic representation can assume different forms. They are initially classified on the basis of their dimension. Accordingly there are diagrams having one dimension, two dimensions as well as three dimensions. The other two types found in use are the cartograms and pictograms. Out of these one-dimensional diagram is the best for accurate visual judgment and three-dimensional diagram is least accurate in this respect. The terms diagram, chart, curve and graph are often loosely used without making much distinction between them. For this section we shall mostly use the words diagram and chart interchangeably.

Diagrams with One-Dimension

One-dimensional diagrams are made of bars. A bar is nothing but a thick line. The width of the bar is not taken into consideration but is shown to make the diagram nice in appearance. Only the height of the bar is taken into consideration and the height represents the data. One-dimensional diagram may be of various types as described below.

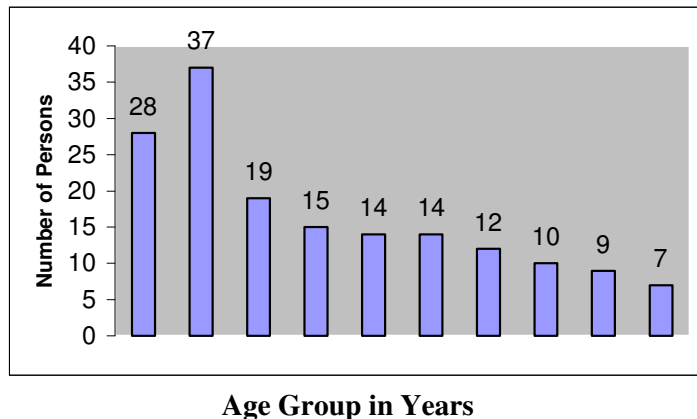
The diagram in which the values of a variable corresponding to different time period or regions or the frequency in different classes are represented by means of vertical bars is called a bar diagram.

(i) Simple Bar-Diagram

The diagram in which the values of a variable corresponding to different time period or regions or the frequency in different classes are represented by means of vertical bars is called a bar diagram. The values of the variable or the frequencies are represented along the vertical axis and the horizontal base represents the time or region or attribute or the variable as the case may be. The scale in the vertical axis is taken in such a way that the highest frequency or the highest value in time series or spatial series is covered. The size of the diagram is adjusted to the highest bar. The height of each bar is proportionate to the frequency or the values in the series. The width of the bars should be equal and gaps should be left between the two bars to make them distinct. Bar diagram can also be drawn on the vertical base in which case the horizontal axis would represent the values for different factors or frequencies of different classes. Bar diagram is found most suitable for representing time and spatial series and frequency distribution of attributes. The following Fig. 3.8 shows the example of simple bar diagram on the horizontal base prepared from Table 3.6 below:

Table 3.7 - The distribution of people according to their age

Age groups in Years	Number of persons
0 – 4	28
5 – 9	37
10 – 14	19
15 – 19	15
20 – 24	14
25 – 29	14
30 – 34	12
35 – 39	10
40 – 44	9
45 – 49	7
TOTAL	165

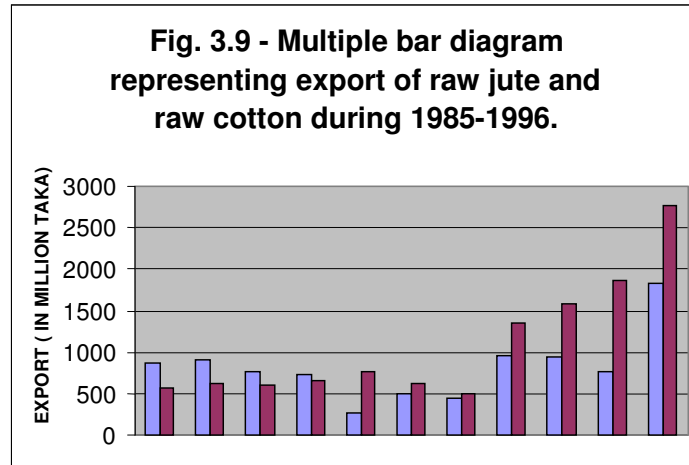
**Fig. 3.8 Bar Diagram showing the distribution of people according to age****(ii) Multiple Bar Diagram**

In multiple bar diagram two or three simple bars are drawn side by side without leaving any gap and each of them is differently coloured or shaded. Each of the bars represents different phenomena relating to a particular time but is shown together. Multiple bar diagram is drawn when it is desired to compare different phenomena relating to the same period of time. Example of a multiple bar diagram is given in Figure 3.9 prepared from the Table 3.8.

In multiple bar diagram two or three simple bars are drawn side by side without leaving any gap and each of them is differently coloured or shaded.

Table 3.8 - The export of raw jute and jute goods during 1994-2005

Year	Export of Raw Jute (in million taka)	Export of Jute Goods (in million taka)
1994-95	863	565
1995-96	898	626
1996-97	769	606
1997-98	731	656
1998-99	262	768
1999-2000	501	627
2000-2001	447	501
2001-2002	967	1353
2002-2003	940	1586
2003-2004	757	1859
2004-2005	1829	2778



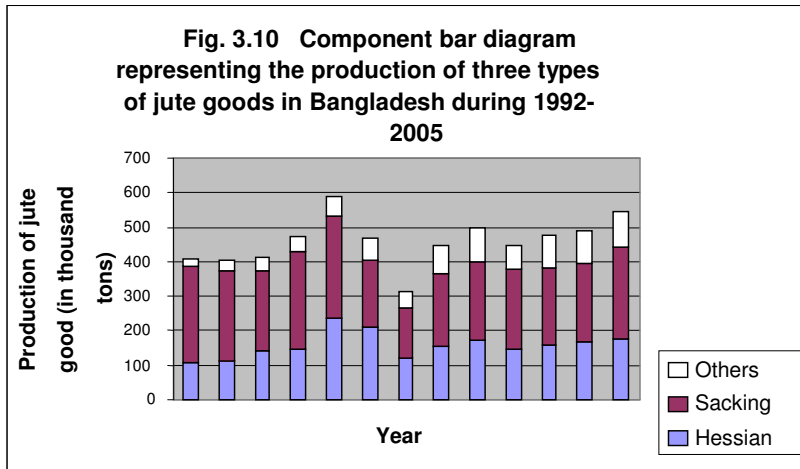
(iii) Component Bar Diagram

The component bar chart may be drawn on the basis of absolute figures as well as percentage figures

In a component bar diagram the total values as well as the various components constituting the total values are shown. The bar is subdivided into as many parts as there are components. Each part of the bar represents each component while the whole bar represents the total value. The component parts are variously coloured or shaded to make them distinct. The component bar diagram permits the comparison of the same factor over different periods of time not only in total but also in terms of the components of the total. The component bar chart may be drawn on the basis of absolute figures as well as percentage figures. The component bar diagram may be drawn to show difference in figures like birth rate and death rate, imports and exports etc. The component bar diagram showing absolute values of production of jute goods is illustrated in Fig. 3.10, prepared from the table 3.8.

**Table 3.9: Production of jute goods in Bangladesh
During 1992-2005**

Year	Production of jute goods (000 tons)			
	Total	Hessian	Sacking	Others
1992-93	409	106	279	24
1993-94	404	110	264	30
1994-95	512	140	234	38
1995-96	473	144	285	44
1996-97	587	235	297	55
1997-98	470	211	191	68
1998-99	315	121	146	48
1999-2000	446	155	210	81
2001-2002	500	172	227	101
2002-2003	446	148	228	70
2003-2004	477	161	221	95
2005-2006	490	166	227	97
2006-2007	546	177	265	104

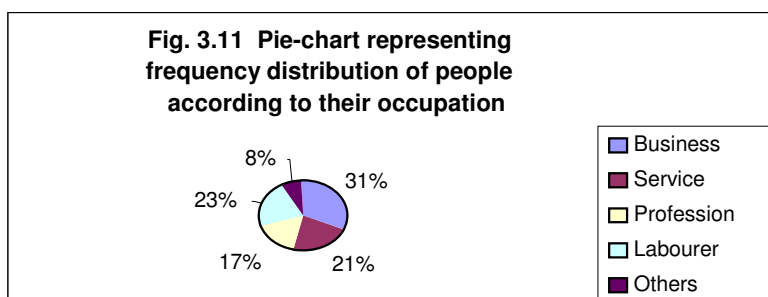


(iv) Pie Diagram

A Pie-Diagram is a circle subdivided into parts to reveal the various components of the data represented in the diagram. This permits comparison between the different components of the same phenomenon. A circle has 360 degrees at the centre, which is divided into sections in proportion to the components of the data. Representation of data by pie-diagram is illustrated in Fig 3.11 from Table 3.10

Table 3.10 Distribution of people according to their occupation and corresponding degree of angles in a pie chart.

Occupation	Number of persons	Percentage of persons	Degree of angles in pie-chart
Business	127	31.8	114.3
Service	82	20.5	73.8
Profession	68	17.0	61.2
Labourer	92	23.0	82.8
Others	31	7.7	27.9
Total	400	100.0	360



Pie-diagram can also be used to represent the different phenomena to make them readily comparable. As the diagram reveals the components it permits the comparison between the components of different phenomena as well as the comparison between the different phenomena as a whole. The area of the circle permits comparison between the component sectors of the phenomena.

Self-Assessment Question:

1. What do you mean by diagram?
2. Define pie diagram.
3. Define multiple bar-diagram.
4. What do you mean by bar diagram
5. What do you mean by one-dimensional diagram?

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) What is the differences between a bar diagram and a histogram?
 - (a) The bars on a bar diagram do not touch while the bars of a histogram do touch.
 - (b) The bars in a bar diagram are all the same width while the bars of a histogram may be of various widths.
 - (c) The bars in a bar chart may be of various widths while the bars of a histogram are all the same width
 - (d) There is no difference between these two graphical displays
- (ii) One characteristic of a bar diagram is:
 - (a) The bars can be displayed either vertically or horizontally
 - (b) There can be no gaps between the bars
 - (c) It is used to display the distribution of a continuous variable
 - (d) Both b and c are correct.

- (iii) Each year advertisers spend millions of taka purchasing commercial time on network sports television. A recent survey listed top 10 leading speeders over a 6 months period.

Company A	Tk. 72.0
Company B	63.1
Company C	54.7
Company D	54.3
Company E	29.0

Which the following could not be used to graphically display the data?

- (i) Pie Chart
 - (ii) Stem Display
 - (iii) Scatters Plot
 - (iv) Histogram
- (iv) The width of each bar in a histogram corresponds to the:
 - (a) Differences between the boundaries of the class
 - (b) Number of observations in each class
 - (c) Midpoint of each class
 - (d) Percentage of observations in each class.
 - (v) A bar diagram is most likely used to display with of the following?
 - (a) A continuous variable
 - (b) A nominal level variable
 - (c) An ordinal level variable
 - (d) Either b or c

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) One drawback of pie chart, dot plots, and histogram is that no measure of reliability can be attached to a graph.
- (ii) Histograms can have gaps between the bars, whereas bar charts cannot have gaps.

- (iii) Bar diagrams can typically be formed with the bars vertical or horizontal without adversely affecting the interpretation.
- (iv) A histogram is used to analyze a single quantitative variable while the bar diagram can display the results of multiple variables simultaneously.
- (v) The height of each bar is proportionate to the frequency or to the values of the series.
- (vi) Apple Computers collected information on the age of their customers. The youngest customers was 12 and oldest was 72. To study the distribution of age among its customer, it is best to use a pie chart.
- (vii) The TN Company monitors customers complaints and organizes there complaints into six districts categories. Over the past year, the company has received 534 complaints. One possible graphical method for representing there data would be a histogram.
- (viii) One of the difference between a bar chart and a histogram is that a bar typically displays data in percentage form.
- (ix) When developing a bar diagram, it is usually preferable to organize the bars in order from high to low.
- (x) One of the advantage of a pie chart is that it clearly shows that the total of all the categories of the pie adds to 100%.

Answer:

Multiple-Choice Question:

1. (i) a (ii) a (iii) a (iv) a (v) d

True/False:

2. (i) T (ii) F (iii) T (iv) T (v) T (vi) F (vii) F
(viii) F (ix) F (x) T

Exercise

1. (a) What is the necessity of graphical representation of statistical data?
(b) Discuss the methods of presentation of statistical data by graphs and diagrams.
2. (a) Discuss the necessity of graphs and diagrams in statistical analysis.
(b) Explain the advantages and limitations of diagrammatical presentation of data.
(c) Represent the following data by an appropriate diagram.

Division	Number of words	Number of household
Rajshahi	137	383000
Barisal	46	83000
Khulna	96	295000
Dhaka	201	1067000
Chittagong	103	445000
Sylhet	25	37000

3. (a) What is the use of histogram?
(b) Write down the differences between histogram and bar diagram, frequency polygon and frequency curve.

4. (a) The following data represent the number of road accidents in different

5 15 12 18 25 10 87 30 42 10 15 18 17 20 22 25 20 21 23 30 35
 5 10 8 7 6 15 18 21 25 2 33 37 36 30 32 34 37 28 29 27 26
 21 19 20 15 16 14 13 9 12 11 18 17 21 23 24 26 32 33 10 35 12
 10 8 38 30 40 19 18 22 25 26 28 23 16 17 15 12 18 24 10 16

- (i) Construct a frequency distribution of the number of road accidents.
 (ii) Draw histogram, ogive and frequency curve to represent the data.
5. (a) Discuss the uses and limitations of graphs and charts used in statistics.
 (b) Explain different graphs usually uses to represent statistical data and compare those.
6. (a) Explain the differences between graphs and diagrams.
 (b) Explain frequency curve and ogive. How does these two differ?
 (c) What is the difference between frequency polygon and frequency curve? Explain some important frequency curves.
7. The following data represent the weights of some fishes caught from a pond.

1.5	2.8	4.6	5.7	2.3	1.7	1.8	1.9	2.0	2.2	2.4	3.5	3.6	4.2	5.0	5.8	4.0
3.6	4.2	1.8	2.0	2.5	2.6	2.7	1.8	2.5	2.6	3.4	3.5	3.0	3.0	2.9	2.0	2.5
3.0	3.5	1.6	1.0	1.5	1.0	0.8	0.7	0.6	1.0	1.8	1.6	1.7	1.8	1.0	1.0	1.8
4.2	4.0	1.0	1.2	1.4	0.6	0.7	0.5	0.6	0.7	0.8	1.7	2.2	2.4	3.0	3.0	3.6
4.2	4.3	1.9	2.2	3.6	3.0	3.8	3.0	2.2	2.5	2.6	2.7	2.8	2.0	2.0	2.0	3.0

- (i) Prepare frequency table with the data of weights of fishes.
 (ii) Draw a frequency curve of the data.
 (iii) Find number of fishes with weights less than 2 kg.
 (iv) Find number of fishes of weights 3 kg and above
 (v) Also draw an ogive of the data.
8. (a) Discuss different graphs used in statistics to represent statistical data.
 (b) The following data represent the production of jute goods

4.5	6.2	8.7	10.4	10.6	11.2	15.7	5.6	6.0	6.2	6.0	8.0	8.5	9.7	10.0
5.6	7.5	7.0	10.6	9.8	8.7	8.0	7.0	7.3	4.2	4.6	5.0	5.8	5.9	6.0
9.7	8.2	8.6	10.0	9.6	9.0	9.8	6.5	6.0	4.0	5.8	7.6	8.2	9.0	10.2
4.8	5.0	6.2	6.0	6.7	6.8	7.2	7.0	7.6	8.0	8.2	9.7	9.0	10.4	8.3
7.6	7.8	7.0	6.4	6.2	7.2	7.0	6.2	6.0	5.9	6.0	9.0	9.8	9.2	9.3

- (i) Draw a frequency distribution table.
 (ii) Draw Histogram and origive curve.

MEASURES OF CENTRAL TENDENCY



In the previous unit, collection of data and presentation of data were discussed. As a next step we can think about to study the characteristics of data, so that they can be compared with each other.

It can be observed that the values of a statistical variable have a tendency to cluster around certain point usually at the centre of the series and most of the values in the series deviate from the central point with certain degree of regularity and smoothness. This tendency of the values to cluster around a certain value is called central tendency. An average is frequency referred to as a measure of central tendency on central values. This unit is dividend into four lessons. The first lesson will concentrate on arithmetic mean as a measure of central tendency and other lessons will discuss in detail other methods of central tendency.

Lesson 1: Measures of Central Tendency

Lesson Objectives:

After studying this lesson, you will be able to explain

- Explain what is meant by central tendency;
- Explain the importance central tendency;
- Know about various measures of central tendency;
- Understand arithmetic mean and its type;
- Calculate arithmetic mean;
- Understand advantages, disadvantages, uses of arithmetic mean.

Introduction

The numerical measure of tendency is the measure of location or the average. This central value usually divides the series into two parts one containing the values less than this value and another containing the values more than the central value. Such a central value becomes representative of the distribution only because of the tendency of the values to concentrate around this value; otherwise, central value would not have any significance. Such a representative value, free from irregularities, permits easy comparison between the two or more distributions. The central value not only permits comparison between the two distributions but also permits comparison within the distribution. The individual values in the distribution can be compared with respect to the central value.

The measure of central tendency gives us a value, which is typical or representative of the entire distribution. In other words, the central value describes the quantitative characteristics revealed by a mass of data. The central value, which is also known as average, gives a precise and definite idea of a large group in terms of a single typical value. Through the measures of central tendency or averages it is possible to set aside abnormal fluctuations that are bound to arise in practical data and to find out a standard value devoid of the irregularities but representative of the data. This central value best describes the characteristics of the distribution than any other value of the series. The most important objective of the measures of central tendency or averages is to have a concise picture of a large mass of data in a few representative numerical expressions and to facilitate comparison between the two or more groups of data thereby. The measure of central tendency by permitting simplification and condensation of statistical data makes a positive and significant contribution towards study of large numbers.

The central value, which is also known as average, gives a precise and definite idea of a large group in terms of a single typical value

Measures of Central Tendency

The following are the different measures of central tendency:

1. Mean:

- (i) Arithmetic mean
- (ii) Geometric mean
- (iii) Harmonic mean

2. Median

3. Mode

4. Quadratic mean

The arithmetic mean is the central value of the items in a series

Arithmetic Mean

The arithmetic mean is also referred to as the average or simply as the mean. The abbreviation of arithmetic mean is 'A.M.' The arithmetic mean is the central value of the items in a series. It is obtained by dividing the total value of the items in a series by the number of items. Let the daily wages received by seven industrial workers be Tk. 30,40,45,50,55,60,70. Then the mean daily wage of those workers would be,

$$\text{Mean} = \frac{30 + 40 + 45 + 50 + 55 + 60 + 70}{7} = \text{Tk.50 per person}$$

If we express these wage figures algebraically by $x_1, x_2, x_3, x_4, x_5, x_6$ and x_7 and arithmetic mean by $(\bar{x})$ and the number of wage earners by n , then the above example can be algebraically expressed as:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7}{n} = \frac{\sum x_i}{n}; i = 1, 2, \dots, n$$

The arithmetic mean may be of two types: (a) the simple arithmetic mean and (b) the weighted arithmetic mean. In the simple arithmetic mean each item in the series is counted only once while in the weighted arithmetic mean each item is assigned some weight in proportion to its importance.

Simple Arithmetic Mean

In a simple arithmetic mean each item in a series is counted once only

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}; i = 1, 2, 3, \dots, n$$

Where $\bar{x}$ = the arithmetic mean

x_i = the value of i th item

n = number of values

Illustration 4.1

Monthly sales of a shop are given below in taka for 12 months. Calculate of the mean of monthly sale.

Monthly sale in thousand Taka

36	62	49	75
50	63	55	53
48	47	61	42

Solution:

$$\begin{aligned}
 \text{Mean} = \bar{x} &= \frac{\sum_{i=1}^n x_i}{n}; i=1, 2, 3, \dots, n \\
 &= \frac{36 + 62 + \dots + 42}{12} \\
 &= \frac{641,000}{12} = \text{Tk.}53,410. \\
 \therefore \text{Mean} &= 53.410 \text{ Tk.}
 \end{aligned}$$

Computation of Arithmetic Mean from Grouped Data

In frequency distribution we have the class intervals and class frequencies and we are to deal with them. The class interval represents certain range of values and the mid-point of the class interval represents the class itself. The mid-point of the class is taken as the representative of the class on the assumption that the values are evenly distributed within a class. For computational purpose we need to have the sum of the individual values, which is to be divided by the number of values. This total value we strive to have by multiplying the mid-value of each class by the frequency of that class and then adding together the resultant products. This total is divided by the number of items, i.e., total frequency in the distribution. The resultant quotient is the arithmetic mean. The formula for computation of arithmetic mean from grouped data by direct method is given in the next page.

Where, x_i = The mid-value of i th class
 f_i = The frequency of i th class
 n = The total frequency
 $\bar{x}$ = The arithmetic mean.

Illustration 4.2

The table showing the frequency distribution of 185 families according to their size.

Family Size	Number of Families
2	3
3	9
4	25
5	62
6	55
7	16
8	7
9	6
10	2

Solution:

Calculation of arithmetic mean

Family Size, X_i	Frequency, f_i	$f_i X_i$
2	3	6
3	9	27
4	25	100
5	62	310
6	55	330
7	16	112
8	7	56
9	6	54
10	2	20
Total	185	1015

$$\text{Arithmetic Mean} = \bar{x} = \frac{\sum f_i X_i}{n} = \frac{1015}{185} = 5.49$$

Illustration 4.3

Calculation of arithmetic mean from the following frequency distribution

Weekly wage in Taka	Number of Workers
50-60	7
60-70	25
70-80	76
80-90	32
90-100	17
100-110	12
110-120	3

N.B. Lower limit excluded.

Solution:

Calculation of A.M.

Weekly wage in taka	Frequency f_i	Mid value X_i	$f_i X_i$
50-60	7	55	385
60-70	25	65	1625
70-80	76	75	5700
80-90	32	85	2720
90-100	17	95	1615
100-110	12	105	1260
110-120	3	115	345
Total	172		13650

$$\text{Arithmetic Mean} = \bar{x} = \frac{\sum f_i X_i}{n} = \frac{13650}{172} = Tk. 79.36.$$

The Weighted Arithmetic Mean

In a simple arithmetic mean each item in the series is regarded as of equal importance. But the items in a series do not always carry the same

importance. In a series having items of varying importance it is necessary to multiply each item by a 'weight' proportion to its importance in the series to have a representative mean. The products are then added together and divided by the total of weights. The result so obtained is called weighted arithmetic mean. Thus the formula for weighted mean is

$$\bar{x} = \frac{\sum wx}{\sum w}$$

Where, w = The weight of each item
 x = The value of each item
 wx = $w_1x_1 + w_2x_2 + \dots + w_nx_n$
 W = $w_1 + w_2 + \dots + w_n$
 $\bar{x}$ = The weighted arithmetic mean

Illustration 4.4

Calculation of mean proportion of literate in the following cities of Bangladesh.

Cities	Percentage of literate (5 years & above) x	Total Population (000) w	wx
Dhaka	46.9	557	26123.3
Mymensingh	59.3	53	3142.9
Saidpur	41.8	60	2508.0
Barisal	49.3	70	3451.0
Comilla	49.9	54	2694.0
Narayanganj	41.5	162	6723.0
Chittagong	48.0	364	17472.0
Khulna	46.1	128	5900.8
Total		1448	68015.6

$$\therefore \text{Weighted Arithmetic Mean} = \bar{X} = \frac{\sum wx}{\sum w} = \frac{68015.6}{1448} = 46.97\%$$

In the above example, we have assigned weights to the items in terms of their actual importance. But these actual quantities may not be known in all cases. In such a case a close estimate of weight is made and items are weighted accordingly. Here instead of actual, approximate weights are used. If the approximate weights are properly selected very little error is observed in the average. The choice of estimated weight creates a definite problem in certain cases. In constructing cost of living index it is necessary to determine the relative role played by the different items of consumption in average family budget. If we are to find out the average per acre yield of jute in Bangladesh and if we have only the per acre production figure of jute district-wise in Bangladesh then we are to assign weights in terms of the average of jute cultivation in each district. These are only but two examples showing the problem involved in estimating weight age. No guideline can be suggested as such, but in every case effort is made to ascertain the best possible relative importance of each item and then the estimate is used as weight.

In computing combined mean it is essential, if not always, obligatory, to give weight to the component means by their respective frequencies.

In computing combined mean it is essential, if not always, obligatory, to give weight to the component means by their respective frequencies. This is necessary when the items in the component series from which the means are derived are varying in number. If we are given three means of three different series each having the same number of items then we can simply add up the three means, divide the total by three and get the mean of combined series. This is because the relative importance of each of the mean is same. But if these three arithmetic means are derived from series having unequal number of items in them then we are to multiply each mean by the number of items of the corresponding series to arrive at the total value of that series.

Illustration 4.5

Mean income of seven groups of families and the number of families in each group are given below. Computation of the mean income of all the families in seven groups taken together.

Group No.	Annual Mean Income (in Tk.)	Number of Families
1	4570	25
2	3425	32
3	5260	21
4	7250	45
5	6500	42
6	9445	35
7	12250	53

Solution:

Calculation of mean for combined series.

Mean Income in Tk. x	Number of Families	nx
4570	25	114250
3425	32	109600
5260	21	110460
7250	45	326250
6500	42	273000
9445	35	330575
12250	53	649250
Total	253	1,913,385

Mean of the combined series,

$$= \frac{\sum n \bar{x}}{\sum n} = \frac{1,913,385}{253} = 7562.79$$

The mean income of all groups is Tk. 7,562.79

Mathematical Properties of the Mean

(i) The arithmetic mean possesses certain mathematical properties. If we calculate the deviation of each item from the mean the sum of negative deviations will be equal to the sum of positive deviations. In other words, the total of all the deviations will be zero. If we express in symbols $(x - \bar{x}) = 0$, For example, if the ages of 5 students are 14 yrs.,

17 yrs., 18 yrs., 20 yrs., 21 yrs., their mean age is 18 yrs. Deviation of each item from the mean is calculated below.

Age in years, x	Deviations, $(x - \bar{x})$
14	-4
17	-1
18	0
20	2
21	3
Total = 90	0

(ii) The second property is that if we replace each item of a series by the mean of the series and take the total of the replaced values we shall have the total of the values in the series. This total is the same as would have been obtained by adding the actual items.

Let,

$$\bar{x} = \frac{\sum x}{n} \text{ or, } n\bar{x} = \sum x$$

$$\therefore n\bar{x} = \sum x$$

(iii) The third property of arithmetic mean is that the total of the squares of deviations of items from the mean is minimum. In other words, the total of squared deviations of the items from any other value would be greater than the total of squared deviations of the items from the mean. Symbolically $(x - \bar{x})^2$ is less than $(x - A)^2$, where A is any value other than the mean. The following example will make the matter more Here mean $\bar{x} = 54$.

Height in inches, x	Deviation from mean, $x - \bar{x}$	Squared deviations $(x - \bar{x})^2$	Deviation from 50 $(x - 50)$	$(x - 50)^2$	Deviation from 55 $(x - 55)$	$(x - 55)^2$
42	-12	144	-8	64	-13	169
46	-8	64	-4	16	-9	81
51	-3	9	1	1	-4	16
56	2	4	6	36	1	1
62	8	64	12	144	7	49
67	13	169	17	289	12	144
324	0	454		550		460

Advantages and Limitations of Arithmetic Mean

Advantages:

- The arithmetic mean is easy to compute and understand. An elementary knowledge of arithmetic is sufficient for its computation. Because of this, it is the most commonly used average.
- The arithmetic mean takes into consideration all the items in the series and is, therefore, more representative of the series and can be determined with mathematical exactness.
- As the arithmetic mean can be determined in exact term it is suitable for further algebraic and arithmetic treatment.
- The arithmetic mean can be computed from the raw data and there turns to be no need of arranging the data in some order as is needed in some other types of statistical averages.

- e) The arithmetic mean can be computed even when the detail values are not available. If we know the total value of the items and the number of them we can compute the arithmetic mean. The total value of the items can be computed if we know the mean and the number of items.
- f) The arithmetic mean provides a good standard for comparison as the abnormal fluctuations in one direction tend to offset the abnormal fluctuations in the other direction provided the number of observations is reasonably large. In other words, the mean is the most stable type of average.
- g) In arithmetic mean due weight can be given to individual items in proportion to their relative importance.
- h) The arithmetic mean permits the computation of combined mean, which is not possible in case of median and mode.
- i) The arithmetic mean is least affected by sampling fluctuations. It is, therefore, the most stable type of average.

Disadvantages:

- a) As the mean makes use of all the values in the series, the extreme values definitely affect the average. This makes the mean less representative of the data having extremely large or small values. This happens mostly in income distribution where individual earnings fluctuate greatly. In such a case arithmetic mean is found to be less suitable.
- b) Arithmetic mean cannot be calculated when the two ends of the distribution are not known as happens in case of frequency distribution having open-ended class intervals. However, in such cases median and mode can be computed.
- c) The arithmetic mean cannot be located by the study of the position of items in the series but the median and mode can be.
- d) The mean obtained from a series may be a value, which may not at all occur in the series.
- e) As the arithmetic mean is not a positional measure it cannot be located graphically, but mode and median can be. For this reason the arithmetic mean is termed as computed average.

Self-Assessment Question

Short Question

- 1. What do you understand by measure of central tendency?
- 2. Define mean with an example.
- 3. Define weighted arithmetic mean.
- 4. Write an important difference between weighted arithmetic mean and arithmetic mean.
- 5. Briefly explain the mathematical properties of mean.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which of the following statistics is not a measure of central tendency?
(a) Arithmetic mean (b) Mode
(c) Median (d) Q_3
- (ii) The scores of the top ten students in a mid-term examination are listed below:
71, 67, 67, 72, 76, 72, 73, 68, 72, 72
Find the mean.
(a) 68 (b) 72 (c) 71 (d) 67
- (iii) Deviation from each observation from the mean is
(a) Greater than zero (b) Equal to zero
(c) Less than zero (d) Equal to mean
- (iv) Which measure of central tendency is not resistant to extremely small or extremely large data value in a numeric data set?
(a) Mode (b) Parameters
(c) Mean (d) Median
- (v) The total of the square of deviations of observation from the mean is
(a) Maximum (b) Minimum
(c) Zero (d) Negative

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) One of the most frequency used measures of spread in a set of data is called the mean.
- (ii) The mean for a population will generally be larger than the mean from a random sample from that population.
- (iii) A distribution is said to be symmetric when the sample mean and the population mean are equal.
- (iv) The geometric mean is a measure of variation or dispersion in a set of data.
- (v) The geometric mean is useful in measuring the rate of change of a variable over time.
- (vi) Data are considered to be right skewed when the mean lies to the right of the median.

Answer:

Multiple-Choice Question:

1. (i) d (ii) c (iii) c (iv) c (v) b

True/False

2. (i) F (ii) F (iii) F (iv) F (v) T (vi) T

Lesson 2: The Median

Lesson Objectives

After studying this lesson, you will be able to:

- Define median;
- Calculate median;
- Understand graphical method of locating median;
- Understand mathematical properties of median;
- Know the advantages, disadvantages and use of median.

Introduction

The median is the second measure of central tendency. It is used to describe the centre or middle of a series of data. The median is defined as the value of the central or middle item, which divides the series into two equal parts when the items of the series are arranged in ascending or descending order of magnitude.

Median

The median divides the series into two equal parts, one part consisting of all the values less than the median and the other comprising of all the values greater than the median. The median is a positional measure of central tendency. For example, let us take the following set of raw data:

The median is a positional measure of central tendency

3, 5, 8, 8, 9, 6, 12

If we arrange them in ascending order the series turns to be

3, 5, 6, 8, 8, 9, 12

Here we find that the value of the middle item is 8, which is the median of this series. In this case we haven't taken odd number of items, and therefore, the middle item can be located in such a way that the series is divided into two equal parts. But had there been even number of items there would not have been a middle item as can be observed from the series given below:

4, 7, 9, 12, 13, 15, 16, 19.

In the above series it can be observed that no one of the items can be taken as the middle item and as such we cannot locate the median by the given definition. For situation like this we are to further qualify our definition. In case of even number of items the median is taken as the mean of the values of the two middle items. Using this definition in the above example the median lies in between 12 and 13 and the mean of these two items, i.e., 12.5 is the median of the series. Thus the median is halfway between the two central values in an even series. Symbolically we can express the median of series having odd number of items by the value of $(n+1)/2$ th items where n stands for the number of items in the series. In a series having even number of items the median is the mean value of $n/2$ th and $(n/2 + 1)$ th items.

Where n stands for number of items.

Illustration 4.6

Marks obtained by 18 students are recorded below:

53, 38, 33, 47, 58, 43, 40, 50, 55, 48, 50, 45, 55, 40, 48, 42, 52, 47.

Median is to be computed:

At first we arrange the data into ascending to the order as follows.

Marks arranged in ascending order																	
33	38	40	40	42	43	45	47	47	48	48	50	50	52	53	55	55	58

Here $n = 18$

$n/2 = 9$ and $n/2 + 1 = 10$

So the mean of 9 th and 10 th values is
the median i.e.,

Median = $(47+48)/2 = 47.5$

47.5

Determination of Median from Grouped Data

In a grouped data much of the detail is lost in the process of grouping and as such the median cannot be found out correctly without recourse to original data. But from grouped data the median can be estimated under certain assumptions. We may be able to locate the class in which the median lies by cumulating the class frequencies. Initially we are to calculate the value of $(n+1)/2$ because $(n+1)/2$ th value is the median. Then we cumulate the class frequencies to find out the class whose cumulative frequency first exceeds or equals the value $(n+1)/2$. This class contains the median and is called the median class. Our next step is to estimate the median within the median class. This is done by using the following formula.

$$\text{Median} = L + \frac{\frac{n+1}{2} - fc}{f_m} \times c$$

Where,

L = The lower limit of the median group

C = Class interval of median group

n = The total frequency

f_m = The frequency of the median group

f_c = The cumulative frequency of the group preceding the median group.

Illustration 4.7

Table showing classification of families according to cultivated holding.
Calculation of median.

Cultivated holding (in acres*)	Number of families	Cumulative Frequency
Upto 1.00	257	257
1.00-2.00	138	395
2.00-3.00	187	582 (median group)
3.00-5.00	243	825
5.00-10.00	169	994
10.00 and above	26	1020

*Lower limit excluded.

Solution:

$$\frac{n+1}{2} = \frac{1020+1}{2} = 510.5$$

So the value of 510.5th item is the median, i.e., in this case the holding of 510.5th family is the median holding. Median holding lies in the group (2.00-3.00) as the cumulative frequency of this group just exceeds 510.5. By using the formula

$$\text{Median} = L + \frac{\frac{n+1}{2} - fc}{fc} \times c$$

$$2.00 + \frac{510.5 - 395}{187} \times 1$$

$$2.00 + \frac{115.5}{187} \times 1$$

$$= 2.00 + 0.62 = 2.62 \text{ acres}$$

Graphic Method of Locating Median

The median can be located with the help of graphs. Two methods of locating median graphically are available. The first one is with the help of cumulative frequency curve, which is known as ogive, and the second one is what is known as Galton's Method.

Locating Median from Ogive

The cumulative class frequencies are plotted in a graph where the horizontal axis represents class intervals and the vertical axis represents cumulative frequencies. The median position $n/2$ is marked in the vertical axis and a straight line is drawn from the median point, parallel to X-axis to intersect the ogive. Then a perpendicular line is drawn from the point of intersection to the base line. The point at which this perpendicular line touches the X-axis indicates the median value. Fig. 4.1 illustrates the location of median from ogive.

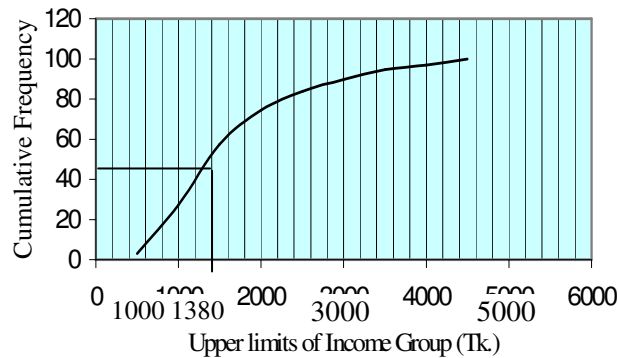
The median can also be located by drawing two ogives, one with ascending cumulative frequency and another with descending cumulative frequency in the same graph. The point of intersection of the two ogives will locate the median. From the point of intersection of the two ogives a perpendicular line is drawn to the base line and the value of the point at which the perpendicular line touches X-axis is the median.

Illustration 4.8 - Location of median graphically

Table showing the percentage distribution of households in a rural area of any one place according to Income.

Income Group per month (in Taka)	Percentage of households	Cumulative percentage frequency
Upto 499	3.0	3.0
500-999	24.2	27.2
1000-1499	30.2	57.4
1500-1999	17.0	74.4
2000-2499	9.4	83.8
2500-2999	6.0	89.8
3000-3499	4.9	94.7
3500-3999	2.2	96.9
4000-4499	3.1	100.0
	100	

Source: National Sample Survey, Third round, 1961.

Solution*Fig. 4.1 - Showing the location of median from ogive*

Median = Tk. 1380.

Mathematical Property of the Median

The median possesses only one mathematical property. Total of the deviations of the individual items of the series from the median is a minimum if we disregard the sign. If the deviations of the items are calculated from any value other than the median the total of such deviations, ignoring sign, will be more than the sum of the absolute deviations from the median. Symbolically the sum of deviations from median, ignoring sign, may be expressed as $\sum (x - \text{median})$. So $\sum (x - \text{median})$ is less than $\sum (x - K)$ where K is any value other than the median. The following example will make it clear.

Illustration 4.9

Weekly income in taka x	x - Median	x - 140	x - 150
125	20	15	25
140	5	0	10
165	20	25	15
155	10	15	5
145	0	5	5
175	30	35	25
135	10	5	15
144	1	4	6
148	3	8	2
Total	99	112	108

Median = Tk. 145.

Above example revealed that $\sum (x - \text{Median})$ always less than $\sum (x - k)$.

Advantages and Disadvantages of the Median

Advantages:

- a) The median is readily comprehensible and easy to calculate.
- b) The median always exists in a series and it can be found from any numerical data. It is capable of exact location and as such is often more representative than the mean.
- c) The median is not affected by extremely high or low values.
- d) The median is especially advantageous in open-ended distribution, skewed distribution and in distribution having varying class intervals.
- e) A distinct advantage of median is that it can be used in defining the median of attributes like efficiency, quality, merit, intelligence, virtue, etc. which are not amenable to quantitative description.
- f) In certain types of distributions median is more stable than the mean and is least affected by sampling fluctuations.
- g) Unlike the mean it can be located graphically.

Disadvantages:

- a) The median involves laborious process of arranging the items before it can be calculated particularly when the number of observations is large.
- b) When the observations are few the median is not likely to be fully representative of the group.
- c) In an irregular frequency distribution the median cannot be obtained exactly since the values of items vary widely in magnitude.
- d) It is not amenable to further mathematical treatment.
- e) In a frequency distribution the median falls within a class and it cannot be located with precision but can be estimated under certain assumption.
- f) The median does not permit the calculation of the median of medians as is possible in case of mean.
- g) In case of mean it is possible to find out the total value provided we know the mean and number of items. This is not possible in case of median.

Summary

The median is particularly used to describe the phenomena, which are not capable of direct quantitative description. So it finds its use mostly in social phenomena and in the description of qualitative characteristics. Even in case of observation capable of quantitative measurements median has got considerable use and in certain types of distributions like frequency distribution with open-ended class interval its use is of vital importance. In a skewed distribution as well as in distribution having unequal class interval the median has got considerable use. Depending upon the nature of the distribution the median has got much use in direct quantitative study too. The use of the median is, however, limited to some extent in business and economic field due to wide fluctuations in business and economic data.

Self-Assessment Questions:

Short Questions:

1. What is mean by median?
2. Write two properties of median
3. Write two disadvantage of median
4. Describe two mean advantages of median

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The most frequency used measure of central tendency is:
(a) Median (b) Mean
(c) Mode (d) All the above
- (ii) The median of a data set for a variable is the data value that:
(a) Appears the most often
(b) Is the average, that is, the sum of the all data values of the variable divided by the number of observations in the data set?
(c) None of there
(d) Lies in the middle of the data when the data is arranged is ascending order.
- (iii) Consider the following sample data:
25, 11, 6, 4, 2, 17, 9, 6
(a) 7.5 (b) 3.5 (c) 10 (d) None of the above
- (iv) In a right skewed distribution
(a) The median equals to the A. M
(b) The median is less than the A. M
(c) The median is larger than the A. M
(d) The median is zero
- (v) Which of the following statistics is not a measure of central tendency?
(a) Arithmetic (b) Median
(c) Mode (d) Q_3

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The median of a data set with 20 items would be the average of the 10th and 11th items in the ordered array.
- (ii) The median of the values 3.4; 4.7; 1.9; 7.6 and 6.5 is 1.9.
- (iii) In a sample of size '10' the sample mean is 15. In this care, the sum of all observations in the sample is $\sum X_i = 600$.
- (iv) If a set of data is perfectly symmetrical, the arithmetic mean must be identical to the median.
- (v) If the arithmetic mean of a numerical data set exceeds the median, the data are considered to be positive on right skewed.

Answer:

Multiple-Choice Question: 1. (i) b (ii) d (iii) a (iv) b (v) d

True/False: 2. (i) T (ii) F (iii) T (iv) T (v) F

Lesson 3: The Mode

Lesson Objectives:

After studying this lesson, you will be able to:

- Define mode;
- Locate mode from graphed data;
- Understand graphical methods of locating mode;
- Know the advantages and disadvantages of mode;
- Understand the use of mode.

Introduction

Mode is like mean and median, a central value of the distribution. The mode is the third measure of central tendency and can be simply stated as the value, which occurs most frequently. This is the most frequency occurring observation in a series. It is the position of highest concentration in the group. The mode thus can be defined by the size of the variable in which most of the cases cluster. The modal position is referred to as the position of highest density. So the mode is the maximum respected observation of the data.

Mode:

In a frequency distribution mode is the point of highest frequency. As the mode is defined as the item of frequent repetition it appears that it should be easy to locate. By mere inspection of a group of data and observing the way of their occurrence it is possible to find the mode. But if there is no repetition of items in groups of data then it is not possible to locate the mode from such a group. In order to locate the mode from ungrouped data we are to arrange the data in an array and observe the value, which occurs most often. For example, if we have the following observations:

5, 7, 6, 9, 10, 15, 9, 7, 9, 11.

We are to arrange them first in an array as follows:

5, 6, 7, 7, 9, 9, 9, 10, 11, 15.

Here we find that 9 occurs most often in the group and, therefore, 9 is the mode of this group. This has been easy to locate only because the observations had a repetitive character. But if the observations do not apparently show any tendency of repetition then the only way is to group the data which is likely to show some concentration. The mode, therefore, can be best located from the grouped data.

Location of Mode from Grouped Data

In a frequency distribution modal class is the class, which has maximum frequency. After finding out the modal class the problem is to find out the exact value of mode within the modal class. If we could assume that the frequencies are evenly distributed in the pre-modal and post-modal classes then we could take the mid-point of the modal group as the mode. Since in practical work the values are not always found to be evenly distributed in the post-modal and pre-modal classes, we resort to neutralize the affect of unevenness by adopting a special formula given below:

Mode can be defined by the size of the variable in which most of the cases cluster

$$\text{Mode } M_o = L_l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

Where,

L_l = The lower limit of the modal class

C = Class interval of modal class

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

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

Illustration 4.10

Compute mode from the following distribution:

Price Groups in Tk.	Frequency
15-30	7
30-35	21
35-40	46
40-45	62
45-50	35
50-55	16
55-60	5

Solution:

(40-45) is the modal group which has highest frequency. By using the formula, we can calculate mode as follows:

$$M_o = L_l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C = 40 + \frac{16}{16 + 27} \times 5$$

$$\therefore \Delta_1 = 62 - 46 = 16$$

$$\Delta_2 = 62 - 35 = 27$$

$$= 40 + (16/43) \times 5$$

$$= 41.86$$

$$\text{Mode} = \text{Tk. } 41.86.$$

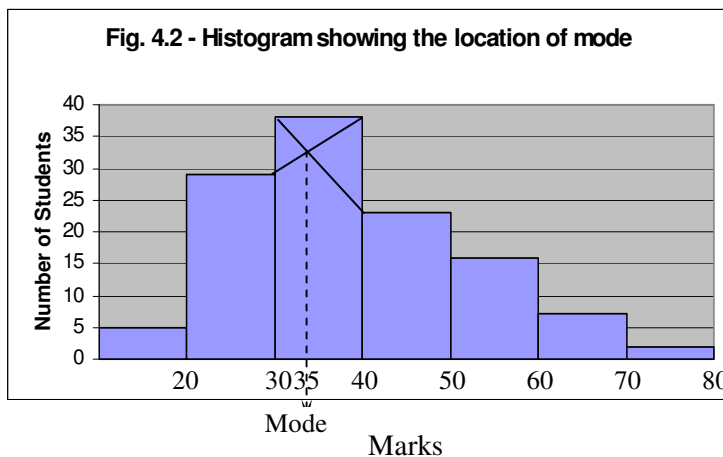
Graphic Method of Locating Mode

Mode can be located graphically from the histogram. In a histogram the highest frequency represents the modal class. Then the task is to find out the value of the mode within the modal class. The location of mode within the modal class is affected by the unevenness of the frequency distribution in the post-modal classes. In a histogram the two rectangles on both sides of the highest rectangle represent the pre-modal and post-modal classes. To neutralize the effect of uneven distribution of items in the pre-modal class and post-modal class we draw two lines from the two top ends of the two rectangles on both sides along the highest rectangle to the two top ends of the highest rectangle. Then from the point of intersection of these two lines perpendicular line is drawn to the X-axis. The value of the variable corresponding to the point at which the perpendicular touches the X-axis is the modal value. The process is illustrated in Fig. 4.2

Illustration 4.11

Table showing the distribution of students according to marks obtained in an examination

Percentage of Marks	Number of Students
10 to below 20	5
20 to below 30	29
30 to below 40	38
40 to below 50	23
50 to below 60	16
60 to below 70	7
70 below 80	2

**Advantages and Disadvantages of Mode****Advantages:**

- The mode is easy to understand and easy to locate. Unlike the arithmetic mean it can be located by inspection in certain cases.
- The mode is not affected by the extremely high or low values. Thus extreme variations are eliminated in the case of mode.
- It can be located when the values of the middle items are known and the detail extreme values are not known.
- In a frequency distribution having open-ended class intervals the mode can, unlike the mean, be ascertained with little difficulty, as the value of the mode is not affected by abnormal values of the extreme class intervals.
- The mode can be advantageously utilized in describing qualitative phenomena.
- The modal value is the actual value of the series and not an isolated example. The modal value possesses highest and dominant frequency and is likely to occur most often if further observations are made.
- Unlike the mean the mode can be located graphically.

Disadvantages:

- a) The mode is often ill defined.
- b) The mode is often uncertain and very difficult to locate exactly. In a frequency distribution, unless the number of observations is reasonably large and the observations show a clear tendency of clustering around certain value, the mode is difficult to locate.
- c) The calculation of mode requires the laborious process of arraying, grouping and sometimes regrouping of the data.
- d) It is not suitable for algebraic manipulation. It is not possible to find out the mode of modes. If we know the modal values of two or more series we cannot calculate the overall mode of the combined series.
- e) In an ungrouped data where no two observations are alike the mode cannot be located since no modal value exists in such a series.
- f) The value of mode is greatly affected by the method employed in computation.
- g) The mode is subjected to more sampling fluctuation than the mean and, therefore, less stable than the mean.
- h) The mode is not useful where it is desired to give weights to individual items. Mode does not give weights to individual items.

Use of Mode

The mode is one of the most commonly understood and used averages. Many people may not be familiar with the term 'mode', yet most of the people understand what it implies.

The mode has got much importance in describing data having qualitative characteristics and not subjected to direct quantitative analysis. In marketing his wares a manufacturer or trader want to know the consumer's preference for different kinds of products. The consumer's preference would be reflected by the modal preferences shown by different sections of the consumers. The use of mode in business and economics was, to some extent, limited in earlier times but the rise in the level of business and economic activities and the growing severity of competition has compelled the manufacturers and traders to get into the task of market and production analysis. With the help of production analysis attempt is made to bring down the cost of production at the competitive level and through market analysis endeavor is made to find out the right demand for the particular type of commodity as well as to undertake sales promotion. Production analysis is greatly augmented by the use of mode. Determination of model output per machine-hour and man-hour enables the management to operate at the level of maximum efficiency. Any cause of deviation from the modal output is rigorously investigated and attempt is made to remove the cause. In this way each element contributing to the production is made to operate with minimum of waste and maximum of efficiency resulting in low cost of operation. On the other hand, through market analysis attempt is made to know the demand of different types of commodity, which is reflected by the consumer's preferences. We have already observed that the elements

like consumer's preference can best be determined with the help of mode, since consumer's preference is nothing but the modal preference shown by different sections of consumers. This knowledge of consumer's preference on the one hand enables a dealer of the commodity to adjust his stock according to the modal preferences and on the other hand enables the manufacturer to adjust his production. Thus we find that the mode has got tremendous potentialities of being used in business and economic activities. In fact, the use of mode in these fields is increasing greatly day by day.

Self-Assessment Questions:

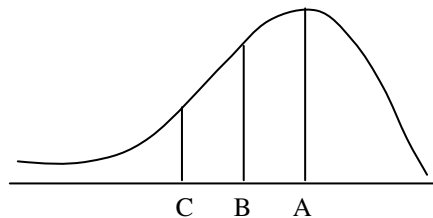
Short Question

1. Define mode with an example.
2. Write two advantage of the mode.
3. Write two disadvantage of the mode

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The scores of the top ten students in a mid-term examination are listed below:
Find the mode score
(a) 68 (b) 67 (c) 65 (d) 66
- (ii) The Maximum repeated observation in a set of data is known as
(a) Mean (b) Variance (c) Median (d) Mode
- (iii) Which of the following is the easiest the compute?
(a) A. M. (b) Mode (c) Median (d) G. M.
- (iv) Which measure of central tendency may have more show one value in a numeric data set?
(a) Mode (b) Median (c) Midrange (d) Mean
- (v)



For the distribution drawn above, identity the mean, median and mode.

- (a) A=Mode, B=Mean, c=Median
- (b) A=Mode, B=Median, C=Mean
- (c) A=Median, B = Mode, C=Mean
- (d) A=Mean, B=Mode, C=Median

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) A set of data in which the mean, median and mode are all equals is said to be a multi-modal distribution.
- (ii) In a symmetric and mound shaped distribution, we expect the values of the mean, median and mode to differ gristly from one another.
- (iii) If the population mean is equal to mode, you can state that the population is symmetric.
- (iv) Suppose a study of hours that have sold recently in your community showed the following frequency distribution for the number of bedrooms:

Bedrooms	Frequency
1	1
2	18
3	140
4	57
5	11

Based on this information the mode for the data is 140.

- (v) It is possible for set of data to have multiple modes as well as multiple medians, but there can be only one mean.

Answer:

Multiple-Choice Question:

1. (i) b (ii) d (iii) c (iv) a (v) b

True/False

2. (i) F (ii) F (iii) F (iv) F (v) F

Lesson 4: Geometric Mean, Harmonic Mean

Lesson Objectives:

After studying this lesson you will be able to:

- Define geometric mean, harmonic mean;
- Calculate geometric mean, harmonic mean;
- Advantages and disadvantages, geometric and harmonic mean;
- Understand use of geometric mean;
- Understand quadratic mean.

Introduction

The geometric mean is a compound average and takes into account all the items in the series. It is geometric rather than arithmetic in character. The geometric mean is computed not on additive basis but in terms of products and ratios.

Geometric mean G. M.

The geometric mean is defined as the n th root of the product of n quantities contained in the series. The values of the series are multiplied and root of the product is taken. If n is taken as the number of items in a series then the formula of geometric mean can be written as,

$$G = \sqrt[n]{x_1 \times x_2 \times x_3 \times x_4 \times \dots \times x_n}$$

$$G = (x_1 x_2 x_3 x_4 x_n)^{1/n}$$

Where, $G =$ The geometric mean
 $x =$ The value of individual items
 $n =$ The number of items.

In practical computation the method of logarithms is used. If we take logarithms on both sides of the above formula we get,

$$\begin{aligned} \text{Log } G &= \log (x_1 x_2 x_3 \dots x_n)^{1/n} \\ &= \frac{1}{n} \log (x_1 x_2 x_3 \dots x_n) \\ &= \frac{1}{n} (\log x_1 + \log x_2 + \log x_3 + \dots + \log x_n) \end{aligned}$$

$$\text{Log } G = \frac{\sum_{i=1}^n \log x_i}{n}$$

Using summation sign,

Illustration 4.12

Percentage of marketable surplus of commodities for 10 families is given below.

2.5, 6.8, 4.9, 64.4, 31.5, 13.0, 43.7, 28.0, 13.9, 5.4.

Calculation of Geometric Mean from Ungrouped Data.

Solution

x	logx
2.5	0.39794
6.8	0.83251
4.9	0.69020
64.4	1.80889
31.5	1.49831
13.0	1.11394
43.7	1.64048
28.0	1.44716
13.9	1.14301
5.4	0.73239
Total	11.30483

$$\begin{aligned}\text{Log } G &= (\log x)/n = 11.30483/10 \\ &= 1.13048\end{aligned}$$

$$G = \text{Antilog of } 1.13048 = 13.51$$

$$\text{Geometric mean} = 13.51$$

Advantages and Disadvantages of Geometric Mean

Advantages:

- The geometric mean takes into account all the items in the series and can be calculated with mathematical exactness if there is no negative or zero value in the series.
- The geometric mean takes into consideration the extreme values and is, to some extent, affected by them.
- In measuring rate of change the geometric mean is the most suitable average. No other average can serve this purpose so accurately.
- The geometric mean is capable of mathematical treatment.
- If the geometric mean of two or more series is known it is possible to find the average of the combined series. Like the arithmetic mean the geometric mean possesses this distinct advantage over mode and median.

Disadvantages:

- The geometric mean is very difficult to calculate and is not commonly understood. The difficult method of computation involving the use of logarithms makes it unpopular.
- The geometric mean cannot be computed where there is any negative or zero value in the series.
- The geometric mean may not be an actual representative of the series and may be found to locate at a point where few or none of the observations may lie.

Use of the Geometric Mean

The geometric mean is mainly useful in taking averages of ratios, percentages and rate. The geometric mean finds its great use in economic field particularly in the construction of index number. Index numbers reveal percentage changes rather than actual changes and as such

averaging them with the help of arithmetic mean would give biased results. Relative changes as in case of index numbers must be measured by using geometric mean. In determining the rate of increase or decrease in the phenomena like population change, amount of compound interest, etc., geometric mean gives us more accurate picture than the arithmetic mean. Distributions having geometric progression should preferably be averaged by geometric mean. The geometric mean due to its intricate method of computation is not used very often unless there is the special need for this type of average.

Harmonic Mean

The harmonic mean of a series of values is given by the reciprocal of the mean of the reciprocals of the individual values. If there are n items in the series the formula of the harmonic mean is given by:

$$H = \frac{n}{\left(\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n} \right)}$$

Where, H = The harmonic mean

$$H = \frac{n}{\sum \frac{1}{x}}$$

x = The value of individual items

n = The total number of items.

In practical computation of harmonic mean of a given series of data we first calculate the reciprocal of each item, sum up the reciprocal values and divided the total by the number of items. Then we compute the reciprocal of the quotient to get the harmonic mean.

Illustration 4.13

A train runs first one mile at the speed of 30 miles per hour, second mile at the speed of 25 miles per hour, third mile at the speed of 40 miles per hour and the fourth mile at the speed of 60 miles per hour. Calculate of the average speed of the train throughout the 4 miles distance.

Solution

In this case harmonic mean of the speeds will be the actual average speed of the train for the whole distance.

$$\begin{aligned} \text{Average Speed} = \text{Harmonic Mean} &= \frac{4}{\left(\frac{1}{30} + \frac{1}{25} + \frac{1}{40} + \frac{1}{60} \right)} \\ &= \frac{4}{(0.0333 + 0.0400 + 0.0250 + 0.0167)} \\ &= \frac{4}{0.115} = 34.78 \text{ miles per hour} \end{aligned}$$

Average speed = 34.78 miles per hour.

Advantages and Disadvantages of Harmonic Mean

Advantages:

- a) It takes into account all the items of the series and can be calculated with mathematical exactness.
- b) It is relatively less affected by extremely large values and gives more weight to small values.
- c) The harmonic mean is amenable to further algebraic treatment.

Disadvantages:

- a) The harmonic mean is not readily comprehensible and is also difficult to compute.
- b) It cannot be determined without detail knowledge of all the items in a series.
- c) It may not be an actual value occurring in the series.

Use of Harmonic Mean

The harmonic mean has got very limited use in general statistical work. The roundabout process of its computation also makes it unpopular to the common people. Occasionally, however, when observations are made in terms of work done per hour, miles covered per hour, quantity of things purchased per taka etc., harmonic mean is found useful.

Quadratic Mean

The quadratic mean is the square root the arithmetic mean of the squares of the values of a series. The formula for calculating quadratic mean is:

$$QM = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

Where, x = The value of the individual items
 n = The number of items.

Use of Quadratic Mean

Quadratic mean is the least important of all the averages and is found very rare in use. It is used in the computation of standard deviation as well as in computation of mean of standard deviations.

Self-Assessment Questions:

Short Questions:

- 1. Define Geometric mean with an example
- 2. Write down about Harmonic mean.
- 3. Write the relation between Geometric and Harmonic mean.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which is computed in terms of product and ratio's
(a) A. M (b) Median (c) Mode (d) G. M.

- (ii) The rate of return on Internet Service Provider over a 10 years period are:

10.25%	12.64%	8.37%	9.29%	-2.35%
6.23%	42.53%	29.23%	15.25%	21.52%

The geometric mean rate of return is:

- (a) 15.30% (b) 17.30% (c) 15.80% (d) 16.30

- (iii) Relationship between G. M. and H. M is

- (a) $G. M \leq H. M$ (b) $G. M. < H. M.$
(c) $G. M. \geq H. M.$ (d) $G. M. = 0$

- (iv) The 12-month rate of return over a nine-year period of a particular share is as follows:

0.099	-0.289	0.089	0.226	0.041	0.161	0.064	-0.029	0.022

The geometric mean rate of return for this share is:

- (a) 4.23% (b) 3.23% (c) 5.11% (d) 4.10%

- (v) The rate of return for a stock over a seven-year period is given below:

0.527	0.145	0.684	1.146	0.564	0.883	0.436

The geometric mean rate of return is:

- (a) 0.5880 (b) 0.6880 (c) 0.5990 (d) 0.4990

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The geometric mean is a measure of variation on dispersion in a set of data.
(ii) The geometric mean is useful in measuring the rate of change of a variable over time.
(iii) The rate of return for the Internet Service Provider over a seven year period is:

-0.224	-0.029	-0.061	-0.493	-0.286	-0.160	-0.186

The geometric mean rate of return is – 22.03%

- (iv) The geometric mean does not takes into consideration the extreme values in the series.
(v) Cities A, B and C are equidistant from each other. A motorist travels from A to B at 30 Km/h; from B to C at 40 Km/h and from C to A at 50 Km/h. The average speed for the entire trip is 38.3 Km/h.

Answer:

Multiple-Choice Question:

1. (i) d (ii) a (iii) b (iv) b (v) c

True/False

2. (i) F (ii) T (iii) T (iv) F (v) T

Lesson 5: Relationship between Different Measures of Central Tendency

Lesson Objectives:

After studying this lesson you will be able to:

- Understand relationship among A. M., G. M & H. M;
- Able to locate mean, median and mode on a symmetrical curve;
- Able to understand properties of a good measure of central tendency;
- Describe comparative study of the measure of central tendency.

Relationship Among Three Mean (A. M., G. M and H. M):

Some mathematical relationship between the different measures of central tendency exists. It is found that the arithmetic mean is greater than the geometric mean and the geometric mean is greater than the harmonic mean of the same set of observations. This happens when the values in the series are unequal. If all the items of a series are of equal value then the arithmetic mean, the geometric mean and the harmonic mean of the series will be the same. Symbolically

$$A.M. \geq G.M. \geq H.M.$$

Proof - For simplicity let us take a series with two values expressed by x_1 and x_2 . Then,

$$\begin{aligned} (x_1 - x_2)^2 &\geq 0 \text{ because a square term cannot be negative.} \\ \text{or } (x_1 + x_2)^2 - 4x_1x_2 &\geq 0 \\ \text{or } (x_1 + x_2)^2 &\geq 4x_1x_2 \\ \text{or } (x_1 + x_2) &\geq 2\sqrt{x_1x_2} \quad [\text{taking square root}] \\ \text{or } (x_1 + x_2)/2 &\geq \sqrt{x_1x_2} \\ \text{i.e., } A.M. &\geq G.M. \end{aligned}$$

Again,

$$\begin{aligned} \frac{x_1 + x_2}{2} &\geq \sqrt{x_1x_2} \\ \text{or, } \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} &\leq \sqrt{x_1x_2} \quad [\text{Taking reciprocal on both side}] \\ \text{or, } \frac{2x_1x_2}{x_1 + x_2} &\leq \sqrt{x_1x_2} \quad [\text{Multiply both side by } x_1x_2] \\ \text{or, } \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} &\leq \frac{x_1x_2\sqrt{x_1x_2}}{\sqrt{x_1x_2}\sqrt{x_1x_2}} \\ \text{or, } \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} &\leq \sqrt{x_1x_2} \end{aligned}$$

i. e., $G. M. \geq H. M.$

Again multiplying both sides by,

So, $A.M. \geq G.M. \geq H.M.$

In a symmetrical distribution the values of the mean, median and mode coincide. In other words, if we have a symmetrical frequency curve we find that the centre of the curve has the highest density and the point of balance. The mean, the median and the mode are found exactly at the centre of the distribution and their values are identical. This can be observed from the following symmetrical curve.

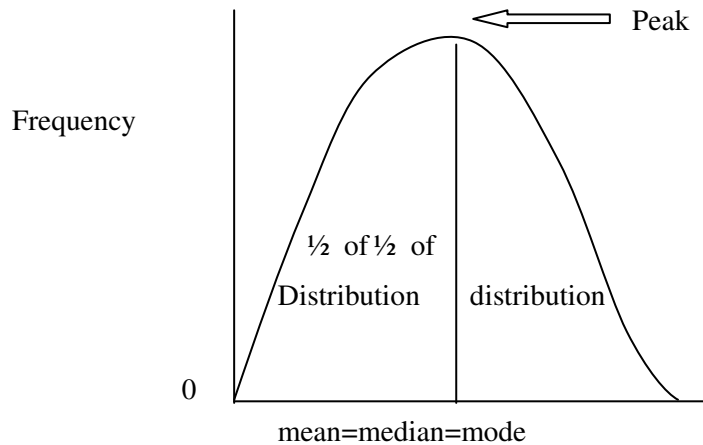


Fig. 4.3 -Location of the Mean, the Median and the Mode on a Symmetrical Curve

From the above curve it can be seen that the peak of the curve-representing mode corresponds to the mean value at the base and the ordinate from the peak of the curve divides the area of the curve into two equal parts. So the ordinate represents the median value at the base. So it can be clearly observed that the values of the mean, the median and the mode are identical.

In a moderately asymmetrical distribution the values of the mean, the median and the mode vary. Karl Pearson has given an empirical formula showing the relationship between the mean, the median and the mode for such type of distribution as given below:

Mean - Mode = 3 (Mean – Median)

In a distribution having positive skewness i.e., when the curve is elongated to the right, the value of mode will be the lowest, the value of the median will be the next highest to the mode and the mean will be the highest value i.e., $\text{mean} > \text{median} > \text{mode}$. On the other hand, in a distribution which skewed to the left, i.e., have negative skewness the value of the mode will be the highest and that of median next highest and mean will be the lowest value i.e., $\text{mean} < \text{median} < \text{mode}$.

Desirable Properties of a Good Measure of Central Tendency

A good measure of central tendency should possess some properties. These are:

- 1) It should be rigidly and unambiguously defined. Unless it is well defined there is the chance of being misunderstood and also there may be the chance of personal bias on the part of the person

computing the same. So an average should be well defined and not to be left to the mere guess or estimation of the observer.

- 2) It should take into account all the items in the series from which it is obtained. An average is taken as the type or the representative of the whole series, which it represents. Unless it is based upon all the observations there is the chance that it may not be wholly representative of the group.
- 3) An average should readily be comprehensible. It should be simple and obvious. It should not be too abstract in concept rendering it difficult to grasp.
- 4) An average should be capable of being computed with reasonable degree of ease and rapidity. An average, even if otherwise good, may render itself unacceptable due to complicated method of its computation. An easily calculated average should be preferred but that should not be done at the cost of other properties.
- 5) An average should be such that it is not affected very much by sampling fluctuations. Some sort of variation in the averages of the different samples drawn from the same population is quite likely. But attempt should be made to adopt an average, which has least sampling fluctuation.
- 6) The average should be readily amenable to algebraic treatment.

A Comparative Study of the Measures of Central Tendency

If we look to the arithmetic mean from the point of view of the above six desirable properties it would be obviously clear that the mean reflects most of them. The arithmetic mean is clearly defined, can be calculated with relative ease and rapidity, takes into account all the observations in the series, is capable of further mathematical treatment and is relatively more stable than other types of averages particularly more than the median in most of the cases. Above all, the arithmetic mean is very easy to understand. Thus from the point of view of all the properties, the arithmetic mean seems to be a satisfactory measure of central tendency and its use is recommended for all general purposes unless there is special reason to choose other types of averages. There are circumstances, which calls for the use of a measure of central tendency other than the mean. Let us, therefore, concentrate on other types of averages.

The median is more easily computed from a given distribution and even in certain cases it is found to be less affected by sampling fluctuation than the mean. But in the case of continuous distribution median may turn to be indeterminate and even in case of continuous frequency distribution the median cannot be located with precision and is only estimated. Again, the median is not suitable for further mathematical treatment and involves laborious process of arraying the data before computation. Of course, the median is readily understood and has an extra advantage of being suitable for graphic location. The median is not found in general use but has its use in certain special circumstances. The data involving attributes cannot be averaged by the mean but the median

Arithmetic mean is very easy to understand. Thus from the point of view of all the properties, the arithmetic mean seems to be a satisfactory measure of central tendency.

The data involving attributes cannot be averaged by the mean but the median can give a measure of average for attributes.

Distribution with varying size of class interval the median is more easily calculable than the mean.

Like the median, mode can be located graphically, which is not possible in case of other types of averages.

can give a measure of average for attributes. Where it is desirable to do away with the extreme values the median is found to be more suitable. In an open-ended distribution it is impossible to find out the mean. But the median is not affected by such open-ended character of the frequency distribution. Again in a distribution with varying size of class interval the median is more easily calculable than the mean.

The mode is found to be less suitable in elementary work due to the difficulty in its precise location. It is not suitable for further algebraic treatment. In frequency distribution the mode can only be estimated. It is less stable than the mean. But the mode is readily understood and in many cases easy to locate. Like the median it can be located graphically, which is not possible in case of other types of averages. Again like the median it can be used to neutralize the effect of extreme values and can be readily computed from the distribution having open-ended class intervals. So these types of distributions call for the use of the mode to show the average value. The mode is most suitable for qualitative study to which the other averages except the medians are completely unsuitable. In determining the consumer's preference on the average size of hats, shoes and the like nature of articles used, the average employed is the mode and not any other type.

The other three measures of central tendency viz., the geometric mean, the harmonic mean and the quadratic mean, although possess certain desirable properties, yet they are the type devised only to suit specific type of investigation. Of course, the last one i.e., the quadratic mean is found very rare in use. Although the harmonic mean is rigidly defined, can be determined accurately, takes all the items into consideration and is subject to mathematical treatment; yet it is not found in common use due to abstract nature of its concept and the difficult process of its computation. It is used in those cases where it is desirable to reduce the influence of the extremely large values and to give more weight to the relatively small values. Like the harmonic mean, the geometric mean possesses many of the properties of a good average but the difficult process of its computation and the difficulty in understanding stand in the way of its general acceptability. However, in certain types of data, involving ratios, rates and percentages the geometric mean is the most suitable average. In the computation of index numbers the geometric mean has got distinct advantages over other measures of central tendency.

Summary:

The above analysis shows that each type of average has got respective advantages and disadvantages and their applicability be different circumstance to be employed. Neither all averages can be applied to different types of data nor they can be employed for a particular purpose. Different types of averages serve the different purposes and the user is to pick up the best one suitable for his purpose. A detailed knowledge of the peculiar characteristics and applicability possessed by the different measures of central tendency is, therefore, a vital need as a guide to selection. A good deal of caution and foresight is needed in selection of the right type of average.

Different types of averages serve the different purposes and the user is to pick up the best one suitable for his purpose

Self-Assessment Questions:

Short Questions:

1. What do you mean by Geometric mean?
2. Prove the relation between Geometric and Harmonic mean i.e. $G.M. \geq H.M.$
3. Is it possible for Geometric mean use when at least one observation is Zero, If not, why?

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which of the following is not a measure of central tendency?
(a) The arithmetic mean (b) The harmonic mean
(c) The geometric mean (d) The interquartile range
- (ii) Which of the following is sensitive to extreme values?
(a) The median (b) The interquartile range
(c) The arithmetic mean (d) The first quartile
- (iii) Who give the empirical formula for showing the relationship among media, median & mode?
(a) Nazrul (b) Cochran (c) K. Pearson (d) R. A. Fisher
- (iv) The distribution of salaries of professional cricket players is skewed to the right. Which measure of central tendency would be the best measure to determine the location of the centre of the distribution?
(a) Mode (b) Mean (c) Frequency (d) Median
- (v) Which measure of the central tendency is more representation of the typical observation if the graph of the data is skewed to the right?
(a) Median (b) Midrange (c) Mean (d) Mode.
- (vi) A distribution that has the right tail longer than the left tail is considered
(a) Skewed right (b) Skewed left
(c) Skewed centrally (d) Not skewed

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) In a distribution, which is skewed to the left, $\text{mean} < \text{median} < \text{mode}$.
- (ii) An Economics Professor bases his final grade on homework, tow midterm examinations, and a final examination. The homework counts 10% toward the final grade, while each midterm examination counts 25%. The remaining portion consists of the final examination. If a student scored 95% in homework, 70% on the first midterm examination, 96% on the second midterm examination, and 72% on the final, his final average is 79.8%.
- (iii) In a distribution with varying size of class interval the median is more early calculated than the mean.
- (iv) In the computation of index numbers geometric mean is more useful than other measures of central tendency
- (v) In a skewed distribution, we expect the values of the mean, median and mode to be approximately equal, since they are all measures of center.

Answer:

Multiple-Choice Question: 1. (i) d (ii) c (iii) c (iv) d (v) a (vi) a

True/False: 2. (i)- F (ii)- T (iii)- T (iv)- T (v)- F

Exercise

1. (a) What do you mean by measures of central tendency? Write down the properties of a good measure of central tendency.
(b) Describe the different measures of central tendency of a frequency distribution, mentioning their merits & demerits.
2. (a) Define measure of central tendency and measure of location. Why they are so called? Write down different between measure of central tendency and measure of location.
(b) What are the desirable properties for an average to possess? Mention the circumstances to use median and mode as a suitable measure of central tendency.
3. Define arithmetic mean, geometric mean and harmonic mean of both ungrouped and grouped data. Compare and contrast the merits and demerits of them. Show that $A.M. \geq G.M. \geq H.M.$
4. (a) What is the difference between measures of central tendency and measures of location?
(b) Find the weighted arithmetic mean of first n natural numbers where weight is the corresponding value of each observation.
5. (a) What are the chief measures of central tendency? Give a comparative study of these. Show that mean deviation from mean is zero.
(b) The means of three sets of observation are 12.8, 15.6 and 14.3 where number of observations in the sets are 50.62, 48, respectively. Find the mean of the combined set of observations.
6. (a) Define weighted mean. Explain clearly the relationship between mean, median and mode in a moderately asymmetrical distribution.
(b) Write down the demerits of mode, geometric mean and harmonic mean.
(c) The weighted geometric mean of 10, 15 and 18 is 16 where the weights of first and second observations are 3 and 4 respectively. Find the weight of the third observation.
7. (a) Write down the merits and demerits of different measures of central tendency.
(b) The arithmetic mean and geometric mean of two observations are 20 and 16, respectively. Find the observations.
8. (a) What are the desirable properties of an average? Discuss the merits and demerits of different measures of central tendency.
(b) The median of the following frequency distribution is 16.56

Class interval	10-12	12-14	14-16	16-18	18-20	20-22	Total
Frequency	17	f_1	32	50	f_2	f_2	152

- (i) Find f_1 and f_2 .
- (ii) Find percentage of observations the value of which is 14 and above.
- (iii) Find percentage of observations the value of which is less than 18.
- (iv) Find maximum value of the first 20% lower values in the data set.

9. (a) Define weighted geometric mean and weighted harmonic mean. Explain the situations where geometric mean and harmonic mean are used profitably as a measure of central tendency.
- (b) A person covers 2 miles on foot at a speed of 3 miles/hour. 10 miles by bus at a speed of 20 miles/hour and 2 miles by boat at a speed of 2 miles/hour. Find the average speed of journey of the person.
10. (a) Define geometric and harmonic means. Write down the uses, merits and demerits of these two means.
- (b) Find arithmetic mean of the series 1, 3, 5, 7 upto n-th term.
- 11 (a) What are measures of location? Why these are so called? Explain the merits and demerits of measures of location.
- (b) Calculate mean, median and mode of the following frequency distribution.

Class interval	100-125	125-150	150-175	175-200	200-225	225+
Frequency	10	37	55	48	35	15

12. (a) What is measures of central tendency?
- (b) Which measure of central tendency will be suitable to compare.
- (i) the grade point average of two groups of student.
 - (ii) productions of two jute industries.
 - (iii) salaries of two groups of workers
 - (iv) rate of change of production in two industries.
 - (v) compare the speed of two industries
 - (vi) per capital income of several countries.
 - (vii) temperature of two seasons.
- (c) The following data represent the distribution of daily wages of some workers:

Class interval of wages	<80	80-90	90-100	100-110	110-120	120-130	130+
No. of workers	5	45	48	52	35	20	45

- (i) Find maximum wage of first 30 percent low paid workers.
 - (ii) Find minimum wage of last 30 percent high paid workers.
 - (iii) Draw Box-and-whisker plot of the distribution
 - (iv) Find the average wage of major group of workers
 - (v) Find P_{80} from graph.
13. (a) What do you mean by central tendency? Describe geometric mean and median
- (b) For n non-zero observations prove that $A.M \geq G.M. \geq H. M.$
- (c) The arithmetic mean and standard deviation of 50 observations are 40 and 10 respectively. A new observation 30 is included with these 50 observations. Find arithmetic mean and standard deviation of new set of observations.

MEASURES OF VARIATION



Before this unit, it has been discussed that, measures of central tendency usually tends to lie in the centre of the arrange. But in practical it is not true. There present some variation or dispersion. Measures of variation help to find how individual observations are dispersed around the mean of a large series.

The term variation means the Scatterdness of observation from some central value as well as mean, median and mode etc. In this unit we have discuss various measures of variation and their uses in business field experiment.

Lesson 1: Variation or Spread or Dispersion

Lesson Objectives:

After completing this lesson we would be able to:

- Understand the meaning of dispersion;
- Understand the need to measure dispersion;
- Explain the uses of dispersion measures;
- Understand the properties of a good measure of dispersion.

Introduction

We know that, measures of central tendency represent the average value of a data set. However, this measures does not indicate any information regarding the natures of scatterness of the observation. One of the important characteristics is to estimate or measuring the scatterness of the observation. Dispersion is the techniques to measures of the scatterdness, known as measures of dispersion.

Variation

Variation or spread or dispersion means scatteredness of individual observations from the central value (s). Significance of measuring variability lies in the fact that two sets of data may have the same central location but one may be more spread out than the other. Further measures of dispersion gives us additional information that enables us to judge the reliability of the measures of central tendency. For example, widely dispersed earnings indicate a higher risk to stockholders and creditors than do earnings remaining relatively stable.

Let us look at the three distributions in Figure (5.1).

The mean of all three curves is the same, but curve A has less spread (or variability) than curve B, and curve B has less variability than curve C. If we measure only the mean of these three distributions, we will miss an important difference among the three curves. Likewise, for any data, the mean, the median, and the mode tell us only part of what we need to know about the characteristics of the data. To increase our understanding of the pattern of the data, we must also measure its dispersion –its spread, or variability.

Why is the dispersion of the distribution such an important characteristic to understand and measure?

First, it gives us additional information that enables us to judge the reliability of our measure of the central tendency. If data are widely dispersed, such as those in curve C in Figure 5.1, the central location is less representative of the data as a whole than it would be for data more closely centered around the mean, as in curve A.

Second, because there are problems peculiar to widely dispersed data, we must be able to recognize that data are widely dispersed before we can tackle those problems.

Measures of dispersion gives us additional information that enables us to judge the reliability of the measures of central tendency

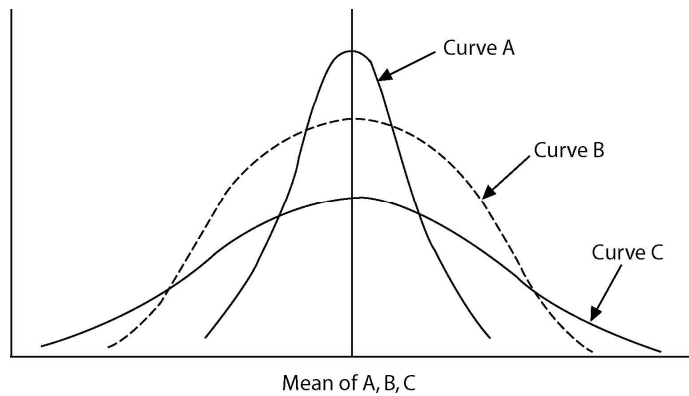


Figure 5.1.1

Third, we may wish to compare dispersions of various samples. If a widespread of values away from the center is undesirable or presents an unacceptable risk, we need to be able to recognize and avoid choosing the distributions with the greatest dispersion.

Properties of a good measure of variation

1. The measure should be easy to understand and easy to calculate.
2. The measure should be rigidly defined. It should have one and only one interpretation so that the personal prejudice or bias of the investigator does not affect the value or its usefulness.
3. The measure should reflect all the values in the data set. If it is calculated from a sample, then the sample should be random enough to be accurately representing the population. This means that if we pick 10 different groups of college students at random and we compute the variation of each group, then we should expect to get approximately the same value from these groups.
4. It should not be affected much by extreme values. If a few very small or very large items in the data, unduly influence the value of the variation measure by shifting it to one side or the other then the measure of dispersion would not be really typical of the entire series.
5. The measure should be suitable for further algebraic treatments.

The measure should reflect all the values in the data set. If it is calculated from a sample, then the sample should be random enough to be accurately representing the population.

Self-Assessment Questions:

Short Questions:

1. Discuss the need for measuring dispersion.
2. Mention the properties of a good measure of dispersion.
3. Choose which of the three curves shown in Figure (3.1.1) best describes the distribution of the following characteristics of various groups. Make your choices only on the basis of the variability of the distributions. Briefly state a reason for each choice.
 - a) The number of points scored by each player in a professional basketball league during an 80-game season.
 - b) The salary of each of 100 people working at roughly equivalent jobs in the government service.
 - c) The grade-point average of each of the 15,000 students at a public university.
 - d) The salary of each of 100 people working at roughly equivalent jobs in private companies.
 - e) The grade-point average of each student at a public university who has been accepted for Ph.D. program.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which descriptive summary measures are considered to be resistant statistics?
 - (a) The arithmetic mean and standard deviation
 - (b) The interquartile range and range
 - (c) The mode and variance
 - (d) The median and interquartile range
- (ii) Which of the following is the most frequently used measure of variation?
 - (a) The Range
 - (b) The Standard Deviation
 - (c) The Median
 - (d) The Mode
- (iii) When extreme values are present in a set of data, which of the following descriptive summary measures are most appropriate.
 - (a) CV and range
 - (b) AM and SD
 - (c) Interquartile range and median
 - (d) Variance and interquartile range
- (iv) Which of the following numerical summary measures cannot be negative?
 - (a) Standard Deviation
 - (b) Q_3
 - (c) Mean
 - (d) Mode
- (v) Consider the following data, which represent the number of miles employees travel from home to work. There are two samples: one for male and one for females.

Male:

13	5	2	23	14
5	1	3	6	7
14	11	7	8	4
13	2	5	8	9

Female:

15	6	3	2	4
6	3	1	7	19
5	3	7	12	4
6	2	18	4	6

Which of the following statements is true?

- (a) The female distribution is more variable since the range for the females is greater than for the males.
- (b) Female is the sample travel further on average than do males.
- (c) The distribution of travel males is symmetrical for both males and females.
- (d) The standard deviation for the males exceeds that of the females in these samples.

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The dispersion of a data set gives insight into the reliability of the measure of central tendency
- (ii) The interquartile range is a measure of variation or dispersion in a set of data.
- (iii) A dairy firm bottles milk in one gallon containers. At a recent mailing, the production manager asked top management for a new filling machine that he argued would assure that all containers had exactly one gallon of milk. Based on sound statistical principles, the top management group should conclude that the production manager could have merit to his argument.
- (iv) One of the most frequently used measures of the spread in a set of data is called the mean.
- (v) The range is a n ideal measure of variation since it is not sensitive to extreme values in the data.
- (vi) One of the advantages of dispersion measures is that any statistic that measures absolute variation also measures relative variation.

Answer:

Multiple-Choice Question:

1. (i)- d (ii)- b (iii)-c (iv)- a (v)-d

True/False

2. (i)- T (ii)- T (iii)- F (iv)- F (v)- F (vi)- F

Lesson 2: Measures of Dispersion (Distance Measures)

Lesson Objectives:

After studying this lesson, you will be able to:

- Define the distance measures of dispersion;
- Describe distance measures range and fractiles;
- Understand graphical presentation of interquartile range.

Introduction

The words dispersion indicate the scatterdness of observations from some central value i.e, the observations may be clustered near to the central value or may be scattered from central value. So, we need such a measure which will helps us to have the idea about the average deviation of observations from some central value and the variability of observations is measured by the measure of dispersion. The measures of dispersion can be divided into two groups

(i) Absolute measures of dispersion.

(ii) Relative measures of dispersion.

Absolute measures of dispersion

It is a measure, which provides the information on scatteredness of average deviation. There are two measures category:

- a) When dispersion is measured in terms of the difference between two values selected from the data set is called the **distance measures**. The range, the interfractile range, the interquartile range are of this category.
- b) The measures of dispersion that deal with the **average deviation** from some measure of central tendency are called the average measures. Variance, Standard deviation and Mean deviation are of this category.

(a) Distance measures of dispersion

(a) Range: Range is the difference between the highest and the lowest observed values i.e. Range (R)= Value of Highest observation – Value of lowest observation, i. e., $R = H - L$.

The range considers only highest and lowest values of a distribution and fails to take account of any other observation in the data set. Thus this measure ignores the nature of the variation among all other observations.

The range considers only highest and lowest values of a distribution and fails to take account of any other observation in the data set.

Advantages:

1. This is the simplest of all measures of dispersion
2. This measure is very easy to understand and easy to calculate
3. This measure does not depend on any measure of central tendency.

Disadvantages

1. This measure is based on the highest and lowest value, so it is affected by extreme values. Thus extreme values at either end or both ends of a data set can move the range markedly upward and as such distort understanding of the data.

This measure is based on the highest and lowest value, so it is affected by extreme values.

2. This measure can't be computed for data sets having open ended class interval(s).
3. It is not based on all observations in the data set.
4. It is sensitive to fluctuations of sampling.
5. This measure is unsuitable for mathematical treatment.

Interpretation Range

The range is no more than a rough measure of dispersion. It gives a comprehensive value for the data in the sense that it includes the limits within which all of the items occurred.

The Range is no more than a rough measure of dispersion. It gives a comprehensive value for the data in the sense that it includes the limits within which all of the items occurred. The range can be interpreted as an intensive measure of variability except in very small samples.

Application of Range

The Range can be used justifiably when we want a quick measure of dispersion or variability and do not have time to compute the other measure of variability. It is also used in the construction of class intervals for setting up a frequency distribution. Since range involves only two extreme values and is influenced strongly by them, it should be applied carefully. Its practical application is more in the situations where, the extreme variation of the values is usually absent or almost negligible as in the manufacturing industries. The chief use of range is in statistical quality control, that is to say, to control the average quality of manufactured products where the variation is limited. It can also be used in the statistical analysis of stock-exchange prices (where the high and low of the stock-prices are shown), daily temperature, weather forecasting etc.

(b) Fractile and Interfractile Range

In frequency distributions, a given fraction or proportion of the data lie at or below a fractile. The median is the 0.5 fractile, because half the data set is less than or equal to this value. Fractiles are similar to percentages. The interfractile range is a measure of the spread between two fractiles in a frequency distribution, that is, the difference between the values of the two fractiles.

Fractiles have special names, depending on the number of equal parts into which we divide the data. Fractiles that divide the data into 10 equal parts are called deciles. Quartiles divide the data into four equal parts and percentiles into 100 equal parts.

Advantages

1. This measure is easy to understand and not very difficult to calculate.
2. For distributions with open-ended class intervals this measure can be computed easily.
3. This measure is not affected by extreme values.

Disadvantages

1. This measure is not based on all observations.
2. This measure is not suitable for further algebraic treatment
3. This measure is affected by sampling fluctuations.

(c) Interquartile Range

The interquartile range measures approximately how far from the median we must go on either side before we can include one-half of the values of the data set. To compute this range, we divide our data into four parts, each of which contains 25 percent of the items in the distribution. The quartiles are then the highest values in each of these four parts, and the interquartile range is the difference between the values of the first and third quartiles:

$$\text{Interquartile range} = Q_3 - Q_1 \text{ where } Q_3 = 3^{\text{rd}} \text{ quartile and } Q_1 = 1^{\text{st}} \text{ quartile}$$

Figure 5.2.1 shows the concept of the interquartile range graphically. Notice in the figure that the widths of the four quartiles need not be the same.

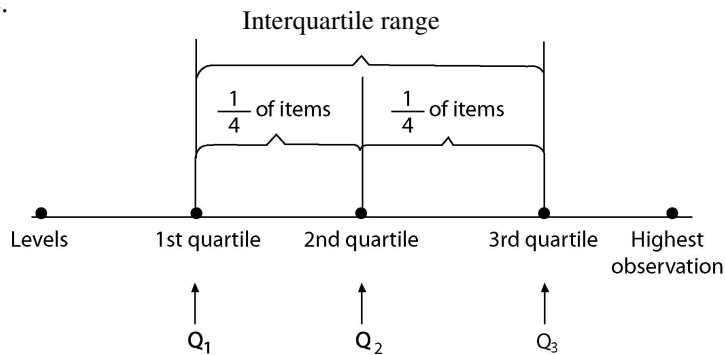


Figure :5.2.1

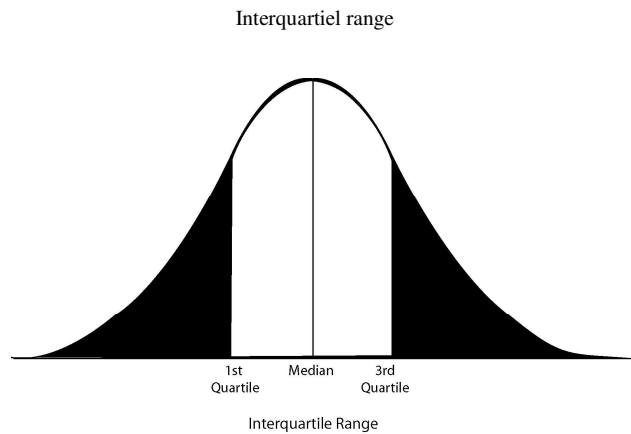


Figure: 5.2.2 Quartiles

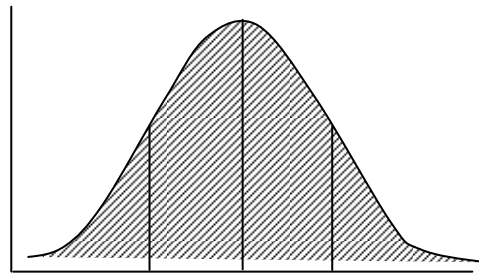
Figure 5.2.2, another illustration of quartiles, the quartiles divide the area under the distribution into four equal parts, each containing 25 percent of the area.

Interpretation of the Quartile Deviation (QD)

A small value of the quartile deviation (QD) reflects a little variation or range uniformity of the middle items. The QD is associated with the median and is considered whenever median is used as a measure of central tendency. This is usually the case in skew distribution. In normal distribution as Fig 5.2.3 (symmetric distribution) if we consider the

median and add and subtract one QD from each side of it, we will cut off approximately 50% of the cases (in the middle of the distribution). It should be noticed that even when the distributions are skewed, the check using the middle 50% of the case would work.

If we measure of 4QDs on each side of the median (ME), we will practically include all the cases. We can, however, state this briefly by saying that 8QDs approximately covers the range, that is, $R=8QDs$.



$-\alpha \dots \dots -IQD \quad ME \quad +IQD$

Fig. 5.2.3: Symmetric distribution

In a perfectly symmetric distribution, it is clear that Q_1 and Q_3 are at equal distances from the median but symmetric distributions are rare in actual life. In case of asymmetric distributions: for a right-skew distribution (Fig 5.2.4), Q_3 is farther away from the median than Q_1 and for a left skew distribution, Q_1 is farther away from the median than Q_3 .

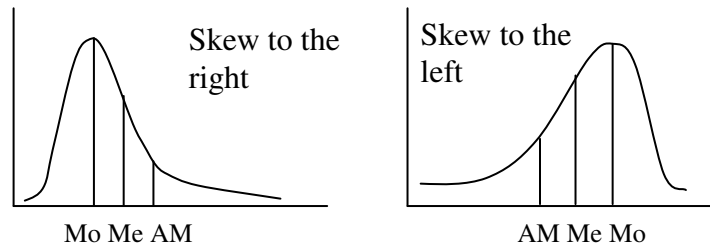


Fig. 5.2.4: Right and left skew distribution

Quartile deviation is defined by

$$Q.D = \frac{1}{2} [Q_3 - Q_1]$$

Where $Q_3 = 3^{rd}$ quartile

$Q_1 = 1^{st}$ quartile

Application of Quartile Deviation (QD):

Some statisticians view QD not as a measure of dispersion but as a measure of position or partition as it does not depend on any particular average and does not show the scatter around an average. However, the QD can be applied in comparing the variability of different distributions.

As it consider only the middle 50% of the observation, it is extremely useful in measuring variations in open-ended distributions. The QD is robust against outliers, easily interpretable and quickly computable without any need to square the numbers or add them up. The QD is used whenever the Median used as a measure of central tendency at the ordinal level of measurement. The QD is little used in statistical theory but is of practical value in dealing with empirical distribution, which are very skew.

Example 3.2.1:

The ages in years of 20 men are given below.

Find Range, 4th decile, 80th percentile & Interquartile of the ages.

50	56	55	49	52	57	56	57	56	59
54	55	61	60	51	59	62	52	54	49

Solution

Let, Range = Highest observation –lowest observation = 62-49 = 13

Arranging the data in ascending order we get

49,49,50,51,52, 52,54,54,55,55, 56,56,56,57,57,59,59,60,61,62

Now first quartile Q_1 is the average of 5th & 6th observation i.e. $Q_1 = 52$

The third quartile Q_3 is the average of 15th & 16th observation i.e. $Q_3 = 58$

So Range = $Q_3 - Q_1 = 58 - 52 = 6$

and Quartile Deviation = $\frac{1}{2} (Q_3 - Q_1)$

$$= \frac{6}{2} = 3$$

Activity:

Given that the total annual rainfall (in. m.m.) recorded in Bangladesh. The rainfall data are as follows: 3860, 3595, 4189, 4438, 4388, 1200, 1540, 1490, 1636, 1540, 2850, 1819.

Find out (i) Range (ii) Quartile Deviation

Self-Assessment Questions:

Short questions:

1. Define distance measures of dispersion.
2. Describe fractiles.
3. Discuss Interquartile range and show it graphically.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) When a distribution is symmetrical and has one mode, the highest point on the curve is called the
 - (a) Range
 - (b) Mode
 - (c) Median
 - (d) Mean
- (ii) Disadvantages of using the range as a measure of dispersion include all of the following except.
 - (a) It is heavily influenced by extreme values;
 - (b) It can change drastically from one sample to the next;
 - (c) It is difficult to calculate;
 - (d) Only two points in the data set determine it.

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The difference between the highest and lowest observations in a data set is called the quartile range_____.
- (ii) The interquartile range is based on only two values taken from the data set_____.
- (iii) A fractile is a location in a frequency distribution that gives a proportion (or fraction) of the data lies at or above_____.
- (iv) One disadvantage of using the range to measure dispersion is that it ignores the nature of the variations among most of the observations_____.
- (v) The interquartile range is a specific example of an interfractile range_____.
- (vi) It is possible to measure the range of an open-ended distribution_____.

Answer:

Multiple Choice Question

- (i) c (ii) b

True/False

- (i) F (ii) T (iii) T (iv) F (v) T (vi) F

Lesson 3: Measure of Dispersion (Average Deviation)

Lesson Objectives:

After studying this lesson, you will be able to:

- Understand the average deviation measures of dispersion;
- Define the average deviation measures variance, standard deviation and mean deviation;
- Discuss units of the measures of dispersion.

Introduction

These measures tell us an average distance of any observation in the data set from the mean or any central value of the distribution. These measures give the most comprehensive descriptions of dispersion. Two of these measures are variance/standard deviation and mean deviation.

Mean Deviation

One of the important absolute measures of dispersion is mean deviation, which can be measured in terms of Mean, Median and Mode. In calculating this measure we take the average of the absolute distances between some value (A) and each item, where A may take the values of Mean, Median, Mode etc. So the calculating formula for mean deviation is

$$\text{Mean Deviation from A} = \frac{\sum_{i=1}^N |X_i - A|}{N} \text{ for ungrouped data.}$$

Where, X_i = Observation item, N = Total Observations

$$\text{and Mean Deviation from A} = \frac{\sum_{i=1}^N f_i |X_i - A|}{N} \text{ for grouped data.}$$

Where

X_i = Mid-value of the i th class

f_i = Frequency of the i th class

K = Number of groups/classes

$$N = \sum_{i=1}^k f_i = \text{Total frequency.}$$

Example: 5.2 Calculate the mean deviation from the following data:

12.7, 14.8, 18.3, 16.1, 22.9, 25.3, 26.8, 26.3, 25.4, 20.6, 28.1, 26.4

Solution: Here, Mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

$$\text{So, mean deviation, } M_e = \frac{\sum |X_i - \bar{X}|}{N}$$

X1	$X_i - \bar{X}$	$ X_i - \bar{X} $
12.7	-9.27	9.27
14.8	-7.17	7.17
18.3	-3.67	3.67
16.1	-5.87	5.87
22.9	0.93	0.93
25.3	3.26	3.26
26.8	4.31	4.31
26.3	4.44	4.33
25.4	3.43	3.43
20.6	1.37	1.37
28.1	6.13	6.13
26.4	4.43	4.43

$$\therefore M_e = \frac{\sum |X_i - \bar{X}|}{n}$$

$$\therefore M_e = 54.76/12$$

$$= 4.56$$

Example 5.3

In an attempt to estimate the potential future demand, a resource organization did a study asking married couples how many colour Televisions a higher middle class family should own. For each couple, the resource organization averaged husband's and wife's response to get the overall couple response. The answers were then tabulated:

Number of TV's (x)	0	0.5	1.0	1.5	2.0	2.5
Frequency (f)	2	14	23	7	4	2

Calculate Mean, Variance, Standard deviation, and Average deviation.

Solution:

(0.5, 1.5) is approximately $\bar{x} \pm s$, so about 68 percent of the data, or $0.68(52)=35.36$ observations should fall in this range. In fact, 44 observations fall into this interval.

(0,2) is approximately $\bar{x} \pm 2s$, so about 95 percent of the data, or $0.95(52)=49.4$ observation should fall in this range. In fact, 50 observations fall into this interval.

Properties of the Mean Deviation

The MD may help us to find the percentage of observations falling in a range of Average $\pm$ MD. If we consider a normal distribution for which $AM=Me=Mo$ and which is symmetric, then the percentage of values falling in the range $(AM \pm MD)$ is the same as that of falling in the range $(Me \pm MD)$.

If the MD is comparatively small, then more than half of the items in the data fall within a small range around the average. This concentration would mean compactness of the distribution.

Application of the Mean Deviation

The application of the MD is overshadowed a large extent by the use of the standard deviation (SD). But the computation of the MD is less difficult. The MD when taken from the Median (Me) is theoretically preferable. The reason behind using the Me sometimes is that the sum of absolute deviations of the observations is minimum when the deviations are taken from the Me. However, despite the fact that the Median (Me) makes the sum of the absolute deviations more stable, mean is more frequently used. A simple average of the absolute deviations from the mean is often sufficient when an informal measure of dispersion is required. Informal in the sense that measure is not to be used in some complex mathematical problems.

Advantages

1. This measure is based on all observations.
2. This measure is rigidly defined
3. This measure is not affected by extreme values.

Disadvantages

1. This measure is not amenable to further algebraic treatment.
2. This measure can't be computed for open-ended class intervals.

(b) Population Variance and Standard Deviation

When the sum of the squared distance between the mean and each item in the population is divided by the total numbers of items in the population we get population variance which is symbolized by σ^2 (Sigma Squared) and the calculating formula for ungrouped data is.

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

$$\sigma^2 = \frac{\sum_{i=1}^N X_i^2}{N} - \mu^2$$

Where σ^2 = Population Variance
 X_i = Observation item
 μ = Population mean
 N = Total number of items or observations in the population.
 Σ = Sum of all items.
 For grouped data variance σ^2 is given by

$$\sigma^2 = \frac{\sum_{i=1}^k f_i (X_i - \mu)^2}{N} = \frac{\sum f_i X_i^2}{N} - \mu^2$$

Where,
 X_i = Mid-value of the i th class
 μ = mean of the observations
 f_i = Frequency of the i th class
 K = Number of classes/groups

$$N = \sum_{i=1}^k f_i = \text{Total frequency}$$

By squaring each distance, we make each number positive and, at the same time, assign more weights to the large deviations.

In equation for variance the middle expression, is the definition of σ^2 . The last expression, is mathematically equivalent to the definition but is often much more convenient to use in computing the value of σ^2 , since it frees us from calculating the deviations from the mean. However, when the x values are large and $x - \mu$ values are small, it may be more convenient to use the middle expression, to compute σ^2 .

Population Standard Deviation:

The population standard deviation is simply the square root of the population variance.

The population standard deviation, or σ , is simply the square root of the population variance. Because the variance is the average of the squared distances of the observation from the mean, the standard deviation is the square root of the average of the squared distances of the observations from the mean. While the variance is expressed in the square of the units used in the data, the standard deviation is in the same units as those used in the data. For ungrouped data the formula for standard deviation is

Population Standard Deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (X - \mu)^2}{N}} = \sqrt{\frac{\sum X^2}{N} - \mu^2}$$

Where,
 X = observation
 μ = population mean
 N = total number of elements in the population
 Σ = sum of all the values
 σ = population standard deviation.

When taking the square root of the variance to calculate the standard deviation, however, only the positive square root is considered.

Sample Measures of Dispersion:

Switching to sample standard deviation

Now we are ready to compute the sample statistics that are analogous to the population variance σ^2 and the population standard deviation σ . On the other hand, the sample variance s^2 and the sample standard deviation s .

Sample Standard Deviation:

To compute the sample variance and the sample standard deviation, we use the same formulas, replacing μ with $\bar{x}$ and N with $n-1$. The formulas for ungrouped data look like this:

Computing the sample standard deviation

Sample Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{\sum x^2}{n - 1} - \frac{n\bar{x}^2}{n - 1} \quad (3-17)$$

Sample Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\left(\frac{\sum (x - \bar{x})^2}{n - 1} \right)} = \sqrt{\frac{\sum x^2}{n - 1} - \frac{n\bar{x}^2}{n - 1}} \quad (3-18)$$

Where,

s^2 = sample variance

s = sample standard deviation

x = value of each of the n observations

$\bar{x}$ = mean of the sample

$n-1$ = number of observations in the sample minus 1

Statisticians has proved that if we take many samples from a given population, find the sample variance (s^2) for each sample, and average each of these together, then this average tends not to equal the population variance, σ^2 , unless we use $n-1$ as the denominator.

Example 5.4

The ages in years of 20 men are given below.

Find the standard deviation and mean deviation of the ages.

50 56 55 49 52 57 56 57 56 59
54 55 61 60 51 59 62 52 54 49

Solution

X	$X - \bar{X}$	$(X - \bar{X})^2$	X	$X - \bar{X}$	$(X - \bar{X})^2$
50	-5.2	27.04	54	-1.2	1.44
56	0.8	0.64	55	0.2	0.04
55	-0.2	0.04	61	5.8	33.64
49	-6.2	38.44	60	4.8	23.04
52	-3.2	10.24	51	-4.2	17.64
57	1.8	3.24	59	3.8	14.44
56	0.8	0.64	62	6.8	46.24
57	1.8	3.24	52	-3.2	10.24
56	0.8	0.64	54	-1.2	1.44
59	3.8	14.44	49	-6.2	38.44
Total		$\sum(X_i - \bar{X})$	1,104	$\sum(X_i - \bar{X})$	285.20

$$\bar{X} = \frac{\sum X}{n} = \frac{1,104}{20} = 55.2 \text{ Years,}$$

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}} = \sqrt{\frac{285.20}{19}} = 3.874 \text{ Years}$$

Mean deviation from mean

$$MD = \frac{\sum |X - \bar{X}|}{N} = \frac{5.0}{20} = 0.25$$

Example 5.5

In an attempt to estimate the potential future demand, a resource organization did a study asking married couples how many colour Televisions a higher middle class family should own. For each couple, the resource organization averaged husband's and wife's response to get the overall couple response. The answers were then tabulated:

Number of TV's (x)	0	0.5	1.0	1.5	2.0	2.5
Frequency (f)	2	14	23	7	4	2

Calculate Mean, Variance, Standard deviation, and Average deviation.

Solution:

(a)

x	f	f × x	x - $\bar{X}$	(x - $\bar{X}$) ²	f(x - $\bar{X}$) ²	f x - $\bar{X}$
0	2	0	-1.0288	1.0585	2.1170	2.0576
0.5	14	7	-0.5288	0.2797	3.9155	7.4032
1	23	23	-0.0288	0.0008	0.0191	0.6624
1.5	7	10.5	0.4712	0.2220	1.5539	3.2984
2	4	8	0.9712	0.9431	3.7726	3.8848
2.5	2	5	1.4712	2.1643	4.3286	2.9424
Total	52	53.5			15.7067	20.2488

$$\bar{X} = \frac{\sum fx}{n} = \frac{53.5}{52} = 1.0288 \text{ TV's} \cong 1.03 \text{ TV's}$$

$$s^2 = \frac{\sum f(x - \bar{X})^2}{n-1} = \frac{15.707}{51} = 0.3080 \quad \text{so} \quad s = \sqrt{0.3080} = 0.55$$

$$\text{Average Deviation} = \frac{\sum f |x - \bar{X}|}{n} = \frac{20.2488}{52} = 0.39 \text{ Tv's.}$$

Mean = 1.028,

Variance = .3080

S.D = .55

and Average deviation = .39

Example 5.6

A company requires that chilled food cabinets in its supermarkets must maintain an average hourly temperature of $3.75^{\circ}\text{C} \pm 0.5^{\circ}\text{C}$. The manager at one of the supermarkets suspects that the performance of one of the shop's cabinets fail to meet this standard and therefore decides to monitor its performance hourly over a 30 day period with the following results:

Temperature (C)	Frequency
0-1	1
1-2	11
2-3	123
3-4	322
5-5	223
6-6	39
6-7	1

Find the mean deviation to assess whether the equipment conforms to the company's policy.

Solution:

The question can be interpreted as: 'What is the hourly mean temperature of the cabinet and by how much, on average, does it deviate from the mean?' In other words find the mean hourly temperature and the mean deviation.

The data are grouped, therefore, we use the following table for computation.

Table for computing Mean Deviation

Temperature (C)	Class mid-point X_i	Frequency f_i	$f_i X_i$	$ X_i - \bar{X} $	$f_i X_i - \bar{X} $
0-1	0.5	1	0.5	3.22	3.22
1-2	1.5	11	16.5	2.22	24.42
2-3	2.5	123	307.5	1.22	150.06
3-4	3.5	322	1127	0.22	70.84
4-5	4.5	223	1003.5	0.78	173.94
5-6	5.5	39	214.5	1.78	69.42
6-7	6.5	1	6.5	2.78	2.78
Total		720	2676		494.68

$$\text{Mean} = \bar{X} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i}$$

Where: f_i = frequency of i th class

X_i = class mid-point

$$\text{So } \bar{X} = 2676/720 = 3.72^{\circ}\text{C}$$

$$\text{MD} = \text{Mean Deviation} = \frac{\sum_{i=1}^n f_i |X_i - \bar{X}|}{\sum_{i=1}^n f_i} = 494.68/720$$

$$\therefore \text{MD} = 0.69^{\circ}\text{C}.$$

Comment:

The mean temperature of 3.72°C, although slightly low, is close to the company's standard. However, the mean deviation of 0.69°C exceed the limit of 0.5°C. Therefore, the equipment does not comply with the company's policy.

Advantages: Standard Deviation (S.D)

1. This measure is based on all observations.
2. This measure is rigidly defined
3. This measure is less affected by sampling fluctuation and has relatively a small sampling error.
4. Further algebraic treatment can be done on this measure.
5. Standard deviation can be easily calculated for coded/changed data.
6. Mathematics of sampling theory is much simpler for this measure than for the other measures of dispersion.

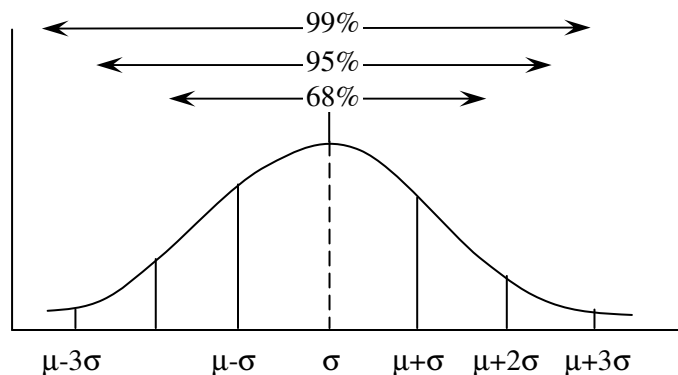
Disadvantages: Standard Deviation

1. Computation of this measure needs basic knowledge of mathematics.
2. This measure is affected by extreme values.
3. This measure can't be computed for distributions having open-ended class intervals.

Interpretation of the Standard deviation

The SD is particularly useful when the population has a special type of frequency distribution, called the normal distribution.

The SD can be viewed as parameter, which can provide a lot of information when combined with other parameters. The SD is particularly useful when the population has a special type of frequency distribution, called the normal distribution. It is possible then to find the percentage of observations falling within distances of one, two or three SDs from the mean (AM). Thus the proportion of observations can be expressed in terms of the SD units. About 68.27 percent, 95.45 percent and 99.73 percent of the observations will lie within the regions ($\mu \pm 1\sigma$), ($\mu \pm 2\sigma$) and ($\mu \pm 3\sigma$), respectively where, μ and σ are the AM and the SD of normal distribution. Thus, in a normal curve, 3 times the SD constitutes practically the whole range of the values in the distribution as shown in the following figure 5.6.



Application of the Standard Deviation (SD)

The SD has manifold applications or uses. Among the various measures of dispersion, the SD is the most important one, mainly because it is used in so many other statistical operations. The SD is the most useful statistical measure in sampling. However, in basic statistics also it has a wide application as a measure of dispersion.

The SD can be used in conjunction with the mean to indicate the percentage of items falling within a specified range. This application is particularly associated with the normal curve.

The SD can compare the extent of variations or degree of uniformity of two or more data sets. But it is to be noted that acceptable comparisons can only be made when the data values are expressed in the same units.

The SD helps us to locate the place of an item in the distribution. The SD can also be used to make comparisons for items in distributions that differ in order of magnitude of units employed. Another important application of the SD is to judge how much representative the AM is as a measure of central tendency. Thus, if two or more comparable distributions have the same mean, then the mean is the most representative in the distribution with the smallest SD.

The range is used in connection with distribution free methods. It is mainly used when a quick measure of variability is required and when we need to find class intervals for the information of a frequency distribution. The QD should be used to measure the variation in open-ended distribution and when we want to eliminate the affect of extreme values. The QD can be used as a rough measure of dispersion, which is superior to the range. The application or use of the MD is overshadowed to a large extent by the use of the SD. But when an informal measure of dispersion is required, which would not be used in complex mathematical problems, and then MD could be used. The SD is the most widely used measure of dispersion. In sampling, the SD is the most useful measure. It is also used widely in basic statistics and in many statistical operations. The SD can be used to compare the variability or degree of uniformity of two or more data sets.

Activity:

The following data are given, find out, (i) mean deviation, (ii) standard deviation and variance

Class interval	<2000	2000-2500	2500-3000	3000-3500	3500-4000
No. of days	15	45	73	30	10

The SD can compare the extent of variations or degree of uniformity of two or more data sets.

Another important application of the SD is to judge how much representative the AM is as a measure of central tendency.

The range is used in connection with distribution free methods.

The QD can be used as a rough measure of dispersion, which is superior to the range.

The SD is the most widely used measure of dispersion

Self-Assessment Questions:

Short Questions:

1. Define average measures of dispersion.
2. Discuss the Measures Variance, Standard Deviation and Mean Deviation.
3. Describe about the units of the average measures of dispersion.

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The most meaningful measure of dispersion is the
 - (a) Variance
 - (b) Mean absolute deviation
 - (c) Range
 - (d) Standard deviations
- (ii) The advantage of using the interquartile range versus the range as a measure of variation is
 - (a) It is easier to computes
 - (b) It utilizes all the data in its computation
 - (c) It gives a value that is closer to the true variation
 - (d) It is less affected by extremes in the data.
- (iii) The following data reflect the number of customers who return merchandise for a refund on Sunday. Following data reflect the population of all 10 Sundays for which data are available.

40	12	17	25	9
46	13	22	16	7

Answer that this same pattern of data were replicated for the next 10 days. How would this affect the standard deviation for the new population with 20 items?

- (a) The standard deviation would be doubled
 - (b) The standard deviation would be in half
 - (c) The standard deviation would not be changed
 - (d) This is no way of knowing the exact impact without knowing how the mean is changed.
- (iv) Income in a particular market area are known to be right-skewed with a mean equal to Tk. 33100. In a report insured recently, a money on Tk. 26700 to Tk. 39500. Given there facts, what is the standard deviation for the income in this makes area?
 - (a) Tk. 6400
 - (b) Tk. 3200
 - (c) Approximately 2533
 - (d) None of the above
- (v) Suppose the book cost for one Somerton's books are given below for a sample of five school students. Calculate the variance of the book costs.

200	256	375	125	250

- (a) 929.65
 - (b) 8642.5
 - (c) 83.4505
 - (d) 6914.0

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) One of the reasons that the standard deviation is preferred as a measure of variation over the variance is that the standard deviation is measured in the original units
- (ii) The standard deviation is equal to the square root of the variance
- (iii) The variance, like the standard deviation, takes into account every observation in the data set
- (iv) The standard deviation is a measure of variation of the data around the median.
- (v) The measure of dispersion most often used by statisticians is the standard deviation

Answer:

Multiple-Choice Question:

1. (i) d (ii) d (iii) c (iv) c (v) c

True/False

2. (i) T (ii) T (iii) F (iv) F (v) T

Lesson 4: Relative Measures of Dispersion

Lesson Objectives:

After studying this lesson, you will be able to:

- Discuss advantages and disadvantages of the measures of dispersion;
- Interpret the measures of dispersion;
- Describe applications of the measures of dispersion;
- Comparative description of the different measures of dispersion.

Introduction

To really know the dispersion of a set of data we must know how the measures of dispersion compare with their central values.

Relative Measures of Dispersion:

The measures of dispersion so far discussed are absolute measures i. e, the measures retain the unit of measurement of the variable. Naturally measures with different units for different distributions cannot be compared. So a sound basis of comparisons can be used by eliminating the disturbing influence of units and it possible to do so by dividing each measure of dispersion by its respective average (not in the cases when quartile deviation and range are used as these are not measured around an average). The result thus obtained gives us a measure of relative variation.

Four measures of relative variations will be discussed here and these are:

Relative measure of dispersion:

Relative measures of dispersion mean the measures of percentage change of variance of a set of observation. This measure is free of unit of variable under study. There are:

- (i) The coefficient of Range (CR)
- (ii) The coefficient of Quartile Variation (CQD)
- (iii) The coefficient of Mean deviation (CMD)
- (iv) The coefficient of Variation (CV)

(i) The Coefficient of Range (CR)

This is the least used measure of relative variation and expressed in terms of the range or the largest and smallest values of a series. The coefficient of range is defined as the ratio of the range to the sum of the largest and smallest values of a series. It is usually expressed as percentage and is denoted by CR. Thus:

$$CR = \frac{X_u - X_l}{X_u + X_l} \times 100$$

Where, X_u and X_l denote the largest and smallest values of the series, respectively.

The coefficient of range is defined as the ratio of the range to the sum of the largest and smallest values of a series.

The coefficient of quartile deviation is defined as the ratio of the difference between the third and first quartiles to the sum of them.

(ii) The Coefficient of Quartile Deviation (CQD)

The measure of relative variation can be expressed in terms of quartiles. The coefficient of quartile deviation is defined as the ratio of the difference between the third and first quartiles to the sum of them. It is usually expressed as percentage and is denoted by CQD.

$$\text{Thus CQD} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$$

Where, Q_1 and Q_3 denote the first and third quartile, respectively.

(iii) The Coefficient of Mean Deviation (CMD)

This is the second important measure of relative variation. The coefficient of mean deviation is defined as the ratio of the Mean deviation (MD) to the arithmetic mean (AM) or the median (Me), depending on which one is used to obtain the mean deviation. It is usually expressed as percentage and is denoted by CMD. Thus:

$$\text{CMD} = \frac{\text{MD}}{\text{AM}} \times 100 \quad \text{or} \quad \text{CMD} = \frac{\text{MD}}{\text{Me}} \times 100$$

(iv) Co-efficient of Variation (CV)

A relative measure will give us a feel for the magnitude of the deviation relative to the magnitude of the central value. Co-efficient of variation is one of such relative measures of dispersion. It expresses standard deviation as percentage of the mean i.e. the unit of measure is “Percent”.

The computing formula for population co-efficient variation is:

$$\text{Popular Co-efficient of variation} = \frac{\sigma}{\mu} \times 100$$

Where σ = standard deviations, μ = Mean

The CV is used to compare distribution where the importance of the variability is related to the average size of the thing or concept measured. In many series the mean and deviation tend to change together, then the CV can be used as a measure.

Example 5.7

The combined grade point average in different semesters of two MBA students of Bangladesh Open University is shown below:

Student A	2.5	2.5	3.0	3.5	3.5	4.0	3.5	3.5
Student B	2.5	3.0	4.0	4.0	4.0	2.0	2.5	4.0

Which student would be considered better throughout the courses of studies?

Solution: The mean and S.D. student A are calculated at first

$$X_A = 0.50$$

$$S.D._A = 3.25$$

$$\therefore C.V._A = .50/3.25 \times 100 = 15.38\%$$

Again mean and S. D for student B:

$$X_B = 0.79$$

$$S. D_B = 3.25$$

$$\therefore C. V_B = 79/3.25 \times 100 = 24.31\%$$

If observe that $C. V_A < C. V_B$ so, Student A is better than B.

Activity:

Two set of data Set A & Set B are given below:

Set A	12	48	70	35	15
Set B	5	38	80	50	7

Which Set of data are making better performance comment on it?

A Few Remarks on the Measures of Relative Variation

The above measures of relative variation can also be expressed in decimal forms but the usual practice is to use them in percentage form.

The measures of relative variation are not some new measures of dispersion and these measures do not introduce new statistical concepts. These measures are the result of some adjustments to the measures of absolute variation so as to enable us to compare the variations among two or more data series or distributions. These are also known as **adapted measures**. However, in comparing the variations in two or more distributions by the measures of relative variation, we must be careful to use the same type of measure. Thus the CV in one distribution must be compared with the CV in other distributions, not with the CMD, CQD, or CR.

The absolute measures of dispersion are used to compare distributions whose means are close to each other. On the other hand Co-efficient of variation is used to compare the amount of variation in data groups that have different means.

The absolute measures of dispersion are used to compare distributions whose means are close to each other.

Standardized Variables: Standard Scores

The variable that measures the deviation from the mean in units of the standard deviation is called a standardized variable, is a dimension less quantity (i.e., independent of the units used), and is given by

$$z = \frac{X - \bar{X}}{S}$$

If the deviations from the mean are given in units of the standard deviation, they are said to be expressed in standard units, or standard scores. These are of great value in the comparison of distributions.

Example-5.8

A brewing company wishes to launch a new canned lager. It has close links with a major supermarket chain which will only permit a promotion of the lager in two of its stores. The selected stores monitor their sales of all brands of canned lager over a weekend period with the following results, which were communicated to the brewery:

Sales value of Lager (Tk.)	Frequency	
	Store A	Store B
0-2.5	27	1
2.5-5	114	3
7.5-10	530	142
10-12.5	504	328
12.5-15	334	498
15-17.5	121	504
17.5-20	29	351
20-22.5	5	110
22.5-25	2	29
25-27.5	1	3
	2000	2000

What advice would you give the management of the brewery on the differences and similarities in the pattern of expenditure in stores A and B.

Solution :

For Supermarket A

Sales value of Lager (Taka)	Frequency f_i	mid-point X_i	$f_i X_i$	$f_i X_i^2$
0-2.5	27	1.25	33.75	42.19
2.5-5	114	3.75	427.5	1603.13
5-7.5	333	6.25	2081.25	13007.81
7.5-10	530	8.75	4637.5	40578.13
10-12.5	504	11.25	5670.0	63787.5
12.5-15	334	13.75	4592.5	63146.88
15-17.5	121	16.25	1966.25	31951.56
17.5-20	29	18.75	543.75	10195.31
20-22.5	5	21.5	106.25	2257.81
22.5-25	2	23.75	47.5	1128.13
25-27.5	1	26.25	26.25	689.06
	2000		20132.5	228387.51

$$\bar{X}_A = \frac{\sum f_i X_i}{N = \sum f_i}, \quad \bar{X}_A = \frac{20132.5}{2000}, \quad \text{Mean} = \text{C } 10.07$$

$$\text{Variance } (s^2) = \frac{\sum f_i X_i^2}{N} - \left[\frac{\sum f_i X_i}{N} \right]^2$$

$$= \frac{228.387.51}{2000} - \left(\frac{20132.5}{2000} \right)^2$$

$$= 114.19 - 101.4$$

$$\text{Variance} = 12.79$$

$$\text{Standard deviation} = \sqrt{12.79} = \text{Tk.} 3.58$$

For Supermarket B

Sales value of Lager (Tk.)	Frequency f_i	mid-point X_i	$f_i X_i$	$f_i X_i^2$
0-2.5	1	1.25	1.25	1.56
2.5-5	3	3.75	11.25	42.19
5-7.5	31	6.25	193.75	1210.94
7.5-10	142	8.75	1242.5	10871.88
10-12.5	328	11.25	3690.0	41512.50
12.5-15	498	13.75	6847.5	94153.13
15-17.5	504	16.25	8190.0	133087.50
17.5-20	351	18.75	6581.25	123398.44
20-22.5	110	21.25	2337.5	49671.88
22.5-25	29	23.75	688.75	16357.81
25-27.5	3	26.25	78.75	2067.19
	2000		29862.50	472375.02

$$\bar{X}_B = \frac{\sum f_i X_i}{N}, \quad \bar{X}_B = \frac{29862.25}{2000}, \quad \text{Mean} = \text{Tk. } 14.93$$

$$\text{Variance } (s^2) = \frac{\sum f_i X_i^2}{N} - \left[\frac{\sum f_i X_i}{N} \right]^2$$

$$= \frac{472375}{2000} - \left(\frac{29862.25}{2000} \right)^2$$

$$= 236.19 - 14.93^2$$

$$= 236.19 - 222.9$$

$$\text{Variance} = 13.29$$

$$\text{Standard deviation} = \sqrt{13.29} = \text{Tk. } 3.65$$

Coefficient of variation

$$\text{CV of } V_A = s/\bar{X} = (\text{Standard deviation})/\text{mean}$$

$$= 3.58/10.07 = 0.36$$

$$\text{CV of } V_B = 3.65/14.93 = 0.24$$

The mean amount spent by those buying lager in store A is lower than that of store B whilst the variability of the data measured by both the variance and standard deviation are almost identical. The coefficients of variation, however, show that the relative dispersion of store A is greater than that of store B. From the data provided from that particular weekend it would appear that store B would be the best location to promote the new lager.

Self-Assessment Questions:

Short questions

1. What are relative measures of dispersion? Mention their necessity.
2. Define different measures of relative dispersion.
3. Mention the situations in which we can use relative measures.

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The smaller the spread of score around the arithmetic mean,
 - (a) The smaller the interquartile range
 - (b) The smaller the standard deviation
 - (c) The smaller the coefficient of variation
 - (d) All the above.
- (ii) If one were to divide the standard deviation of a population by mean of the same population and multiply this value by 100, one would have calculated the:
 - (a) Population standard score
 - (b) Population variance
 - (c) Population standard deviation
 - (d) Population coefficient of variation
- (iii) The heights (in inches) of 10 adults males are listed below. Find the sample standard deviation.

70	72	71	70	69	73	69	68	70	71

- (a) 70
 - (b) 3
 - (c) 2.38
 - (d) 1.49
- (iv) If nothing is known about the shape of a distribution, what percentage of the observations fall within 2 standard deviation of the mean?
 - (a) Approximately 5%
 - (b) Approximately 95%
 - (c) At least 75%
 - (d) At most 25%
- (v) The following data are the yields, in crops, from a farmer's last 10 years.

375	210	150	147	429	189	320	580	407	180

Find the interquartile Range

- (a) 433
 - (b) 279
 - (c) 265
 - (d) 227
- 2. Write "T" if the statement is true and "F" if the statement is false:**
- (i) The variance of a population is the arithmetic average of the squared deviations about the population mean.
 - (ii) A relative measures given the magnitude of the deviation relative to the central value.
 - (iii) Relative measures can be expressed in decimal form.
 - (iv) The sample standard deviation for the group of data items, 10, 10, 10, 13, 16, 16, 16, is 3.
 - (v) The Z score for the value 91, when the mean is 94 and the standard deviation is 4 will be : $Z = - 0.75$.

Answer:

Multiple-Choice Question: 1. (i) d (ii) d (iii) d (iv) c (v) d

True/False: 2. (i) T (ii) F (iii) T (iv) T (v) T

Exercise

1. (a) Define measure of variation. What are the different measures of variation? Explain the utility of measures of variation.
- (b) The following frequency table relates with the distribution of number of passengers travel by bus in different days from a rural area to an urban area.

Class interval of passengers	<2000	2000-2500	2500-3000	3000-3500	3500-4000
No. of days	15	45	73	30	10

Calculate standard deviation and co-efficient of variation.

2. (a) Define mean deviation, standard deviation and coefficient of variation. Illustrate the merits and demerits of these three measures of dispersion.
- (b) The means and standard deviations of three sets of samples are 10.5, 18.6, 9.3, 2.5, 4.6, 7.2 respectively. The frequencies of three sets are 40, 62, 58. Find the coefficient of variation of combined sample observations.
3. (a) Find mean and variance of first n natural numbers
- (b) Give a comparative study of important measures of dispersion.
4. (a) What is meant by mean deviation? Show that mean deviation is least when deviation is measured from median.
- (b) Show that mean square deviation is minimum if deviation is measured from mean.
- (c) Given two frequency distributions as follows:

Class interval	20-25	25-30	30-35	35-40	40-45	45-50
Frequency, f_{1i}	4	17	32	18	10	6
Frequency, f_{2i}	15	18	21	33	47	12

Calculate co-efficient of variation and comment on it.

5. The distribution of computer sells in two different shops in different days are shown below:

Class interval of Computers	4-6	6-8	8-10	10-12	12-14	14 ⁺
No. of days of shop-1	5	12	15	18	25	12
No. of days of shop-2	18	42	20	13	7	2

Calculate co-efficient of variation on both days and comment on it?

6.

Activity

The head chef of the flying Taco has just received two dozen tomatoes from her supplier, but she isn't ready to accept them. She knows from the invoice that the average weight of a tomato is 7.5 ounces, but she insists that all be of uniform weight. She will accept them only if they average weight is 7.5 ounces and the standard deviation is less than 0.5 ounce. Here are the weights of the tomatoes:

6.3	7.2	7.3	8.1	7.8	3.8	7.5	7.8	7.2	7.5	8.1	8.2
8.0	7.4	7.6	7.7	7.6	7.4	7.5	8.4	7.4	7.6	6.2	7.4

What is the chef's decision and why?

7. (a) What is standard deviation? Write down the merits and demerits of standard deviation.
- (b) Find percentage change of variation of the observations from mean, median and mode of the following distribution.

Class interval	50-60	60-70	70-80	80-90	90-100
Frequency	28	42	64	32	16

CORRELATION ANALYSIS



In the preceding units, we considered descriptions or inferences using data collected with respect to some single measurable characteristic of the elements of a statistical universe. Since only one variable was involved, this type of statistical investigation is called *univariate* analysis. However, the elements in a statistical universe usually have more than one measurable characteristic, and it often happens that some are known or easily measured while others may, for a variety of reasons, be difficult to measure. If a variable which is easily measured is closely related to one which is difficult to measure, and if the relationship can be expressed mathematically, it should be possible to use the score obtained for the easily measured variables as a basis for estimating the characteristic which is hard to measure. This type of statistical study is called *bivariate* analysis. If the variable which is hard to measure is estimated from two or more related variables, the technique is called *multivariate* analysis.

This unit is primarily concerned with the logic, procedures, and business application of bivariate analysis

Lesson 1: Fundamentals Concept of Correlation

Lesson Objectives:

After completing this lesson, you will be able to:

- To know the historical development of concept of correlation;
- Define and discuss meaning of correlation analysis;
- To understand utilities of correlation analysis;
- To understand difference between correlation and causation;
- Understand why reasons the existence of correlation;
- Understand the various types of correlation with examples.

Introduction

In business, the key to decision making often lies in the understanding of the relationships between two or more variables. For example, Financial experts, in studying the behavior of the share market, might find it useful to know if the interest rates on bonds are related to the price of shares, a marketing executive might want to know how strong the relationship is between advertising amount and sales amount for a product. A company engaged in the distribution business may determine that there is a relationship between the price of fuel and their own transportation costs. In this unit, we will study the concept of correlation and how it can be used to estimate the relationship between two variables. Correlation is a measure of the degree of relatedness of variables.

Correlation analysis has its roots in the 19th century and is primarily attributed to the work of **Sir Francis Galton** (1822–1911), a British polymath. He introduced the concept of correlation while studying the relationship between parents' and offspring's traits, particularly height. His work laid the foundation for the statistical study of relationships between variables.

Building upon Galton's ideas, **Karl Pearson** (1857–1936) formalized the mathematical framework for correlation. He developed the **Pearson correlation coefficient (r)**, a measure that quantifies the strength and direction of a linear relationship between two continuous variables. This coefficient remains widely used today in statistics.

In the 20th century, other types of correlation measures were developed to handle non-linear relationships, categorical data, and rank-based data, which includes **Spearman's rank correlation**.

Today, correlation analysis is a fundamental statistical tool used in various fields, including business, economics, finance, psychology, and the social sciences, to assess relationships between variables and support decision-making.

Definition of Correlation:

So far, we have studied problems relating to one variable only. However, in practice we come across a large number of problems involving the use of two or more than two variables. For example, we may be interested in

Correlation is a measure of the degree of relatedness of variables.

Pearson correlation coefficient (r), a measure that quantifies the strength and direction of a linear relationship between two continuous variables.

Other types of correlation measures were developed to handle non-linear relationships, categorical data, and rank-based data, which includes Spearman's rank correlation.

The statistical tool with the help of which relationship between two variables is studied is called Correlation.

finding out the relationship between the heights of husbands and wives at the time of marriage. If the height of the bridegroom is represented by x in general and that of bride by y , then to each marriage there corresponds a pair of values (x_1, y_1) of the variables x and y . We may be interested in finding out whether tall men tend on the average to wed tall women, or they choose short women? The statistical tool with the help of which this relationship between two variables is studied is called **Correlation**. There may be other variates also, such as heights and weights of students in a class, advertisement and sales, and price and demand of a commodity, records of rainfall and yields of crops etc.

Whenever two variables are so related that a change in the value of one is accompanied by a change in the value of the other, in such a way, that

- (i) an increase in the one is accompanied by an increase or decrease in the other, or
- (ii) a decrease in the one is accompanied by a decrease or increase in the other then the variables are said to be correlated. Let us consider some other examples:

Examples of (i)

- (a) an increase in the amount of rainfall accompanied by an increase in the sales of raincoats;
- (b) an increase in the price of a commodity accompanied by a decrease in its demand;
- (c) increase in the heights of children accompanied by increase in their weights.

Examples of (ii)

- (a) a decrease in price of a commodity accompanied by an increase in its demand;
- (b) a decrease in price of a commodity accompanied by a decrease in its supply.

A.M. Tuttle gives a very simple definition of correlation as: "An analysis of the correlation of two or more variables is usually called correlation."

The extent to which individual cases of one variable co-vary with those of another is represented by a coefficient of correlation. A correlation coefficient quantifies the degree of correlation between two or more variables. Correlation pertains to the relationship between two variables. Correlation quantifies the extent of association between variables. It does not examine a singular series. Thus far, our focus has predominantly been on the attributes of a singular variable, including central tendency and variability. We will now examine the relevant statistical approaches for analyzing the relationship between two or more variables.

Utility of the Study of Correlation

The utility of the study of correlation is obvious from the following:

- (i) The determination of the existence and extent of the relationship between two phenomena is one of the most important problems in statistics and the answers to many practical questions turn on the

An analysis of the correlation of two or more variables is usually called correlation.

connection between two factors, e.g., association between smoking and the incidence of lung cancer. The correlation coefficient helps us in measuring the extent of relationship in one figure only.

(ii) Sometimes existence of relationship between two or more variables enables us to predict what will happen in future, e.g., if production of wheat has increased, other factors remaining constant, we may expect a fall in price.

(iii) It is possible to estimate the value of one variable using the value of another, provided that we know the degree of relationship between the two variables. The regression equations covered in the upcoming unit will be useful for this purpose. However, analysis based on correlation can be deceiving if not applied with caution. Consequently, this statistical method necessitates extreme caution.

Correlation and Causation

The correlation coefficient alone quantifies covariation. It is a quantitative assessment of the degree of correlation among two or more variables. It does not establish causation. The causal relationship is not established by it. Correlation does not inherently imply causality. Causation invariably leads to correlation. For instance, there exists a direct causal relationship between activating a switch and the illumination of a light. Light is the consequence of activating the switch.

The Following table clearly highlights the differences between **correlation** and **causation** in terms of their nature, examples, and evidence required.

Aspect	Correlation	Causation
Definition	A statistical relationship or association between two variables. Example: There is a correlation between ice cream sales and drowning deaths (both increase in summer).	A direct cause-and-effect relationship between two variables. Example: Smoking causes an increased risk of lung cancer.
Nature of Relationship	Shows that two variables change together, but one does not necessarily cause the other.	One variable directly causes a change in another.
Direction	Can be positive, negative, or zero (no relationship).	One variable causes the other to change.
Purpose	To measure the strength and direction of a relationship between variables.	To establish a direct cause-and-effect link between variables.
Establishing Evidence	Can be observed through statistical methods (e.g., Pearson's correlation coefficient).	Requires experimental evidence, often with controlled conditions (e.g., randomized controlled trials).
Causality	Does not imply one variable causes the other.	Directly implies that one variable causes the other to change.

Correlation does not inherently imply causality. Causation invariably leads to correlation.

Correlation measure the strength and direction of a relationship between variables and on the other hand causation establish a direct cause-and effect link between variables.

Aspect	Correlation	Causation
Example of Fallacy	Assuming that because two things are correlated, one causes the other. (e.g., higher ice cream sales cause drowning deaths).	Understanding that one thing causes another only when proven by research and evidence.
Used	Descriptive studies, exploratory research, and analysis of patterns.	Experimental research, clinical trials, and scientific studies to prove causal relationships.

Correlation may exist because of any one, or a combination of the following reasons:

(i) **The Correlation may be due to pure chance.** There may be correlation between the price of sugar and height of students. This correlation between the price of sugar and height of students has no significance whatever in indicating causal connection. This type of relationship is known as "*spurious correlation*" or "*nonsense correlation*." Another case of spurious relationship between two variables is when they do not operate in the same physical or social systems, i.e., they have nothing to do with each other. For example: there has been a very close relationship between the average salary of an Government Official and the price of Ice-cream in Dhaka city.

(ii) **Both the correlated variables may be influenced by one or more other variables.** Sometimes relationship between two variables is not due to cause and effect. Changes in both the variables are due to their separate relationship to a third variable. For example, the correlation between the price of sugar and of shoes may be due to the fact that the two are related to a third variable of money value, i.e., a general rise in retail prices of all goods. This correlation between the price of sugar and shoes has no significance whatever in indicating causal connection.

Let us take another example. If a student, after putting in ten hours of preparation, obtains 60 percent marks in a statistics test, then perhaps we cannot conclude that there is a causal relationship between the hours of study and the marks. This is so because scoring of marks at an examination depends upon various factors. Still there may be correlation between the two (as is shown in the following Table 6.1), as existence of correlation does not necessarily imply causation, though causation will always result in correlation.

Table 6.1: Marks in a Statistics Test

Serial No of Students	Number of H ours of Study	Marks obtained
1	10	50
2	11	55
3	12	60
4	13	65
5	14	70
6	15	75

(iii) Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect. It is occasionally challenging to determine which variable is the cause and which is the consequence, even when a link exists between the two. The demand and supply factors of a product may interact reciprocally. According to an economic principle, when the price of a commodity rises, its demand diminishes. The price is the cause, while demand is the effect. However, the need may remain unchanged, similar to that of salt. The increase in population or the escalation of the general price level may compel prices to ascend. Consequently, the cause is the heightened demand, and the impact is the price. The two factors are interdependent, making it challenging to determine which is the cause and which is the impact.

Consequently, the determination of whether the alterations in variables signify causality must rely on evidence beyond the extent of correlation. The existence of correlation between two variables does not inherently indicate direct causation, since causality invariably leads to correlation.

Types of Correlation

Correlation is described in the following three ways:

- (i) Positive or negative
- (ii) Simple, partial or multiple
- (iii) Linear or non-linear

(i) **Positive or negative correlation:** Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both the variables are moving (i.e., changing) in the same direction, i.e., if with an increase in one variable, the other also increases or with a decrease in one variable, the other also decreases, correlation is said to be *positive*. On the other hand, if they are moving in opposite directions, i.e. with an increase in one variable, there is a decrease in the other variable or vice versa correlation is said to be *negative*. Following examples would clarify the above points.

An increase in one variable, the other also increases or with a decrease in one variable, the other also decreases, correlation is said to be positive.

Positive Correlation

(i)

X	Y
5	10
7	12
9	14
10	16
12	17
15	20

(ii)

X	Y
40	10
50	8
30	7
20	5
10	4
5	2

Negative Correlation

(i)

X	Y
5	20
6	17
9	16
10	14
12	12
15	10

(ii)

X	Y
40	2
50	4
30	5
7	5
10	8
5	10

Multiple correlation consists of the measurement of the relationship between a dependent variable and two or more independent variables.

Partial correlation measure the correlation between a dependent variable and one particular independent variable when all other variables involved are kept constant.

The amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then the correlation is said to be linear.

If the two variables are such that the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable, then the correlation would be non-linear or curvilinear.

(ii) Simple, Partial and Multiple Correlations: The study of correlation for two variable involves application of simple correlation. When more than two variables are involved then the problem may be of either multiple or partial correlation. In multiple correlation three or more variables are studied simultaneously. Multiple correlation consists of the measurement of the relationship between a dependent variable and two or more independent variables. For example, when we study the relationship between the age in years and both the weight and height of a group of persons, it is a problem of *multiple* correlation.

On the other hand, in partial correlation, we measure the correlation between a dependent variable and one particular independent variable when all other variables involved are kept constant, i.e., when the effects of all other variables are removed (often indicated by the phrase "other things being equal"). For example, in the above age and height problem, if we limit our correlation analysis of age and height to girls only, it becomes a problem of *partial* correlation. In this unit we shall concentrate on simple correlation only.

(iii) Linear and non-linear (curvilinear) correlation: In the above type of correlation, the distinction was based upon the number of variables studied. Here the distinction is based upon the constancy of the ratio of change between the variables. Where the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then the correlation is said to be *linear*. Let us illustrate this point. Suppose we have two series X and Y as follows:

X	Y
5	40
10	80
15	120
20	16
25	200
30	210

In both the variables the ratio of change is the same. A straight line would be obtained if all the points are plotted on a graph paper.

On the other hand, if the two variables are such that the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable, then the correlation would be *non-linear* or *curvilinear* and we would not get a straight line on a, graph paper. For example, if the age of a person is doubled over a period of time, the height or weight would not necessarily be doubled. In social sciences, generally, we find non-linear relationship between the variables. However, since techniques of analysis for measuring non-linear correlation are far more complicated than those for linear correlation, we generally make an assumption that the relationship between the variables is of the linear type.

The difference between linear and a non- linear correlation is best illustrated by scatter diagrams:

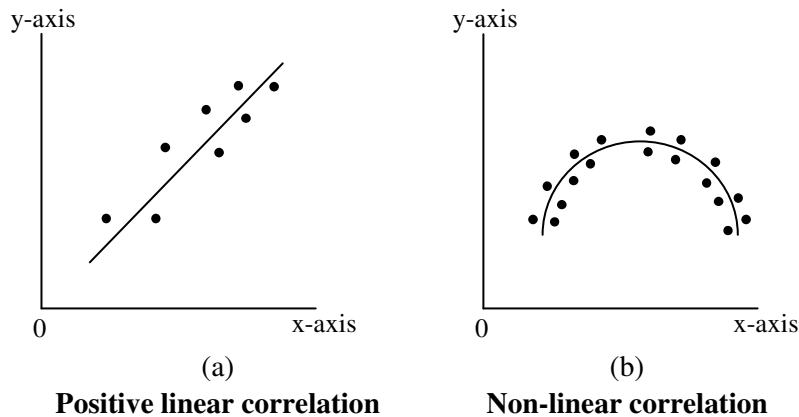


Figure: 6.1

The following points may be remembered in interpreting the scatter diagram regarding the correlation between the two variables:

- (i) If the points are very dense *i.e.*, very close to each other, a fairly good amount of correlation may be expected between the two variables. On the other hand, if the points are widely scattered, a poor correlation may be expected between them.
- (ii) If the points on the scatter diagram reveal any trend (either upward or downward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.
- (iii) If there is an upward trend rising from lower left-hand corner and going upward to the upper right-hand corner, the correlation is positive since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from the upper left-hand corner to the lower right-hand corner, the correlation is negative since in this case the values of the two variables move in the opposite directions.
- (iv) In particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points lie on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative.

Self-Assessment Questions:

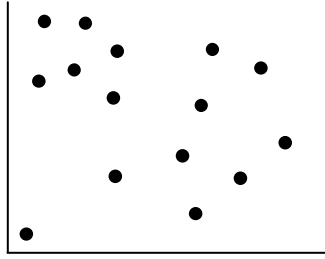
Short Questions

1. What is correlation analysis?
2. State the differences between Correlation and Causation.
3. Mention the nature of relationships that may exist between variables.
4. Define correlation. Explain various types of correlation with suitable examples.
5. What are the utilities of correlation analysis?
6. Define direct and inverse relationships.
7. State the nature of the following correlations (positive, negative or no correlation):
 - (i) Sale of woollen garments and the day temperature;
 - (ii) The color of the saree and the intelligence of the lady who wears it; and
 - (iii) Amount of rainfall and yield of crop.
8. Define correlation. Discuss its significance. Does correlation always signify causal relationship between two variables? Explain with illustration.
9. Does the high degree of correlation between the two variables signify the existence of cause-and-effect relationship between the two variables?
10. Does correlation imply causation between two variables? Explain.
11. What is '*spurious correlation*' and '*non-sense or chance correlation*'? Explain with the help of an example.
12. Comment on the following statement: "A high degree of positive correlation between the 'size of the shoe' and the 'intelligence' of a group of individuals implies that people with bigger shoe size are more intelligent than the people with lower shoe size".

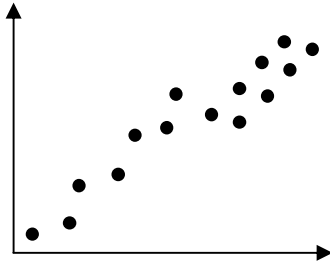
1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Who is credited with the development of the concept of correlation?
 - a) Karl Pearson
 - b) Sir Ronald A. Fisher
 - c) Francis Galton
 - d) John Tukey
- (ii) Which of the following best describes a perfect negative correlation?
 - a) -1
 - b) 0
 - c) +1
 - d) 0.5
- (vii) Considering the dimensions of a human brain, x , and the corresponding result on an IQ test, y , would you anticipate a positive association, a negative correlation, or an absence of correlation?
 - a) No correlation
 - b) Positive correlation
 - c) Inverse correlation
 - d) Negative correlation

- (ix) Determine which nature of association is shown in the following diagram



- (a) Little or no association
 (b) Moderately strong association
 (c) Positive association, linear association
 (d) Negative association, linear association
- (xv) Determine which nature of association is shown in the following diagram



- (a) Positive association, linear association, very strong association
 (b) Positive association, moderately strong association
 (c) Positive association, linear association
 (d) Linear association, very strong association
2. Write “T” if the statement is true and “F” if the statement is false:
- (i) If two variables are related in a positive linear manner, the scatter plot will show points on the X, Y space that are generally moving from the lower left to the upper right.
 (ii) If two variables are highly correlated, it is not only means that they are linearly related, it also means that a change in one variable will cause a change in the other variable.
 (iii) A correlation coefficient of 0 means that there is a perfect relationship between two variables.
 (iv) The correlation coefficient ranges from -1 to +1, with 0 indicating no relationship between variables.
 (v) If two variables have a correlation coefficient equal to +0.75, the scatter plot will have an upward slope moving from left to right.

Answer:

Multiple-Choice Question:

1. (i) c (ii) a (iii) a (iv) a (v) a

True/False

2. (i) F (ii) T (iii) F (iv) T (v) T

Lesson 2: Methods of Studying Correlation

Lesson Objectives:

After studying this lesson, you will be able to:

- Understand the methods of correlation analysis;
- Effects of change of scales and origin in correlation analysis;
- Methods of calculating coefficient of correlation of grouped and ungrouped data;
- Define and interpret correlation coefficient (r);
- Understand coefficient of correlation and probable errors;
- Define and interpret co-efficient of determination (R^2);
- Understand and explain the properties of r .

Introduction

As stated earlier in correlation analysis, we employ a symmetric methodology, treating independent and dependent variables equivalently. The correlation between two variables quantifies their linear relationship. The correlation indicates the degree to which the two variables move in a linear relationship. The correlation between X and Y is same to the correlation between Y and X . We shall now articulate a more explicit definition of correlation: The correlation between two random variables X and Y quantifies the extent of their linear relationship. A variety of correlation metrics exist, with the choice primarily determined by the data level under analysis. Researchers aim to determine ρ , the population correlation coefficient. This section introduces the often-utilized sample coefficient of correlation, r , as researchers predominantly work with sample data. This metric is usable alone if both evaluated variables possess a minimum interval level of data.

In correlation analysis, we employ a symmetric methodology, treating independent and dependent variables equivalently.

Methods of Studying Correlation

There are two steps in correlation analysis:

- (i) To visualize the relationship and
- (ii) To measure the extent of relationship.

In the first part of this lesson, we shall discuss Scatter Diagram and Graphic methods which help us in visualizing the relationship between two variables. They are based on the knowledge of diagrams and graphs.

Following are other methods which are known as mathematical methods:

1. Karl Pearson's Coefficient of Correlation.
2. Rank Method
3. Concurrent Deviation Method.
4. Method of Least Squares.

Scatter Diagram or Correlation Chart

The graphical representation of two variables may establish the fact that correlation between two variables exists. For each unit of observation in correlation analysis, there is a pair of figures. The two sets of figures are known as subject and relative. The more important set which is being

used as the standard is known as "Subject" and the one of less importance, the "Relative". The values of 'subject' are plotted along the X-axis and the corresponding value of the 'relative' along the Y-axis. When every dot representing a pair of figures has been plotted, we get what is called. a scatter diagram. The term scatter refers to the dispersion of the dots on the graph.

Positive and Negative Correlation

The dots in a scatter may fall on a line or they may be scattered over the graph without any system or they may form a path. ' The way the dots scatter gives an indication of the kind of relationship which exists between two variables. Let us now take some examples of scatter diagrams.

The points plotted form a path and cluster along a straight line. Since the movements of the variables are in the same direction, the correlation is positive.

The less perfect the relationship between the two sets of data, the greater will be the departures from the line of the best fit. These departures are known as "Scatters".

In Figure 6.2, weights of students are plotted against their heights. It is seen that in most cases, but not in all, the more is the height, the more is the weight. It will also be noticed that for two students of the same height, there may be different weights. However, there is a distinct tendency for the heights and weights of the students to move in the same direction. The points plotted form a path and cluster along a straight line. Since the movements of the variables are in the same direction, the correlation is positive.

When the points form a path then it is possible to draw a line by inspection. A good method is to stretch a piece of thread and try out the best position for the line. This line is the line of "best fit". When, however, the points are too scattered to be judged, recourse must be had to mathematical methods e.g., to fit a straight line ($Y = a + bx$) to the data. The less perfect the relationship between the two sets of data, the greater will be the departures from the line of the best fit. These departures are known as "Scatters".



Fig. 6.2 Positive Correlation
(An increase in Y is associated with of increase in X)

If higher values of one variable are associated with higher, values of the other or if low values of one variable go with low values of the

other, then we have a path that runs roughly from the lower left, corner to the upper right corner (Fig.6.2). This type of relationship is direct and is termed positive correlation. On the other hand, if high values of one variable are associated with low values of the other (i.e. when the movements of the two variables are in opposite direction), the path of dots runs roughly from the upper left corner to the lower right corner (Fig. 6.3). For instance, there may be a negative correlation between price of and demand for a commodity.

If a scatter diagram is drawn and no path is formed, i.e., all the pints are scattered over without any system, then there is no association between the variables (Fig. 6...4).

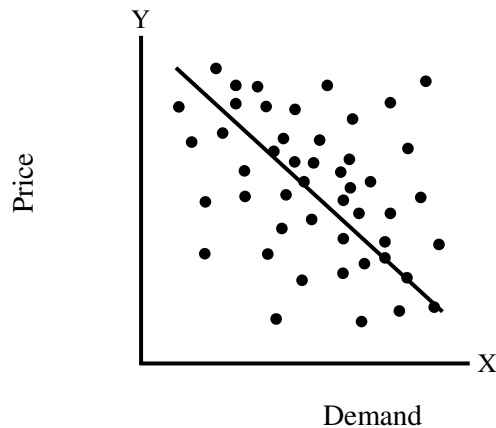


Fig. 6.3 Negative Correlation
(A decrease in Y is associated with of increase in X)

For instance, there is no relationship between the weight of a student and his marks in statistic. The two sets of figures do not appear to move in any direction. This complete dis-association - where one variable does not help us to establish the value of the other - is termed as "no" or "zero" correlation.

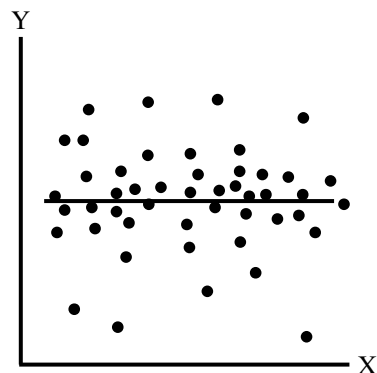


Fig. 6.4 Zero Correlation

On the other hand, two variables may be so closely associated with each other that one is inclined to think that one phenomenon is the function of the other. In such a situation, the relationship between the two variables is perfect and therefore for every given value on the X-axis, there would

All the pints are scattered over without any system, then there is no association between the variables.

When one variable does not help us to establish the value of the other is termed as "no" or "zero" correlation.

always be indicated a certain value on the Y-axis. All the points would coincide with a curve or line instead of forming a path across the face of the scatter diagram. It is to be noted that such a situation is never found in practice. This is known as perfect correlation which might be either positive or negative (Figures 6.5 and 6.6). Examples of perfect positive correlation are: (i) the circumference of a circle increases in a perfectly definite ratio with an increase in the length of its diameter, (ii) the amount of electricity bill increase in a perfectly definite ratio with an increase in the number of units consumed.

Example of a perfect Negative Correlation: In the case of gases obeying Bowley's law the volume

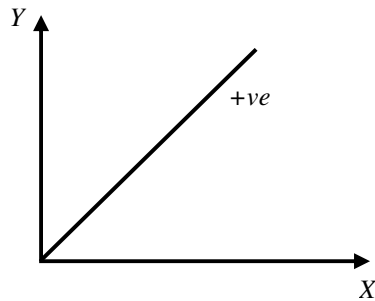


Figure 6.5

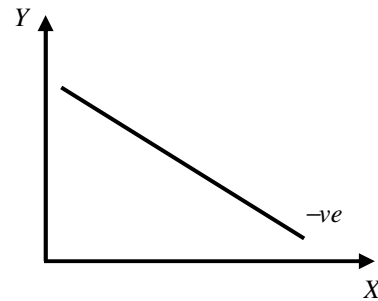


Figure 6.6

Perfect Functional Relationship

Varies inversely with the pressure (at constant temperature). If the pressure be doubled the volume will be halved.

The instances of perfect correlation are to be found in physical sciences and not in business and social sciences. In business, correlation between two variable is never perfect, but a definite value of Y will result when a given value of X is selected. On an average, an increase in one is accompanied by a corresponding increase (or decrease) in the other. Thus correlation between two variables may not be perfect but exists only to a limited extent, i.e., a change in one variable brings about a change in the other variable, but the change in the latter bears no definite ratio to the change in the former.

Example 6.1 Given the following pairs of values of the variable X and Y:

X	2	3	5	6	8	9
Y	6	5	7	8	12	11

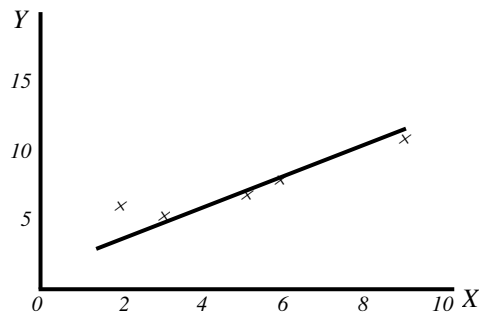
Required:

- Make a scatter diagram.
- Do you think that there is any correlation between the variables X and Y? Is it positive or negative? Is it high or low?
- By graphic inspection draw an estimating line.

Solution:

The pairs of values of the variables X and Y are plotted on the graph (Fig. 6.7). There is correlation between the two variables, and it is positive. The correlation is high as the points are not far off from the estimating line, which is drawn by graphic inspection.

It will be observed that the scatter diagram shows only the type of correlation between the two variables. To some extent, the degree of correlation may also be guessed from it. But the exact degree of correlation cannot be obtained from it. Subsequently in this lesson, we shall discuss the coefficient of correlation which is the measurement of the degree of correlation between two variables.

**Figure 6.7****Merits and Limitations of the Method:****Merits:**

1. This method is very easy and simple as it is non-mathematical method of studying correlation between the variables. To some extent, the degree of correlation may also be guessed from it.
2. It is not influenced by the size of extreme items. But we shall see later that most of the mathematical methods of finding correlation lack this quality.
3. Drawing a scatter diagram usually is the first step in investigating the relationship between two variables.

Scatter diagram method is very easy and simple as it is non-mathematical method of studying correlation between the variables.

Limitations:

The scatter diagram only shows the type of correlation between the two variables. To some extent, degree of correlation may also be guessed from it. But the exact degree of correlation cannot be obtained from it. As pointed out earlier we shall discuss later the coefficient of correlation which is the measurement of the degree of correlation between two variables.

Graphic Method

It is the simplest method of ascertaining the presence of correlation between the two variables. The values of the two variables are plotted on a graph paper, keeping in view the values of the X-axis and Y-axis as the same for both. The two curves thus obtained form the basis of comparison. if both of them move in one direction, there is correlation; if the curves do not move in the same direction; there is no correlation.

The two curves thus obtained form the basis of comparison. if both of them move in one direction, there is correlation; if the curves do not move in the same direction; there is no correlation.

Example 6.2 From the following data ascertain whether there is correlation between exports of raw cotton (X) and import of manufactured cotton goods (Y).

Year	I	II	III	IV	V	VII	VIII
X	42	44	58	55	89	90	66
Y	46	49	53	58	65	76	58

[in crores of Taka]

Solution:

The graph shows that the variables, X and Y are quite closely related. In the graph, the dotted line shows the X variable and solid line shows the Y variable.

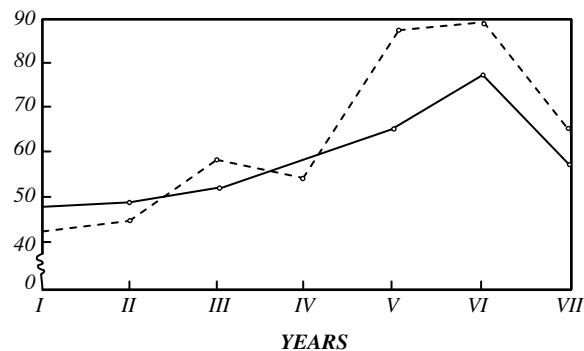


Figure: 6.8

Merits and Limitations of the Graphic Method

Like a scatter diagram, it is possible to ascertain the nature and extent of correlation from a graph. But the exact degree of correlation cannot be ascertained either from the scatter diagram or the graphic method. This method is generally used where we are given data over a period of time and where the exact degree of correlation is not required. In other cases, in general, other method, i.e., Mathematical methods are adopted which describe the correlation by a numerical value.

Mathematical Methods

The following four methods are generally discussed for studying correlation mathematically:

1. Karl Pearson's Coefficient of Correlation.
2. Rank Correlation.
3. Concurrent Deviation Method.
4. Method of Least Squares.

Acknowledging the breadth of the discussion, only the Karl Pearson's Coefficient of Correlation will be continued in this lesson two. Thereafter in the lesson number three Rank correlation and in the lesson number Four Concurrent deviation method and Least Square method will be discussed respectively.

The Karl Pearson Correlation Coefficient

Of the above methods of measuring correlation numerically the Karl Pearson's method, popularly known as correlation coefficient, is most widely used. It is denoted by the symbol r .

Symbolically,

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

Where;

$$x = (X - \bar{X})$$

$$y = (Y - \bar{Y})$$

σ_x = Standard deviation of series X

σ_y = Standard deviation of series Y

N = Number of paired observations.

It may be noted that this method is to be applied only where the deviations of items are taken from actual means and not from assumed means.

Example 6.3. Calculate Karl Pearson's Correlation Coefficient if the following paired data:

X=	28	37	40	38	35	33	40	32	34	33
Y=	23	32	33	34	30	26	29	31	34	38

Required:

What inference would you draw from, the estimate?

Solution:

$$\bar{X} = \frac{\sum X}{N} = \frac{350}{10} = 35$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{310}{10} = 31$$

Table 6.2: Computation of Coefficient of Correlation

X	x	x^2	Y	y	y^2	xy
28	-7	49	23	-8	64	+56
37	+2	4	32	+1	1	+2
40	+5	25	33	+2	4	+10
38	+3	9	34	+3	9	+9
35	0	0	30	-1	1	0
33	-2	4	26	-5	25	+10
40	+5	25	29	-2	4	-10
32	-3	9	31	0	0	0
34	-1	1	34	+3	9	-3
33	-2	4	38	+7	49	-14
$\sum X = 350$		$\sum x^2 = 130$	$\sum Y = 310$		$\sum y^2 = 166$	$\sum xy = +60$

$$\sigma_x = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{130}{10}} = 3.6$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{166}{10}} = 4.1$$

Therefore:

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

$$r = \frac{+60}{10(3.6)(4.1)} = \frac{+60}{147.6} = +0.407$$

The degree of correlation varies between +1 and -1.

The advantage of Correlation Coefficient is that it describes not only the magnitude of correlation but also its direction.

The Degree of Correlation

The degree of correlation varies between +1 and -1. If there is perfect correlation, the degree of correlation will be +1, as in the case of positive correlation and -1 in the case of negative correlation. In business and social science, it is most likely to be less than 1 and therefore it may vary between -1 and +1. When there is no relationship between two variables, then it is known as zero correlation. However, in practice such values of r as +1, -1 and 0 are rare. For instance, in the above example, the answer was +0.407. It shows that the extent of correlation is 0.407 and its direction is positive. Thus, the advantage of Correlation Coefficient is that it describes not only the magnitude of correlation but also its direction.

Limitations of the Method

It may be noted here that it is a lengthy process because the true means of both the series are first to be ascertained and all the deviations are to be found out. The original values of the standard deviations, too, are to be known. Only then the final formula computing the correlation coefficient may be used. Thus it may be noted that this method is to be applied only where the deviations of items are taken from assumed means.

The above formula for computing Pearsonian coefficient of correlation can be transformed into another form which is much easier to apply. This is given below.

Product Moment Formula

The original formula that Pearson developed is commonly known as the Product Moment Method.

Symbolically,

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

This formula can easily be transformed from the first formula applied in the above example:

i.e.,

$$r = \frac{\sum xy}{N\sigma_x\sigma_y}$$

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

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

The product moment formula is much easier to apply as it separately. Substituting the values of Example 6.3 in this formula, we get;

$$r = \frac{+60}{\sqrt{(130 \times 166)}} = \frac{+60}{\sqrt{21580}} = \frac{+60}{147} = +0.408 \text{ appx.}$$

$$x = (X - \bar{X})$$

$$y = (Y - \bar{Y})$$

Procedures: The computation of correlation coefficient may be summarized as under:

- Take the deviations of X series from $\bar{X}$ and denote the deviations by x.
- Square these deviations and find out $\sum x^2$. Similarly find out $\sum y^2$.
- Multiply the deviations x and y and obtain $\sum xy$.
- Substitute the values of $\sum xy$, $\sum x^2$ and $\sum y^2$ in the product moment formula.

Example 6.4. Making use of the data summarized below calculate the coefficient of correlation, r_{12} .

Case	X_1	X_2	Case	X_1	X_2
A	10	9	E	12	11
B	6	4	F	13	13
C	9	6	G	11	8
D	10	9	H	9	4

Solution:

Calculation of Coefficient of Correlation.

Case	X_1	$(X_1 - \bar{X}_1) = x_1$	x_1^2	X_2	$(X_2 - \bar{X}_2) = x_2$	x_2^2	x_1x_2
A	10	0	0	9	1	1	0
B	6	-4	16	4	-4	16	16
C	9	-1	1	6	-2	4	2
D	10	0	0	9	+1	1	0
E	12	2	4	11	+3	9	6
F	13	3	9	13	+5	25	15
G	11	1	1	8	0	0	0
H	9	-1	1	4	-4	16	4
	80	0	32	64	0	72	43

$$x_1 = \frac{\sum X_1}{N} = \frac{80}{8} = 10 \text{ and } x_2 = \frac{\sum X_2}{N} = \frac{64}{8} = 8$$

$$r_{12} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2 \times \sum x_2^2}} = \frac{43}{\sqrt{32 \times 72}} = \frac{43}{\sqrt{2304}} = \frac{43}{48}$$

$$= 0.896$$

Example 6.5. From the following data compute the coefficient of correlation between X and Y.

	X series	Y series
Number of items	15	15
Arithmetic mean	25	18
Squares of deviations from mean	136	138

Summation of product deviations of X and Y series from their respective Arithmetic Means = 122.

Solution:

Denoting deviations of X and Y from their Arithmetic Means by x and y respectively, the given data are:

$$\sum x^2 = 136, \quad \sum xy = 122, \quad \sum y^2 = 138$$

Applying the Product Moment Formula, the coefficient of correlation is given by,

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} = \frac{122}{\sqrt{136 \times 138}} = \frac{122}{137} = +0.89$$

Example 6.6. A computer while calculating the Coefficient of Correlation between two variates X and Y from 25 pairs of observations obtained the following.

$$n=25, \sum dx=125, \sum dx^2 = 650, \sum dy=100, \sum y^2=960 \text{ and } \sum dxdy=508.$$

It was, however, later discovered, at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Required:

Obtain the correct value of the Coefficient of Correlation.

Solution:

As while copying the problem, some mistakes were committed, it is necessary to make adjustments in two values before calculating the coefficient of correlation. They are:

Σdy^2 and $\Sigma dx.dy$. The values of n , Σdx , Σdy and Σdx^2 remain unchanged.

dy^2				$dx dy$					
<i>Original</i>		<i>New</i>		<i>Original</i>			<i>New</i>		
14	196	12	144	6	14	84	8	12	96
6	36	8	64	8	6	48	6	8	48
232		208		132			144		

The corrected value of $\Sigma dy^2 = 960 - 232 + 208 = 936$ and the corrected value of $\Sigma dx dy = 508 - 132 + 144 = 520$

$\therefore$ Therefore; Correct value of $r = \frac{520}{\sqrt{650 \times 936}} = +.66$ approx.

Change of Scale and Origin

Computations of correlation coefficient can be simplified by dividing the given data by a common factor. This procedure was also adopted while computing arithmetic mean and standard deviation. But the ultimate result is not multiplied by the common factor as is done in the case of arithmetic mean and standard deviation because correlation coefficient is independent of change of scale and origin.

Example 6.7. Calculate coefficient of correlation from the following data:

X	50	100	150	200	250	300	350
Y	10	20	30	40	50	60	70

Solution:

X	$X - \bar{X}$	$\frac{X - \bar{X}}{50} = x$	x^2	Y	$Y - \bar{Y}$	$\frac{Y - \bar{Y}}{10} = y$	y^2	xy
50	-150	-3	9	10	-30	-3	9	9
100	-100	-2	4	20	-20	-2	4	4
150	-50	-1	1	30	-10	-1	1	1
200	0	0	0	40	0	0	0	0
250	+50	+1	1	50	+10	+1	1	1
300	+100	+2	4	60	+20	+2	4	4
350	+150	+3	9	70	+30	+3	9	9
ΣX =1400		$\Sigma x = 0$	Σx^2 =28	ΣY =280		$\Sigma y = 0$	Σy^2 =28	Σxy =28

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$r = \frac{28}{\sqrt{28 \times 28}} = \frac{28}{28} = 1$$

Calculation of r without taking deviations:

Where values of X and Y are small, r can be calculated without taking deviations from actual or assumed mean.

Example 6.8. Find out the coefficient of correlation between X and Y.

X	2	2	4	5	5
Y	6	3	2	6	4

Solution

X	X ²	Y	Y ²	XY
2	4	6	36	12
2	4	3	9	6
4	16	2	4	8
5	25	6	36	30
5	25	4	16	20
18	74	21	101	76

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

$$r = \frac{5(76) - (18)(21)}{\sqrt{[5(74) - (18)^2][5(101) - (21)^2]}}$$

$$r = \frac{380 - 378}{\sqrt{(370 - 324)(505 - 441)}}$$

$$r = \frac{2}{\sqrt{(64)(64)}} = \frac{2}{64} = 0.04$$

When Deviations are taken from Assumed Mean**(Karl Pearson's Short cut Method)**

In our preceding Example 6.3, means were whole numbers, (ie., $\bar{X} = 35$) and ($\bar{Y} = 31$). In practice actual means are in fractions. To avoid difficult calculations, deviations are taken from assumed means. Also the formula is modified as regards Standard Deviation because deviations are taken from assumed mean. Thus when deviations 'are taken from an assumed mean the following formula is applicable:

$$\frac{\sum d_x d_y - \frac{\sum d_x \times \sum d_y}{N}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{N}} \times \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{N}}}$$

Where;

d_x = refers to deviations of X series from an assumed mean, i.e., $(X-A)$

d_y = refers to deviations of Y series from an assumed mean: i.e., $(Y-A)$.

$\Sigma d_x d_y$ = denotes the sum of the product of the deviations of X and Y series from their assumed means.

Σd_x^2 = denotes the sum of the squares of the deviations of X series from an assumed mean.

Σd_y^2 = denotes the sum of the squares of the deviations of Y series from an assumed mean.

It may be noted that there are many variations of the above formula. The above formula may also be written as:

$$r = \frac{\Sigma d_x d_y - \frac{\Sigma d_x}{N} \times \frac{\Sigma d_y}{N}}{\sqrt{(\frac{\Sigma d_x^2}{N} - (\frac{\Sigma d_x}{N})^2) \times (\frac{\Sigma d_y^2}{N} - (\frac{\Sigma d_y}{N})^2)}}$$

or

$$r = \frac{\Sigma d_x d_y \times N - \Sigma d_x \times \Sigma d_y}{\sqrt{\Sigma d_x^2 \times N - (\Sigma d_x)^2 \times (\Sigma d_y^2 \times N - (\Sigma d_y)^2)}}$$

But the form given earlier is easiest to apply.

Procedures: In the computation of correlation coefficient by the above method are:

- (i) Find out the deviations of X series from an assumed mean and denote these deviations by d_x and find out Σd_x . Similarly in the case of Y series find out Σd_y .
- (ii) Square d_x and obtain the total Σd_x^2 . Similarly find out Σd_y^2
- (iii) Multiply d_x with d_y and obtain the total $\Sigma d_x d_y$
- (iv) Substitute the values found out in the formula for coefficient of correlation.

Example 6.9. Calculate the coefficient of correlation from the following data:

Marks in Statistics:	20	30	28	17	19	23	35	13	16	38
Marks in Accounting:	18	35	20	18	25	28	33	18	20	40

Solution: According to the formula of coefficient correlation:

$$r = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \cdot \Sigma d_y}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N} \times \Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

Table 6.3: Computation of Coefficient of Correlation between Marks in Statistics and Accounting

Marks in Statistics X	$X - 30 = d_x$	d_x^2	Marks in Law Y	$X - 30 = d_y$	d_y^2	$d_x d_y$
20	-10	100	18	-12	144	120
30	0	0	35	+5	25	0
28	-2	4	20	-10	100	20
17	13	169	18	-12	144	156
19	-11	121	25	-5	25	55
23	-7	49	28	-2	4	14
35	+5	25	33	+3	9	15
13	-17	289	18	-12	144	204
16	-14	196	20	-10	100	140
38	+8	64	40	+10	100	80
	$\Sigma d_x = -61$	$\Sigma d_x^2 = 1017$		$\Sigma d_y = -45$	$\Sigma d_y^2 = 795$	$\Sigma d_x d_y = 804$

Substituting the values in the formula for r, we get

$$r = \frac{804 - \frac{(-61) \times (-45)}{10}}{\sqrt{\left[1017 - \frac{(-61)^2}{10}\right] \times \left[795 - \frac{(-45)^2}{10}\right]}}$$

$$r = \frac{804 - 3355}{\sqrt{[1017 - 3721] \times [7950 - 3025]}}$$

$$r = \frac{4685}{\sqrt{[6449] \times [4925]}} = \frac{4685}{5635.8} = +0.85$$

Example: 6.10. The following table gives the distribution of the total population and those who are wholly or partially blind among them.

Age	No. of Persons ('000)	Blind
0-10	100	55
10-20	60	40
20-30	40	40
30-40	36	40
40-50	24	36
50-60	11	22
60-70	6	18
70-80	3	15

Required:

Find out if there is any relation between age and blindness.

Solution: In this case, blinds per lakh are to be calculated as follows: For age group 0-10 it will be $\frac{55}{1000} \times 100000 = 55$; for 10-20, it will be $\frac{40}{60,000} \times 10,000 = 67$ and so on for other age groups.

Solution:

Age groups	Age			Blindness			Product of deviations $d_x d_y$
	Mid. Value X	$d_x = \frac{X-45}{10}$	d_x^2	Blind per lakh	$d_y = Y-185$	d_y^2	
0-10	5	-4	16	55	-130	16900	520
10-20	15	-3	9	67	-118	13924	354
20-30	25	-2	4	100	-85	7225	170
30-40	35	-1	1	111	-74	5476	74
40-50	45	0	0	150	-35	1225	0
50-60	55	+1	1	200	-115	13225	230
60-70	65	+2	4	300	15	225	4500
70-80	75	+	9	500	315	99225	157500
N=8		-4	44		+3	157425	2308

It may be noted that frequency of blinds has also been convert into lakhs so as to facilitate comparison with Age.

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \cdot \sum d_y}{N}}{\sqrt{\left[\sum d_x^2 - \frac{(\sum d_x)^2}{N} \right] \left[\sum d_y^2 - \frac{(\sum d_y)^2}{N} \right]}}$$

Substituting the values in the above formula by Karl Pearson we get:

$$r = \frac{2308 - \frac{(-4)(+3)}{8}}{\sqrt{\left[44 - \frac{16}{1} \right] \left[157425 - \frac{916}{8} \right]}}$$

$$r = \frac{2308 + 1.5}{\sqrt{[42][157425]}} = \frac{2309.5}{\sqrt{6611841}}$$

$$= \frac{2309.5}{2571} = +0.898$$

Example 6.11. Consider the following information:

	X series	Y series
Number of items	10	10
Total of the Deviations	-170	-20
Total of the square of Deviations	8288	2264

Total of the multiplication of deviations of X and Y series = 3044.

Required:

Find out the coefficient of correlation when the assumed means, of X series and Y series are 82 and 68 respectively.

Solution: Denoting deviations of X and Y from assumed means by d_x and d_y , the given data are

$$\Sigma d_x d_y = 3044; \Sigma d_x = -170; \Sigma d_y = -20$$

$$N = 10; d_x^2 = 8288; d_y^2 = 2264.$$

As the deviations are from assumed means, therefore the formula to be applied would be:

$$r = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \cdot \Sigma d_y}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} \times \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$$r = \frac{3044 - \frac{(-170)(-20)}{10}}{\sqrt{8288 - \frac{(-170)^2}{10}} \times \sqrt{2264 - \frac{(-20)^2}{10}}}$$

$$= \frac{3044 - 340}{\sqrt{8288 - 2890} \times \sqrt{2264 - 40}} = \frac{2704}{\sqrt{5398} \sqrt{2224}}$$

$$= \frac{2704}{3464.8} = +0.78$$

Another form of the Formula where deviations are taken from assumed mean

There is another form of the formula for correlation coefficient where deviations are taken from assumed mean. It is

$$r = \frac{\Sigma xy - N(\bar{X} - A_x)(\bar{Y} - A_y)}{N\sigma_x\sigma_y}$$

where; A_x and A_y denote assumed means of X and Y series respectively.

The following example shall illustrate the application of the above formula.

Example 6.12. Given the following information:

Number of pairs of observations X and Y series = 8

X Series Arithmetic Average = 74.5

X Series Assumed Average = 69.0

X Series Standard Deviation = 13.07

Y Series Arithmetic Average = 125.5

Y Series Assumed Average = 112

Y Series Standard Deviation = 15.85

Summation of products of corresponding deviations of X and series = 2176.

Required:

Calculate the Pearson's coefficient of correlation.

Solution:

In this example, we are given both the actual and assumed means of X and Y series. In such a case, it is appropriate to apply this formula:

$$r = \frac{\sum xy - N(\bar{X} - A_x)(\bar{Y} - A_y)}{N\sigma_x\sigma_y}$$

where A_x and A_y denote assumed means of X and Y series respectively.

Substituting the values in the above formula, we get

$$\begin{aligned} r &= \frac{2176 - 8(74.5 - 69)(125.5 - 112)}{8(13.07)(15.85)} \\ &= \frac{2176 - 594}{1657.276} = \frac{1572}{1657.276} = +0.955. \end{aligned}$$

Coefficient of Correlation of Grouped Data

Generally, the data of the continuous series are classified in a two-way frequency table. Such tables are termed as bivariate or 'contingency' or 'Correlation' tables. The class intervals for Y are listed in the column headings and those for X are listed in the stubs at the left of the table. However, the order can also be reversed.

The formula for calculating the coefficient of correlation is

$$r = \frac{\sum fd_x d_y - \frac{\sum fd_x \cdot \sum fd_y}{N}}{\sqrt{\sum fd_x^2 - \frac{(\sum fd_x)^2}{N}} \times \sqrt{\sum fd_y^2 - \frac{(\sum fd_y)^2}{N}}}$$

This formula is the same as the one discussed above for assumed mean. But as in the grouped data frequencies are involved, the formula has been modified accordingly.

There are some variations of the above formula.

$$r = \frac{\frac{\sum [fd_x d_y]}{N} - \frac{\sum fd_x}{N} \cdot \frac{\sum fd_y}{N}}{\sqrt{\frac{\sum (fd_x^2)}{N} - \left(\frac{\sum fd_x}{N}\right)^2} \times \sqrt{\frac{\sum (fd_y^2)}{N} - \left(\frac{\sum fd_y}{N}\right)^2}}$$

or $r = \frac{\sum (fd_x d_y) \cdot N - \sum fd_x \cdot \sum fd_y}{\sqrt{\sum (fd_x^2) \cdot N - (\sum fd_x)^2} \times \sqrt{\sum (fd_y^2) \cdot N - (\sum fd_y)^2}}$

But the formula given earlier is invariably used.

This formula is explained with the help of the following example:

Example 6.13. The following table gives the ages of husbands and wives at the time of their marriages.

Ages of Wives

		10-20	20-30	30-40	40-50
Ages of husbands	10-20	20	26	-	-
	20-30	8	14	37	-
	30-40	-	4	18	3
	40-50	-	-	4	6

Required:

Calculate the correlation coefficient between the ages of husbands and wife.

Solution: It will be observed from Table 6.4, that all the values necessary for calculation of coefficient of correlation have been computed therein. The box →

d_x/d_y

Signifies the step deviation. The sum of $fd_x d_y$ is obtained by:

- (1) securing the cross product of the indicated value of d_x and d_y for each box.
- (2) multiplying the values obtained by the frequency contained in the box (the result is entered in upper left-hand corner of the box).
- (3) adding up for all boxes in the table on the next page.

$$r = \frac{76 - \frac{(49)(-1)}{140}}{\sqrt{123 - \frac{(49)^2}{140}} \sqrt{111 - \frac{(-1)^2}{140}}}$$

$$= \frac{76 + 0.35}{\sqrt{123 - 17.15} \sqrt{111 - 0.0007}}$$

$$= \frac{76.35}{\sqrt{105.85} \sqrt{110.993}} = \frac{76.35}{10.3 \times 10.54}$$

$$= \frac{76.35}{108.56} = +0.703$$

TABLE 6.4.**Ages of Wives**

			X	10–20	20–30	30–40	40–50				
			Mid-Point	15	25	35	45				
			Deviation	–10	0	+10	+20				
Y	M.P.	Deviation	$\frac{dx}{dy}$	–1	0	+1	+2	f	f dy	f dy ²	f dx dy
10–20	15	–10	–1	$\frac{20}{20}$	$\frac{0}{26}$			46	–46	46	20
20–30	25	0	0	$\frac{0}{8}$	$\frac{0}{14}$	$\frac{0}{37}$		59	0	0	0
30–40	35	+10	+1		$\frac{0}{4}$	$\frac{18}{18}$	$\frac{6}{3}$	25	25	25	24
40–50	45	+20	2			$\frac{8}{4}$	$\frac{24}{6}$	10	20	40	32
			f	28	44	59	9	140	–1	111	76
			f dx	–28	0	59	18	49			
			f dx ²	28	0	59	36	123			
			f dx dy	20	0	26	30	76			

Example 6.14 The following table gives the frequency according to age group of marks obtained by 67 students in an intelligence test.

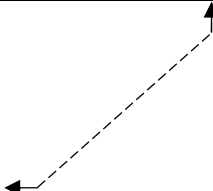
Test Marks	Age in years				
	18	19	20	21	Total
200-250	4	4	2	1	11
250-300	3	5	4	2	14
300-350	2	6	8	5	21
350-400	1	4	6	10	21
Total	10	19	20	18	67

Required:

Is there any relationship between age an intelligence?

Solution:

Denoting age by X and test marks by Y, the calculations are done in the following table:

Y \ X		X				f	fdy	fdy ²	fdxdy
			18	19	20	21			
			dx						
			dy						
200–250	225	–1	$\frac{4}{4}$	$\frac{0}{4}$	$\frac{-1}{2}$	$\frac{-2}{1}$	11	–11	11
250–300	275	0	$\frac{0}{3}$	$\frac{0}{5}$	$\frac{0}{4}$	$\frac{0}{2}$	14	0	0
300–350	325	1	$\frac{-2}{2}$	$\frac{0}{6}$	$\frac{8}{8}$	$\frac{10}{5}$	21	21	21
350–400	375	2	$\frac{-2}{1}$	$\frac{0}{4}$	$\frac{12}{6}$	$\frac{40}{10}$	21	42	84
		f	10	19	20	18	N=67	Σfdy = 52	Σfdy ² = 116
		fdx	–10	0	20	36	Σfdx = 46		
		Σfdx ²	10	0	20	72	Σfdx = 102		
		Σfdxdy	0	0	19	48	Σfdxdy = 67		

$$67 - \frac{46 \times 52}{67}$$

$$r = \frac{67 - \frac{46 \times 52}{67}}{\sqrt{102 - \frac{(46)^2}{67}} \sqrt{102 - \frac{(52)^2}{67}}}$$

$$r = \frac{67 - 35.7}{\sqrt{102 - 31.6} \sqrt{102 - 40.4}} = \frac{31.3}{\sqrt{70.4} \sqrt{75.6}}$$

$$= \frac{31.3}{72.9} = +0.429$$

Characteristics of Pearson's Coefficient

1. It varies between +1 and -1 in the case of presence of correlation. Absence of correlation is denoted by zero.
2. It takes into account all items of the variable and is thus based on a suitable measure of variation.
3. It measures both the direction as well as degree of change.
4. The coefficient of correlation might lead to fallacious conclusions in cases of accidental correlation. As mentioned earlier this is known as non-sense or spurious correlation and may be found in space as well as time.
5. As explained earlier, the coefficient of correlation does not prove causation but it is simply a measure of covariation, because a variation in two series may be due to (i) some causation of the subject to the relative, (ii) some common cause, (iii) some interacting causes or, (iv) some chance.

It varies between +1 and -1 in the case of presence of correlation. Absence of correlation is denoted by zero.

Assumptions of Pearson's Correlation: Karl Pearson's coefficient of correlation is based on the following assumptions:

1. There is linear relationship between the two variables, i.e. by plotting them on a graph paper, a straight line would be obtained.
2. A large number of independent causes are operating in both the variables correlated so as to produce a normal distribution.
3. The operating forces are not independent of each other and their relation is one of cause and effect. If such forces are independent there cannot be any correlation. For example, there is no relationship between height of a man and that of a tree because the forces that affect these variables are not common.

Merits and Limitations of the Pearson's Correlation Coefficient

The correlation coefficient summarizes in one figure not only the degree of correlation but also the direction, i.e., whether correlation is positive or negative. The value of correlation coefficient cannot exceed +1 or -1.

However, the use of this method has some limitations. They are:

1. It always assumes linear relationship between the variables regardless of the fact that assumption is correct or not.
2. It is not easy to interpret the significance of correlation coefficient. Very often it is misinterpreted.
3. As compared with some other methods it is time consuming.
4. It is affected by the extreme items.

It always assumes linear relationship between the variables regardless of the fact that assumption is correct or not.

Interpreting the Coefficient of Correlation:

The coefficient of correlation measures the degree of relationship between two or more variables. As the reliability of estimates depends upon the closeness of the relationship, it is necessary that great care must be exercised in interpreting the coefficient of correlation, otherwise fallacious conclusions may be drawn. The coefficient of correlation can be interpreted in terms of (i) degrees and (ii) significance.

As regards the degree of correlation, it is perfect positive if r is +1 and perfect negative if it is -1. If there is an absence of correlation the result is zero. If r is greater than 0.95 there is a high degree of correlation between the variables and one of them may be quite accurately estimated from a known value of the other. If r is greater than 0.75 but less than 0.85 there probably is a decided amount of association between the two series and one of the variables may be estimated roughly from a known value of the other. If r is more than 0.40 but less than 0.60 there is a fair degree of correlation but any estimate of one variable from a known value of the other would ordinarily be of but little practical value. If r is less than 0.35 the correlation is not at all marked. There is but little association present between the two series, and one of the series cannot be used as the basis for estimating the value of the other.

It may be concluded that the closer r is to + 1 or -1, the closer the relationship between the variables and the closer r is to 0, the less close the relationship. Nothing more than this can be said about the interpretation of r . The full interpretation of r depends upon circumstances one of which is the size of the sample. While estimating the value of one variable from the value of another, the higher the value of r the better the estimate. But it may be noted that the closeness of the relationship is not proportional to r . If the value of r is 0.45 it does not indicate a relationship one-half as close as that of 0.90. If r is 0.45, it is also very close and marked.

Coefficient of Correlation and Probable Error: The Probable Error of r helps in interpreting its value. As coefficient of correlation are often computed from samples, they, like other statistical quantities, are subject to errors of sampling and from the interpretation point of view Probable Error is very useful.

The Probable Error of r helps in interpreting its value.

The Probable Error of the Karl Pearson's coefficient of correlation is computed by the following formula:

$$P.E. = 0.6745 \frac{1-r^2}{\sqrt{n}}$$

It is used in interpreting whether r is significant or not. If r is less than its P.E., it is not at all significant. Perhaps there is no evidence of correlation. If r is more than P.E., there is correlation and it is significant if r is more than six times P.E.

Probable error is used in interpreting whether r is significant or not.

The P.E. is also used for testing the reliability of a particular value of r . If P.E. is added to and subtracted from the average r , then we shall have two limits within which the probability is that a coefficient of correlation from series selected at random from the same population will lie. Symbolically,

$$r \pm P.E.$$

The standard error is defined as

$$S.E. = \frac{1-r^2}{\sqrt{n}}$$

Thus, it is obvious that $P.E. = 0.6745$ of standard error of coefficient of correlation.

It may be noted that the P. E. of a correlation coefficient varies inversely with the number of pairs of items and the size of the coefficient.

Conditions for the Use of P.E.: The concept of P.E. can be properly applied only when the following three conditions exist:

- (i) The whole of the data are symmetrical and give a normal bell-shaped curve.
- (ii) The standard for which the P.E. has been computed must have been derived from a sample data, and
- (iii) The sample must have been taken out in an unbiased manner.

However, these conditions are generally not satisfied and as such the reliability of the correlation coefficient is determined largely on the basis of exterior tests of reasonableness which are often of a non-statistical character.

The following example illustrates the above formulae.

Example 6.15. Calculate the correlation coefficient from the following data. Also find out the limits within which coefficient of correlation of a sample drawn from the same universe will lie.

Number of Oars sold	Value ('000 Tk.)
21	41
18	34
23	38
34	67
36	68
38	84
38	76
36	72
32	99
33	67
32	58

Solution:

X	$x = X - \bar{X}$	x^2	Y	$y = Y - \bar{Y}$	y^2	xy
21	-10	100	41	-23	529	230
18	-13	169	34	-30	900	390
23	-8	64	38	-26	676	208
34	+3	9	67	+3	9	9
36	+5	25	68	+4	16	20
38	+7	49	84	+20	400	140
38	+7	49	76	+12	144	84
36	+5	25	72	+8	64	40
32	+1	1	99	+35	1225	35
33	+2	4	67	+3	9	6
32	+1	1	58	-6	36	-6
$\Sigma X=341$		$\Sigma x^2=496$	$\Sigma Y=704$		$\Sigma y^2=4008$	$\Sigma xy=1156$

$$\bar{X} = \frac{\sum X}{N} = \frac{341}{11} = 31 \text{ and } \bar{Y} = \frac{\sum Y}{N} = \frac{704}{11} = 64$$

$$\sigma_x = \sqrt{\frac{\sum X^2}{N}} = \sqrt{\frac{496}{11}} = 6.71$$

$$\sigma_y = \sqrt{\frac{\sum Y^2}{N}} = \sqrt{\frac{4008}{11}} = 19.09$$

$$r = \frac{\sum xy}{n\sigma_x\sigma_y}$$
$$= \frac{1156}{11(6.71)(19.09)} = +0.82 \text{ Approx}$$

As r is more than 0.5, it is significant.

$$P.E. = 0.6745 \frac{1 - r^2}{\sqrt{n}}$$

$$P.E. = 0.6745 \frac{1 - (0.82)^2}{\sqrt{11}}$$

$$P.E. = \frac{(0.6745)(0.3276)}{3.32} = 0.07 \text{ Approx.}$$

The limits of r are given by $r \pm P.E.$

$$0.82 \pm 0.07 = 0.75 \text{ and } 0.89.$$

Coefficient of Determination

Another useful way of interpreting the value of r between two variables is to use coefficient of determination which will be explained in the next unit on Regression analysis.

Properties of the Coefficient of Correlation

The following are the important properties of the coefficient of correlation:

1. It is independent of changes of scale and origin of the variables X and Y.

Proof: By change of origin means subtracting of some constant from every given value of X and Y and by change of scale means dividing (or multiplying) every given value of X and Y by some constant.

Given

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

where $\bar{X}$ and $\bar{Y}$ represent means of X and Y series respectively.

Now we shall change the scale and origin. Let the constant to be subtracted from X be 'a' and from Y be 'b'. Also divide X and Y series by a constant, i.e., 'c' and 'i'. The new values x and y obtained from X and Y after changes of scale and origin shall be

$$x = \frac{X - a}{c}$$

$$\text{and } y = \frac{Y - b}{i}$$

$$\text{mean of } x = \frac{\sum x}{n} = \frac{\sum \left(\frac{x - a}{c} \right)}{n} = \frac{\sum x - na}{nc} = \frac{\sum x - Na}{nc}$$

$$\text{But } \frac{\sum x - Na}{nc} = \frac{\bar{X} - a}{c}$$

$$\text{Therefore; mean of } x = \frac{\bar{X} - a}{c}$$

$$\text{and similarly, it can be shown that mean of } y = \frac{\bar{Y} - b}{i}$$

The value of r for the new set of values will be:

$$\begin{aligned} &= \frac{\sum \left(\frac{X - a}{c} - \frac{\bar{X} - a}{c} \right) \left(\frac{Y - b}{i} - \frac{\bar{Y} - b}{i} \right)}{\sqrt{\sum \left(\frac{X - a}{c} - \frac{\bar{X} - a}{c} \right)^2 \times \sum \left(\frac{Y - b}{i} - \frac{\bar{Y} - b}{i} \right)^2}} \\ &= \frac{\sum \left(\frac{X - a - \bar{X} + a}{c} \right) \left(\frac{Y - b - \bar{Y} + b}{i} \right)}{\sqrt{\sum \left(\frac{X - a - \bar{X} + a}{c} \right)^2 \times \sum \left(\frac{Y - b - \bar{Y} + b}{i} \right)^2}} \\ &= \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{ci}}{\sqrt{\frac{\sum (X - \bar{X})^2}{c^2} \times \frac{\sum (Y - \bar{Y})^2}{i^2}}} \\ &= \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{ci}}{\sqrt{\frac{\sum (X - \bar{X})^2 \times \sum (Y - \bar{Y})^2}{c^2 i^2}}} \\ &= \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \times \sum (Y - \bar{Y})^2}} \end{aligned}$$

which is same as r between original values of X and Y.

Thus the coefficient of correlation is independent of change of scale and origin.

2. The Value of coefficient of correlation lies between -1 and +1.

Proof: We have seen above that r is independent of change of origin. So assuming x and y as the deviations of X and Y series from their means and σ_x and σ_y as standard deviations of x and y and.

Expanding $\sum \left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y} \right)^2$ we get

$$\begin{aligned} &= \sum \left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} + \frac{2xy}{\sigma_x \sigma_y} \right] \\ &= \frac{\sum x^2}{\sigma_x^2} + \frac{\sum y^2}{\sigma_y^2} + \frac{2\sum xy}{\sigma_x \sigma_y} \end{aligned}$$

But the definition

$$(i) \sqrt{\frac{\sum x^2}{n}} = \sigma_x, \text{ i.e., } \sum x^2 = n\sigma_x^2$$

$$(ii) \sqrt{\frac{\sum y^2}{n}} = \sigma_y, \text{ i.e., } \sum y^2 = n\sigma_y^2$$

$$(iii) \frac{\sum xy}{n\sigma_x \sigma_y} = r, \text{ i.e., } \frac{\sum xy}{\sigma_x \sigma_y} = nr$$

$$\begin{aligned} \text{Hence } \sum \left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y} \right)^2 &= n + n + 2nr \\ &= 2n + 2nr \\ &= 2n(1 + r) \end{aligned}$$

But $\sum \left(\frac{x}{\sigma_x} - \frac{y}{\sigma_y} \right)^2$ is the sum of squares of real quantities and as such it cannot be negative. At the most it can be zero.

$\therefore 2n(1 + r)$ cannot be negative and at the most it can be zero.

Hence r cannot be less than -1 and at the most it can be -1.

Similarly by expanding $\sum \left(\frac{x}{\sigma_x} - \frac{y}{\sigma_y} \right)^2$, it can be shown that this is equal

to $2n(1 - r)$ which cannot be negative and at the most it can be zero.

Hence r cannot be greater than +1 and at the most it can be +1

Hence $-1 \leq r \leq +1$ or

$$|r| \leq 1.$$

3. The value of r is very much influenced by large items, if they are present in data.

Proof: Consider the values

X	2	2	4	5	5
Y	6	3	2	6	4

The coefficient of correlation between X and Y by calculations is 0.04 (see Example 6.8).

Let there be one more entry for $X = 100$ and corresponding to it $Y=120$. Now we shall calculate the value of r .

X	X ²	Y	Y ²	XY
2	4	6	36	12
2	4	3	9	6
4	16	2	4	8
5	25	6	36	30
5	25	4	16	20
100	10000	120	14400	12000
118	10074	141	14501	12076

$$\begin{aligned}
 r &= \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[N \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}} \\
 &= \frac{6(12076) - (118)(141)}{\sqrt{[(6)(10074) - (118)^2][(6)(14501) - (141)^2]}} \\
 &= \frac{72456 - 16638}{\sqrt{[60444 - 13924][87006 - 19881]}} \\
 &= \frac{55818}{\sqrt{[46520][67125]}} \\
 &= \frac{55818}{55887} \\
 &= +0.99
 \end{aligned}$$

The value of r for five items was +0.04 but with an addition of only one pair of observations, the value of r becomes almost 1. With introduction of one large pair of values, r is affected. Therefore it is advisable to ignore very large values in the series while calculating coefficient of correlation.

4. Coefficient of correlation is the geometric mean of two regression coefficients.

$$\text{Symbolically, } r = \sqrt{b_{xy}b_{yx}}$$

Proof for this property will be given in the next unit on Regression analysis.

Self-Assessment Questions:

Short Questions

1. Explain the meaning of the concept of correlation. Does correlation always signify casual relationships between two variables? Explain with illustration on what basis can the following correlation be criticized?
 - (a) Over a period of time there has been an increased financial aid to under developed countries and also an increase in comedy act television shows. The correlation is almost perfect.
 - (b) the correlation between salaries of school teachers and amount of liquor sold during the period 1940 – 1980 was found to be 0.96.
2. What is a scatter diagram? How does it help in studying correlation between two variables, in respect of both its nature and extent?
3. Explain the meaning and significance of the concept of correlation. How will you calculate it from statistical point of view?
4.
 - (a) Define Karl Pearson's coefficient of correlation. What is it intended to measure?
 - (b) What are the special characteristics of Karl Pearson's coefficient of correlation? What are the underlying assumptions on which this formula is based?
 - (c) How do you interpret a calculated value of Karl Pearson's coefficient of correlation? Discuss in particular the values of $r = 0$, $r = -1$ and $r = +1$.
5.
 - (a) Explain what is meant by coefficient of correlation between two variables. What are the different methods of finding correlation? Distinguish between Positive and Negative correlation.
 - (b) Write down an expression for the Karl Pearson's coefficient of linear correlation. Why is it termed as the coefficient of linear correlation? Explain.
6. Define product moment correlation coefficient between two variables x and y . State its limits. Draw the scatter diagram for the extreme cases.
7.
 - (a) If x and y are independent variates then prove that they are uncorrelated. Is the converse true? Explain your answer with the help of an example.
 - (b) Prove that two independent variables are uncorrelated. By giving an example, show that the converse is not true. Explain the reason?
 - (c) Comment on the following statement:
"If the coefficient of correlation between two variables is zero, it does not mean that the variables are unrelated".
8. Discuss the statistical validity of the following statements:
 - (a) "High positive coefficient of correlation between increase in the sale of a newspaper and increase in the number of crimes, leads to the conclusion that newspaper reading may be responsible for the increase in the number of crimes."
 - (b) "A high positive value of r between the increase in cigarette smoking and increase in lung cancer establishes that cigarette smoking is responsible for lung cancer."
 - (c) If the coefficient of correlation between the annual value of exports during the last ten years and the annual number of children born during the same period is $+0.9$, what inference, if any, would you draw?

9. Comment on the following:
 - (a) "Positive correlation $r = 0.9$, is found between the number of children born and exports over last decade."
 - (b) The correlation coefficient between the railway accidents in a particular year and the babies born in that year was found to be 0.8.
10. (a) Define a scatter diagram. Draw the scatter diagram when (i) $r = +1$, (ii) $r = -1$ and (iii) $r = 0$, where r is the correlation coefficient.
 (b) What is a scatter diagram? Give the procedure of drawing a scatter diagram. Draw scatter diagrams when the coefficient of correlation $r = +1$ and $r = -1$.
11. What is correlation? What is a scatter diagram? How does it help in studying correlation between two variables, in respect of both its nature and extent?
12. What is correlation? Explain the implications of positive and negative correlation. Show by means of scatter diagram, the presence of perfect positive and perfect negative correlation.
13. What is a scatter diagram? How is it useful in the study of correlation between two variables? Explain with suitable examples.
14. Write a note on scatter diagram. Draw sketches of scatter diagram to show the following correlation between two variables x and y :
 - (i) linear; (ii) linear and perfect;
 - (iii) non-linear; (iv) x and y uncorrelated.
15. While drawing a scatter diagram, if all points appear to form a straight line going downward from left to right, then it is inferred that there is:
 - (i) Perfect positive correlation; (ii) Simple positive correlation;
 - (iii) Perfect negative correlation; (iv) No correlation.
16. Write short not on the following
 - (a) Spurious correlation
 - (b) Positive and negative correlation
 - (c) Linear and non-linear correlation
 - (d) Simple, multiple and partial correlation
17. Write short note on the following
 - (a) Karl Pearson's coefficient of correlation
 - (b) Probable Error
 - (c) Spearman's Rank Correlation Coefficient

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The _____ denotes both the magnitude and orientation of the linear correlation between the independent and dependent variables.
 - a) regression intercept b) regression gradient
 - c) Coefficient of Determination d) Correlation Coefficient
- (ii) The table below shows the number of cars sold last month by seven employees at Concord Motors and their number of years of sales experience.

Experience	Sales
1	8
2	6
2	7
4	14
5	9
6	13
8	10

The correlation coefficient for this data is _____.

- a) -0.251 b) 0.360
c) 0.553 d) 0.744

- (iii) The table below shows the number of cars sold last month by seven employees at Concord Motors and their number of years of sales experience.

Experience	Sales
1	8
2	6
2	7
4	14
5	9
6	13
8	10

The test statistic for testing whether the population correlation coefficient is greater than zero is _____.

- a) 1.48 b) 2.25
c) 3.09 d) 3.71

- (iv) What is the range of the Pearson correlation coefficient?

- a) -1 to 1 b) 0 to 1
c) -1 to 0 d) 0 to 100

- (v) Pearson's correlation coefficient is most appropriate when:

- a) The relationship between the two variables is curvilinear.
b) The two variables are measured on a nominal scale.
c) The variables are linearly related and normally distributed.
d) The relationship between the two variables is non-linear.

2. Write "T" if the statement is true and "F" if the statement is false:

- (i) Pearson's correlation coefficient (r) measures the strength and direction of a linear relationship between two variables.
(ii) The Pearson correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, not positive.
(iii) A coefficient of determination equal to zero indicates that there is no relationship between the independent and dependent variables.
(iv) If two variables have a correlation coefficient equal to +0.75, the scatter plot will have an upward slope moving from left to right.
(v) Pearson's correlation coefficient measures the strength and direction of a linear relationship, but it does not test for causation.

Answer:

Multiple-Choice Question:

1. (i) d (ii) c (iii) a (iv) a (v) c

True/False

2. (i) T (ii) F (iii) T (viii) T (v) F

Lesson 3: Rank Correlation

Lesson Objectives:

After studying this lesson, you will be able to:

- Define and discuss meaning of rank correlation;
- Understand calculation procedures of rank correlation;
- Distinguish between regression and correlation analysis;
- Define and interpret correlation coefficient (r);
- Understand the properties of r.

Introduction

The simple correlation analysis, no attempt is made to estimate one variable from another, and it makes no difference which variable is labeled X or which is labeled Y. Both are considered random variables. As mentioned earlier lesson, the purpose of correlation is to provide a mathematical statement of the degree, or closeness, of the relationship existing between variables.

Sometimes we come across some intangible qualities in statistical series in which the variables under consideration are not capable of quantitative measurement but can be arranged in serial order. This happens when we are dealing with qualitative characteristics (attributes) such as leadership quality, personality, honesty, beauty, character, morality, value of employees to the firm etc., cannot be easily measured and assigned scores, but it is often possible to judge these qualities, to rank them, and to compare ranks assigned with ranks given by others or with ranked scores made on some other basis. Such type of problems cannot be solved with the help of Karl Pearson's coefficient of correlation. Charles Edward Spearman, a British psychologist, developed a formula in 1904 which consists in obtaining the correlation coefficient between the ranks of n individuals in the two attributes under study. This method is especially useful in cases when the actual magnitudes or item-values are not given and simply their ranks in the series are known. Spearman's coefficient of correlation is computed by ranking various item-values in the two variables, finding out the difference in ranks, squaring them and finding out the aggregate of the squared differences.

Rank correlation is especially useful in cases when the actual magnitudes or item-values are not given and simply their ranks in the series are known.

Spearman's coefficient of correlation is computed by ranking various item-values in the two variables, finding out the difference in ranks, squaring them and finding out the aggregate of the squared differences.

$$\text{Symbolically, } \rho = 1 - \frac{6 \sum d^2}{N(N^2-1)} \text{ or} \\ \rho = 1 - \frac{6 \sum d^2}{N^3-N}$$

where ρ (Greek letter Rho) denotes the coefficient of correlation; $\sum d^2$ denotes the sum of the squared differences and the number of pairs of the observations.

The value of Spearman's coefficient also ranges between + 1 and -1. When ρ is +1, the concordance between rankings is perfect and the ranks are in the same direction. When ρ is -1, there is also perfect concordance between rankings but the ranks are in opposite direction. If the

concordance between rankings is perfect, Σd^2 will be equal to zero and ρ will be +1. These points are explained in the table below.

Ranks in		Differenced ($R_1 - R_2$)	d^2	Ranks in		d ($R_1 - R_2$)	d^2
Statistics (R_1)	Economics (R_2)			Stat. (R_1)	Eco. (R_2)		
1	1	0	0	1	4	-3	9
2	2	0	0	2	3	-1	1
3	3	0	0	3	2	1	1
4	4	0	0	4	1	3	9
			$\Sigma d^2 =$ 0				$\Sigma d^2 =$ 20

$$\begin{aligned}\rho &= 1 - \frac{6 \Sigma d^2}{N^3 - N} \\ &= 1 - \frac{6(0)}{4^3 - 4} \\ &= 1 - 0 \\ &= 1\end{aligned}$$

$$\begin{aligned}\rho &= 1 - \frac{6 \Sigma d^2}{N^3 - N} \\ &= 1 - \frac{6(20)}{4^3 - 4} \\ &= 1 - \frac{120}{60} \\ &= 1 - 2 = -1\end{aligned}$$

Calculation of ρ

The following procedure may be followed for calculating ρ .

1. Arrange the various item-values in the two series according to their ranks. If there are two items having the same value say 6, in a series, then both of them should be assigned 6.5th rank $= \frac{(6+7)}{2}$ and the succeeding value the 8th rank, and so on.
2. Find out the differences in the ranks of both the series.
3. Square the differences in the ranks and sum them up.
4. Apply the Spearman's formula for rank correlation, i.e.

$$\rho = 1 - \frac{6 \Sigma d^2}{N(N^2 - 1)}$$

where ρ (Greek letter Rho) denotes the correlation and Σd^2 , the sum of the squared differences and N the number of pairs of the series.

If the concordance between rankings is perfect, Σd^2 will be equal to zero and ρ will be 1. If the concordance is perfect ρ will be equal to -1. In other cases, ρ lies between +1 and -1.

The following examples clarify the formula.

Example 6.16. The rankings of ten students in Statistics and Economics are as follows:

Statistics:	3	5	8	4	7	10	2	1	6	9
Economics:	6	4	9	8	1	2	3	10	5	7

Required:

What is the coefficient of rank correlation?

Solution:**Calculation of Spearman's correlation coefficient**

Ranks		Rank Difference $d = (R_1 - R_2)$	Square of ranks difference d^2
Statistics (R_1)	Statistics (R_2)		
3	6	-3	9
5	4	+1	1
8	9	-1	1
4	8	-4	16
7	1	+6	36
10	2	+8	64
2	3	-1	1
1	10	-9	81
6	5	+1	1
9	7	+2	4
Total		$\Sigma d = 0$	$\Sigma d^2 = 214$

$$\rho = 1 - \frac{6 \Sigma d^2}{N(N^2 - 1)}$$

In this example ranks are given to us. Where actual values are given, then we have to find out ranks. The following example illustrates this point.

Example 6.17. Calculate Spearman's rank correlation coefficient between advertisement cost and sales from the following data:

Advertisement cost ('000 Tk.): 39	65	62	90	82	75	25	98	36	78
Sales (lakhs Tk.) 47	53	58	86	62	68	60	91	51	84

Solution: Let X denote the advertisement cost ('000 Tk.) and Y denote the sales (lakhs Tk.).

Calculation of Rank Correlation Coefficient

X	Y	Rank of $X (x)$	Rank of $Y (y)$	$d = x - y$	d^2
39	47	8	10	-2	4
65	53	6	8	-2	4
62	58	7	7	0	0
90	86	2	2	0	0
82	62	3	5	-2	4
75	68	5	4	1	1
25	60	10	6	4	16
98	91	1	1	0	0
36	51	9	9	0	0
78	84	4	3	1	1
				$\Sigma d = 0$	$\Sigma d^2 = 30$

Example 6.18. Compute Rank correlation from. the following table of Index Numbers of supply (X) and price (Y) :

X	115	134	120	130	124	128
Y	130	132	128	130	127	125

Solution: Calculation of Spearman's correlation coefficient

X	Ranks R_1	Y	R_2	d $R_1 - R_2$	d^2
135	6	130	2.5	3.5	12.25
134	1	132	1	0	0
120	5	128	4	1	1
130	2	130	2.5	-0.5	0.25
124	4	127	5	-1	1
128	3	125	6	-3	9
					$\Sigma d^2 = 23.50$

$$\rho = 1 - \frac{6 \Sigma d^2}{N(N^2 - 1)} = 1 - \frac{6(23.50)}{6^3 - 6}$$

$$= 1 - \frac{141}{210} = \frac{69}{210} = 0.33$$

Example 6.19. The contestants in a beauty contest are ranked by three judges in the following order:

First Judge	1	6	5	10	3	2	4	9	7	8
Second Judge	3	5	8	4	7	10	2	1	6	9
Third Judge	6	4	9	8	1	2	3	10	5	7

Required:

Use the rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

Solution: Let us calculate three sets of correlation between the ranks of (i) First and second judges; (ii) First and third judges; (iii) Second and third judges.

First Judge Opinions d_1	Second Judge Opinions d_2	Third Judge Opinions d_3	$d_{12} = (d_1 - d_2)$	d_{12}^2	$d_{13} = (d_1 - d_3)$	d_{13}^2	$d_{23} = (d_3 - d_2)$	d_{23}^2
1	3	6	-2	4	-5	25	-3	9
6	5	4	+1	1	+2	4	+1	1
5	8	9	-3	9	-4	16	-1	1
10	4	8	+6	36	+2	4	-4	16
3	7	1	-4	16	+2	4	+6	36
2	10	2	-8	64	0	0	+8	64
4	2	3	+2	4	+1	1	-1	1
9	1	10	+8	64	-1	1	9	81
7	6	5	+1	1	+2	4	+1	1
8	9	7	-1	1	+1	1	+2	4
N=10	Totals		0	200	0	60	0	214

(i) Rank correlation between the opinions of the first and second judges:

$$\rho = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$$

$$\text{where } d_{12}^2 = 200, N = 10$$

$$\begin{aligned} \therefore \rho_{12} &= 1 - \frac{6 \times 200}{10(10^2 - 1)} = 1 - \frac{1200}{10(100 - 1)} \\ &= 1 - \frac{1200}{10(99)} = 1 - \frac{1200}{990} = \frac{210}{990} = +0.23 \end{aligned}$$

(ii) Ranks correlation between the opinions of the first and third Judges.

$$\begin{aligned} \rho_{13} &= 1 - \frac{6 \sum d^2}{N(N^2 - 1)} \\ &= 1 - \frac{6(60)}{10(10^2 - 1)} \\ &= 1 - \frac{360}{990} = \frac{630}{990} = +0.64 \end{aligned}$$

(iii) Rank correlation between second and third Judges:

$$\begin{aligned} \rho_{23} &= 1 - \frac{6 \sum d^2}{N(N^2 - 1)} \\ &= 1 - \frac{6 \times 214}{10(10^2 - 1)} \\ &= 1 - \frac{1284}{990} = \frac{294}{990} = -0.3 \end{aligned}$$

Thus we conclude that the second pair of judges have the nearest approach to common taste of beauty.

Merits and Demerits of the Rank Method

Merits

1. We always have $\sum d = 0$, which provides a check for numerical calculations.
2. Since Spearman's rank correlation coefficient ρ is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.
3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure which is distribution-free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's ρ is such a measure (*i.e.*, distribution-free), since no strict assumptions are made about the form of the population from which sample observations are drawn.

Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics.

4. Spearman's formula is easy to understand and apply as compared with Karl Pearson's formula. The values obtained by the two formulae, viz., Pearsonian r and Spearman's ρ are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.
5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time consuming.

Demerits

1. The method cannot be employed for finding out correlation in a grouped frequency distribution.
2. If the number of item-values exceed 30, it becomes a difficult task to find out ranks and their differences. Therefore, it is advisable not to resort to rank correlation where N is exceeding 30 unless ranks are given to us.
3. As all the information concerning the variables is not utilized, this method lacks precision as compared to Pearson's method.

Remarks on Spearman's Rank Correlation Coefficient

1. We always have $\sum d = 0$, which provides a check for numerical calculations.
2. Since Spearman's rank correlation coefficient ρ is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.
3. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure which is distribution-free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's ρ is such a measure (i.e., distribution-free), since no strict assumptions are made about the form of the population from which sample observations are drawn.
4. Spearman's formula is easy to understand and apply as compared with Karl Pearson's formula. The values obtained by the two formulae, viz., Pearsonian r and Spearman's ρ are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is always some loss of information. Unless many ties exist, the coefficient of rank

correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.
6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time consuming.

Self-Assessment Questions:

Short Questions

1. What is Spearman's rank correlation coefficient? Discuss its usefulness.
 2. Explain the difference between Karl Pearson's (product moment) correlation coefficient and rank correlation coefficient.
 3. What are the advantages of Spearman's rank correlation coefficient over Karl Pearson's correlation coefficient? Explain the method of calculating Spearman's correlation coefficient.
 4. Define rank correlation coefficient. When is it preferred to Karl Pearson's coefficient of correlation?
 5. Distinguish between Karl Pearson's coefficient of correlation and Spearman's rank correlation coefficient. Explain with the help of an example when Spearman rank correlation coefficient results to $+1$, -1 and between -1 to $+1$.
 6. Define Rank Correlation. Write down Spearman's formula for rank correlation coefficient. What are the limits of ρ ? Interpret the case when ρ assumes the minimum value.
- 2. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:**

- (i) Which of the following is the correct range for Spearman's rank correlation coefficient?
a) -1 to 1 b) -2 to 2 c) 0 to 1 d) 0 to 100
- (ii) What does a Spearman's rank correlation coefficient of 1 indicate?
a) Perfect negative correlation b) No relationship
c) Perfect positive correlation d) Weak positive correlation
- (iii) Which of the following is the key assumption of rank correlation methods?
a) Data must be on an interval scale
b) The relationship between the variables must be linear
c) The relationship must be monotonic
d) There should be no ties in the ranks

- (iv) Which of the following is the range of the Spearman's rank correlation coefficient?
 - a) -2 to 2 b) -1 to 1 c) 0 to 1 d) -5 to 5
- (v) In rank correlation, what happens when the ranks are tied?
 - a) The method ignores the ties
 - b) Each tied value gets the same rank
 - c) The ranks are adjusted for the number of ties
 - d) Tied ranks are assigned the highest rank

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) Spearman's rank correlation coefficient is only applicable to nominal data.
- (ii) A Spearman's rank correlation coefficient of 0 means there is no monotonic relationship between the variables.
- (iii) Spearman's rank correlation can be used even when the data is not normally distributed.
- (iv) Rank correlation methods assume that the relationship between variables is linear.
- (v) Spearman's rank correlation can be used for both ordinal and interval data.

Answer:

Multiple-Choice Question:

1. (i) a. (ii) c (iii) c (iv) b (v) b

True/False

2. (i) F (ii) T (iii) T (iv) F (v) T

Lesson 4: Concurrent Deviation and Least Squares Method

Lesson Objectives:

After studying this lesson, you will be able to:

- Define and discuss meaning of the concurrent deviation method;
- Understand of computation of concurrent method;
- Explain the merits and demerits of concurrent method;
- Define methods of least squares.

Introduction:

This is a casual method of determining the correlation between two series when we are not very serious about its precision. This is based on the signs of the deviations (*i.e.* the direction of the change) of the values of the variable from its preceding value and does not take into account the exact magnitude of the values of the variables. Thus, we put a plus (+) sign, minus (-) sign or equality (=) sign for the deviation if the value of the variable is greater than, less than or equal to the preceding value respectively. The deviations in the values of two variables are said to be concurrent if they have the same sign (either both deviations are positive or both are negative or both are equal).

Concurrent deviation is a casual method of determining the correlation between two series when we are not very serious about its precision.

The Concurrent Deviation Method

This method is one of the ways of ascertaining the coefficient of correlation by an extremely simple calculation. It is based on the direction of change or variation in the two paired variables. It like Pearson's r and Spearman's ρ , varies between +1 and -1.

This method is one of the ways of ascertaining the coefficient of correlation by an extremely simple calculation.

The formula applicable is:

$$r_c = \pm \sqrt{\frac{\pm 2c - n}{n}}$$

where r_c represents coefficient of correlation by the Concurrent Deviation method; c denotes the number of concurrent deviations or the number of positive signs obtained after multiplying D_x with D_y and n equals one less than the number of pairs of items or equal to the total number of deviations.

It should be noted that under this method deviations are not calculated from the arithmetic mean but are taken from the preceding items. But the direction of the deviation (*i.e.*, positive or negative) and not the extent of deviation is considered.

The steps in computation of this coefficient are:

1. First of all, find out the direction of change in the values of X variable. The first item-value will be the starting point and from it the direction of movement will be marked as plus or minus in the case of second and in the successive item values from their preceding figures. For instance, in Example 6.20, the second value of X (36) is compared with the first value (8) and as it exceeds it, therefore + sign is put against the second value. Similarly the third value is compared with the second and so on. Denote this column of deviations by d_x .

2. Repeat the process for Y variable and find out the deviations and denote this column by d_y .

3. Compare the deviations in both the columns and find out concurrences and disagreements. The plus of X series and the minus of Y series will show disagreement represented by a minus sign. The minus of X series and the minus of Y series or similar plus signs of both the series will indicate concurrences.

4. Coefficient of concurrent deviations is primarily based on the following principle:

“If the short time fluctuations of the time series are positively correlated or in other words, if their deviations are concurrent, their curves would move in the same direction and would indicate positive correlation between them”

5. Determine the value of c, i.e., the number of positive signs or concurrences.

6. Apply the above formula, i.e.,

$$r_c = \pm \sqrt{\pm \left(\frac{2c - n}{n} \right)}$$

It may be noted that n in this case will always mean $(\sum f - 1)$ or $N-1$ where N =total number of items, because the first item-value in both the series serve as starting points only.

Example 6.20. The following are the marks obtained by a group of 10 students in Economics and Statistics:

Students	1	2	3	4	5	6	7	8	9	10
Marks in Economics	8	36	98	25	75	82	90	62	65	39
Marks in Statistics	84	51	91	60	68	62	86	58	53	47

Calculate r by the method of Concurrent Deviations.

Solution:

Computation of Coefficient of Concurrent Deviation between Marks in Economics and Statistics.

Students	Marks in Economics (X)	Deviation from Preceding item d_x	Marks in Statistics (Y)	Deviation from Preceding item d_y	$d_x d_y$
1	8		84		
2	36	+	51	-	-
3	98	+	91	+	+
4	25	-	60	-	+
5	75	+	68	+	+
6	82	+	62	-	-
7	90	+	86	+	+
8	62	-	58	-	+
9	65	+	53	-	-
10	39	-	47	-	+
		N=9		N=9	C=6

$$r_c = \pm \sqrt{\pm \frac{2c-n}{n}} = + \sqrt{\frac{2(6)-9}{9}}$$

$$= + \sqrt{\frac{12-9}{9}} = + \sqrt{\frac{3}{9}} = 0.58$$

Example 6.21. Calculate the coefficient of concurrent deviations for the following data :

Supply :	65	40	35	75	63	80	35	20	80	60	50
Demand :	60	55	50	56	30	70	40	35	80	75	80

Solution

Calculations for Coefficient of Concurrent Deviations

Supply (X)	Sign of the deviation from preceding value (x)	Demand (Y)	Sign of the deviation from preceding value (y)	Product of deviations (xy)
65		60		
40	-	55	-	+
35	-	50	-	+
75	+	56	+	+
63	-	30	-	+
80	+	70	+	+
35	-	40	-	+
20	-	35	-	+
80	+	80	+	+
60	-	75	-	+
50	-	80	+	-

Here we have: n = Number of pairs of deviations = 11-1=10

c = Number of pairs of deviations having like signs = 9

The coefficient of concurrent deviations is given by:

$$r = \pm \sqrt{\pm \frac{(2c-n)}{n}} = \pm \sqrt{\pm \frac{2 \times 9 - 10}{10}} = \pm \sqrt{\pm 0.8}$$

since (2c-n) = 8, is positive, we take positive sign inside and outside the square root so that: $r = +\sqrt{0.8} = 0.89$

Example 6.22: Calculate r_c from the following table:

X	368	384	384	385	361	347	384	395	403	385
Y	22	21	24	20	22	26	26	29	28	27

Solution:

X	d _x	Y	d _y	d _x d _y
368		22		
384	+	21	-	-
385	+	24	+	+
361	-	20	-	+
347	-	22	+	-
384	+	26	+	+
395	+	26	0	0
403	+	29	+	+
400	-	28	-	+
385	-	27	-	+
				Σd _x d _y =6

$$\begin{aligned}
 rc &= \pm \sqrt{\pm \frac{2c-9}{9}} \\
 &= \pm \sqrt{\pm \frac{3}{9}} \\
 &= +0.58
 \end{aligned}$$

Merits and Demerits of Concurrent Deviation Method**Merits**

1. It is simple to compute and easy to understand. As compared to other methods, it is the simplest and the easiest.
2. It is quite suitable for measuring the extent of correlation if the variable includes item-values having short-term oscillations or fluctuations.

Demerits

1. It is not useful if long-term changes are to be considered because it ignores the general trend of variation. The method does not differentiate between small and big variations. For example, if X increases from 50 to 51 the sign will be plus and if Y increases from 50 to 151 the sign will be plus. Both get equal weight when they are in the same direction.
2. It indicates the direction of change only and therefore the results obtained by this method can be only a rough indicator of the presence or absence of correlation.

However, it may be noted that results obtained by this method are not very much different from those obtained by the use of Karl Pearson's coefficient in the case of short-term oscillations only.

Method of Least Squares

The least squares method is a mathematical procedure used to identify the linear equation that best fits a set of ordered pairs. The line that best fits the ordered pairs is called the regression line. The procedure can be used to find the values for b_0 (the y-intercept) and b_1 (the slope of the line). The goal of the least squares method is to minimize the total squared error between the values of y and $\hat{y}$, which provides us with the best fitting line through our data points.

This method will be discussed in details in our next unit Regression Analysis.

The least squares method is a mathematical procedure used to identify the linear equation that best fits a set of ordered pairs. The line that best fits the ordered pairs is called the regression line.

Miscellaneous Problems

1. Two series X and Y with 50 items each have standard deviations 4.5 and 3.5 respectively. If the summation of products of deviations of X and Y series from their respective arithmetic means be 420, find the r between X and Y.

Solution:

Given $N = 50$; $\sigma_x = 4.5$; $\sigma_y = 3.5$; and $\sum xy = 420$

$$r = \frac{\sum xy}{\sigma_x \sigma_y}$$

$$r = \frac{420}{50(4.5)(3.5)} = \frac{420 \times 2 \times 2}{50 \times 9 \times 7} = \frac{8}{15} = +0.53$$

2. Use the formula

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y}$$

where σ_{x-y} denote standard deviation of (X-Y) series.

2. To compute the coefficient of correlation between X and Y series.

X:	2	5	7	10	19	17
Y:	26	29	26	30	34	35

Solution:

X	$d_x = X - 10$	d_x^2	Y	$d_y = Y - 10$	d_y^2	$D = Y - X$	$D' = Y - X$	
2	-8	64	26	-4	16	24	4	16
5	-5	25	29	-1	1	24	4	16
7	-3	9	26	-4	16	19	-1	1
10	0	0	30	0	0	20	0	0
19	+9	81	34	+4	16	15	-5	25
17	+7	49	35	+5	25	18	-2	4
60		228	180	0	74	12	0	62

$$\sigma_x^2 = \frac{\sum d_x^2}{n} - \left(\frac{\sum d_x}{n}\right)^2$$

$$= \frac{228}{6} - \left(\frac{0}{6}\right)^2 = 38$$

$$\sigma_y^2 = \frac{\sum d_y^2}{n} - \left(\frac{\sum d_y}{n}\right)^2$$

$$= \frac{74}{6} - \left(\frac{0}{6}\right)^2 = 12.34$$

$$D' = \frac{\sum D}{n} = \frac{120}{6} = 20$$

$$= \sigma_{x-y}^2 = \frac{\sum D'^2}{n} = \frac{62}{6} = 10.34$$

$$r = \frac{38 + 12.34 - 10.34}{2\sqrt{38 \times 12.34}} = \frac{40}{2(6.2)(3.5)}$$

$$= +0.92.$$

3. Two series X and Y have standard deviations 2.5 and 3.5 respectively. If the coefficient of correlation between X and Y is 0.6, what is the standard deviation of the difference between X and Y?

Solution:

In this type of a problem, we have to employ the formula given in Problem No. 2 above, i.e.,

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$$

Substituting the value, we get

$$0.6 = \frac{2.5^2 + 3.5^2 - \sigma_{x-y}^2}{2(2.5)(3.5)}$$

$$0.6 \times 17.50 = 6.25 + 12.25 - \sigma_{x-y}^2$$

$$\sigma_{x-y}^2 = 18.50 - 10.50 = 8$$

$$\sigma_{x-y} = \sqrt{8} = 2.9 \text{ approx.}$$

4. From the data given below find the number of items, i.e., n,

$$r = 0.5, \sum xy = 120, \sigma_y = 8, \sum x^2 = 90$$

Solution:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}; \sigma_y = \sqrt{\frac{\sum y^2}{n}} \text{ or } \sum y^2 = n\sigma_y^2$$

$$0.5 = \frac{120}{\sqrt{90 \times n \times 8^2}} \text{ or } 0.5 = \frac{120}{8\sqrt{90}\sqrt{n}}$$

$$\therefore \sqrt{n} = \frac{120}{0.5 \times 8 \times \sqrt{90}}$$

$$= \frac{120 \times 120}{0.25 \times 64 \times 90} = 10$$

5. Coefficient of correlation between two variates X and Y is 0.48. Their co-variance is 36. The co-variance of X is 16, find the standard deviation of Y series.

Solution:

Covariance means $\frac{\sum xy}{n}$ where x and y denote the deviations of values from arithmetic mean.

$$\text{Variance of X} = 16 \quad \therefore \sigma_x = 4$$

$$\text{Now } r = \frac{\sum xy}{n\sigma_x\sigma_y} = \left(\frac{\sum xy}{n}\right) \sigma_x\sigma_y$$

Substituting the given values,

$$0.48 = \frac{36}{4(0.48)} = \left(\frac{36}{1.92}\right) = 18.7$$

6. A student calculates the value of r as $+0.7$ when the value of N is 5 and concludes that r is highly significant. Is he correct?

Solution: The value of r can be correctly interpreted with the help of probable error. We know that if the value of r is more than 6 times the P.E. it is considered to be significant.

$$\text{P.E.} = \frac{2}{3} \frac{(1 - r^2)}{\sqrt{N}} = \frac{2}{3} \frac{(1 - 0.7^2)}{\sqrt{5}} = \frac{2}{3} \frac{0.51}{2.24}$$

$$= \frac{0.34}{2.24} = 0.15$$

$$\frac{r}{\text{P.E.}} = \frac{0.7}{0.15} = 4.5$$

Since r is less than six times the P.E., it is not significant.

(**Note:** In the solution, we assumed that $0.6745 = \frac{2}{3}$)

7. In a question on correlation based on 25 observations the value of P.E. was calculated to be 0.245. Find the value of r .

Solution:

$$\text{P.E. of } r = \frac{2}{3} \frac{(1 - r^2)}{\sqrt{N}}$$

(where $\frac{2}{3} = 0.6745$)

Given P.E.=0.245 and $N= 25$

Substituting the values in the formula we get

$$0.245 = \frac{\frac{2}{3}(1 - r^2)}{\sqrt{25}}$$

$$1.225 = \frac{2}{3} - \frac{2}{3}r^2$$

$$1.225 - \frac{2}{3} = -\frac{2}{3}r^2$$

$$-\frac{2}{3}r^2 = \frac{2}{3} - 1.225$$

$$-\frac{2}{3}r^2 = -0.558$$

$$r^2 = \frac{0.558 \times 3}{2} = 0.837$$

$$r = \sqrt{0.837} = 0.915$$

8. In a question on correlation the value of r is 0.64 and its P.E.=.1312. What was the value of N ?

Solution:

$$P.E. = .6745 \frac{1 - r^2}{\sqrt{N}}$$

Given P.E.=.1312, and $r=.64$. Substituting the values in the given formula

$$.1312 = \frac{.6745(1 - .64^2)}{\sqrt{N}}$$

$$.1312 = \frac{.6745(1 - .4096)}{\sqrt{N}}$$

$$.1312\sqrt{N} = .6745(.5904) = .3936$$

$$\sqrt{N} = \frac{.3936}{.1312} = 3$$

$$N = 9$$

9. What is the significance of the coefficient of correlation, r for the following values based on the number of observation (a) 50, (b) 500 :

$$r=0.2 ; 0.4 ; 0.9$$

Solution:

(a) Where $n = 50$.

$r = 0.2$	$r = 0.4$	$r = 0.9$
$P.E. = .6745 \frac{1 - .2^2}{\sqrt{50}}$ $P.E. = .09$ $3 P.E. = 0.27$ The value of r is less than 3 P.E. Therefore there is no evidence that r is significant.	$P.E. = .6745 \frac{1 - .4^2}{\sqrt{50}}$ $P.E. = .08$ $3 P.E. = 0.24$ $5 P.E. = 0.40$ The value of r lies between 3 P.E. and 5 P.E. We can say with reasonable confidence that correlation is significant.	$P.E. = .6745 \frac{1 - .9^2}{\sqrt{50}}$ $P.E. = .018$ $5 P.E. = 0.09$ The value of r is much greater than 5 P.E., therefore we can say that correlation is significant.

(b) When $n=500$.

$r = 0.2$	$r = 0.4$	$r = 0.9$
$P.E. = .6745 \frac{1 - .2^2}{\sqrt{500}}$ $P.E. = .96 \times .03 = 0.29$ $5 P.E. = 0.145$ The given value of r is greater than 5 P.E., r is quite significant	$P.E. = .6745 \frac{1 - .4^2}{\sqrt{500}}$ $P.E. = .84 \times .03 = 0.025$ $5 P.E. = 0.125$ The value of r is much greater than 5 P.E., correlation is quite significant.	$P.E. = .6745 \frac{1 - .9^2}{\sqrt{500}}$ $P.E. = .19 \times .03 = .006$ $5 P.E. = 0.03$ The value of r is very much greater than 5 P.E., correlation is highly significant.

10. Which value of r is more significant?

(i) $r = .6$ P.E. = 0.05

(ii) $r = .9$ P.E. = 0.09

Solution

(i) $\frac{r}{\text{P.E.}} = \frac{.6}{.05} = 12$

(ii) $\frac{r}{\text{P.E.}} = \frac{.9}{.09} = 10$

It is obvious that r is more significant in (i).

11. The coefficient of rank correlation of the marks obtained by 10 students in English and Economics was found to be 0.5. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was roughly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Solution:

Given $r_k = 0.5$; $N = 10$

The formula for $r_k = 1 - \frac{6 \sum D^2}{N^3 - N}$

Substituting the given data in the formula,

$$0.5 = 1 - \frac{6 \sum D^2}{10^3 - 10}$$

$$= 1 - \frac{6 \sum D^2}{990}$$

$$\frac{6 \sum D^2}{990} = .5$$

$$6 \sum D^2 = 495$$

$$\sum D^2 = 82.5$$

But this is not the correct $\sum D^2$

The correct $\sum D^2 = 82.5 - 3^2 + 7^2$

$$= 82.5 - 9 + 49 = 122.5$$

$$r_k = 1 - \frac{6 \times 122.5}{990} = 1 - \frac{735}{990}$$

$$= 1 - 0.742 = +0.258$$

12. The coefficient of rank correlation between marks in statistics and marks in Accountancy obtained by a certain group of students is 0.8. If the sum of the squares of the differences in ranks is given to be 33, find the number of students in the group.

Solution:

Given $r=0.8$; $\sum D^2 = 33$

The formula for rank correlation $r_k = 1 - \frac{6 \sum D^2}{N^3 - N}$ Substituting the values in the formula, we get

$$0.8 = 1 - \frac{6 \times 33}{N^3 - N}$$

$$\frac{198}{N^3 - N} = 1 - .8 \text{ or } \frac{198}{N^3 - N} = 0.2$$

$$2(N^3 - N) = 1980$$

$$N^3 - N = 990$$

$$N^3 - N - 990 = 0$$

$$N^3 - N - 1000 - 10 = 0$$

$$N^3 - 10^3 - (N - 10) = 0$$

$$(N - 10)(N^2 + 100 + 10N) - (N - 10) = 0$$

$$(N - 10)(N^2 + 99 + 10N) = 0$$

$$\therefore N = 10 \text{ or } N^2 + 99 + 10N = 0$$

which gives imaginary values of N because $b^2 - 4ac$ is negative.

Therefore, the number of items in the group is 10.

Self-Assessment Questions:

Short Questions

1. Explain the method of concurrent deviations for computing the correlation between two variable series.
2. Give the points of strength and weakness of finding out the relationship between two variables by the method of concurrent deviations.
3. Define least square method. State the procedure of this method.

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which of the following is the primary goal of the Concurrent Deviation Method?
 - a) To find the exact functional relationship between two variables
 - b) To compare the deviations of two variables from their means
 - c) To create a regression line for predictions
 - d) To calculate the standard deviation of the data

- (ii) What does the Concurrent Deviation Method primarily analyze?
 - a) Linear relationship between variables
 - b) Correlation between two variables
 - c) Regression model fitting
 - d) The mean difference between variables
 - (iii) In the Concurrent Deviation Method, if one variable increases while the other decreases, what type of relationship is inferred?
 - a) Positive correlation
 - b) No correlation
 - c) Negative correlation
 - d) No relationship
 - (iv) Which scale of measurement is appropriate for the Concurrent Deviation Method?
 - a) Nominal scale
 - b) Ordinal scale
 - c) Interval or ratio scale
 - d) Any scale
 - (v) Which of the following is NOT a limitation of the Concurrent Deviation Method?
 - a) It assumes a linear relationship
 - b) It requires data to be on an interval or ratio scale
 - c) It is highly accurate in predicting exact values
 - d) It is not suitable for non-linear relationships
- 2. Write “T” if the statement is true and “F” if the statement is false:**
- (i) The Concurrent Deviation Method is used to find the relationship between two variables in time series data.
 - (ii) In the Concurrent Deviation Method, if both variables show similar deviations, a positive correlation exists.
 - (iii) The Concurrent Deviation Method assumes that the relationship between variables is linear.
 - (iv) The Method of Least Square is used to minimize the sum of the squared differences between observed and predicted values.
 - (v) The Method of Least Squares is applicable to both simple and multiple regression models.

Answer:

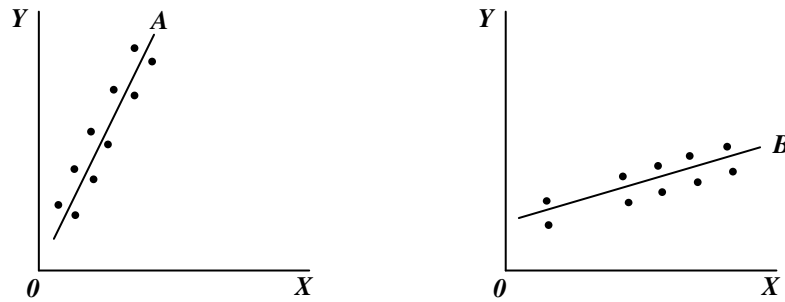
Multiple-Choice Question: 1. (i) b (ii) d (iii) c (iv) c (v) c

True/False: 2. (i) T (ii) T (iii) F (iv) T (v) T

Exercise

1. How will you interpret the following values of the coefficient of correlation?
 - (i) +.5, (ii) -.29 (iii) +.45.
2. Define the coefficient of correlation between x and y . State what precautions are to be taken in interpreting it. ·what inferences would you draw if the correlation coefficient were:
 - (i) +0.9,
 - (ii) -9.8 and
 - (iii) + 0.01?

3. From the scatter diagram decide which of the two pairs of variables show the greater correlation. Explain your answer.



4. Prove that $-1 \leq r \leq +1$, where r is correlation coefficient.
 5. Following are the heights and weights of 10 students of B. Com. Com:

Height (inches)	62	72	68	58	65	70	66	63	60	72
Weight (kgs)	50	65	63	50	54	60	61	55	54	65

Required:

Draw a scatter diagram and indicate whether the correlation is positive or negative.

6. A sample of five items is taken from the production of a firm Length and weight of the five items are given below:

Length (Inches) :	3	4	6	7	10
Weight (Ounces) :	9	11	14	15	16

Required:

Calculate Karl Pearson: correlation coefficient between length and weight in the above sample and interpret the value of this coefficient.

7. A random sample of the following two variables was obtained:

x	29	48	28	22	28	42	33	26	48	44
y	16	46	34	26	49	11	41	13	47	16

Required: Calculate the correlation between these two variables.

8. Following marks were obtained by 12 students in Mathematics and Statistics:

Math:	50	54	56	59	60	62	61	65	67	71	71	74
Stat:	22	25	34	28	26	30	32	30	28	34	36	40

Required: Find r .

9. The following table gives the marks obtained by a group of 12 students in two examinations A and B. Calculate the correlation coefficient between the marks obtained in the two examinations and interpret the result:

Student	1	2	3	4	5	6	7	8	9	10	11	12
A	15	13	17	14	18	12	20	16	18	17	19	21
B	18	16	18	15	19	16	18	15	21	17	18	20

10. You are given the following data for variables x and y :

x	y
3.0	1.5
2.0	0.5
2.5	1.0
3.0	1.8
2.5	1.2
4.0	2.2
1.5	0.4
1.0	0.3
2.0	1.3
2.5	1.0

Required:

- Plot these variables in scatter plot format. Based on this plot, what type of relationship appears to exist between the two variables?
 - Compute the correlation coefficient for these sample data. Indicate what the correlation coefficient measures.
11. The following table gives the index numbers of industrial production of Great Britain and the number of registered unemployed persons in the same country during 1924-31:

Year	Industrial Production Index Number	Number of Registered Unemployed (Hundred Thousand)
1924	100	11.3
1925	102	12.4
1926	104	14.0
1927	107	11.1
1928	1058	12.3
1929	112	12.2
1930	103	19.1
1931	94	26.4

Required:

Calculate coefficient of correlation between production and the number of unemployed.

12. In ten areas, the infant mortality (y) and birth rate (x) are found to be

x :	23	18	21	21	21	21	18	14	23	18
y :	44	46	56	42	32	47	38	45	41	52

Required: Calculate the correlation coefficient between x and y .

13. The facilities management department at a university wants to analyze the expenditure (in \$thousands) on the university facilities with the increasing number of students for the past 10 years. This is to ensure that all university facilities and lands support the academic, research and administrative utilities of the university. The data were collected from the past 10 years' financial report, and a correlation coefficient of 0.95 between the expenditure (in \$thousands) and the number of students was obtained.

Required:

Interpret the correlation coefficient value to indicate the strength and direction between the two variables.

14. Find the correlation between height of father (X) and height of son (Y) from the following data and comment on its value:

X	65	66	67	67	68	69	70	72	64	61
Y	67	68	65	66	72	69	71	68	65	60

15. Given the bivariate data:

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	5

Required:

Calculate Karl Pearson's correlation coefficient.

16. Calculate the coefficient of correlation between income and weight from the following data. What conclusions do you draw from the estimate?

Income	Weight
Taka	(lbs.)
150	172
154	180
160	160
172	180
160	170
165	190
180	200

17. Find out the Pearson's coefficient of correlation from the following table

Year	Number of Vehicles with licences ('000)	Motor vehicles accidents ('00)
1964	5.2	11.8
1965	5.6	12.0
1966	5.8	12.4
1967	6.2	12.4
1968	6.4	15.2
1969	4.6	14.0
1970	5.0	14.0
1971	3.6	11.0

18. How does the positive correlation differ from the negative? Computer of the short-term oscillation from the following data:

Year	Supply	Price
1921	80	146
1922	82	140
1923	86	130
1924	91	117
1925	83	133
1926	85	127
1927	89	115
1928	96	95
1929	93	100

(Assume a three-year cycle and ignore decimals)

19. Calculate the coefficient of correlation between price of wheat and average annual rainfall in a region.

Year	Price of what per 5 kilo Rs.	Average annual rainfall Inches
1	3.60	36
2	2.60	39
3	3.40	39
4	3.10	32
5	2.70	35
6	3.0	40
7	2.00	33
8	2.30	59
9	3.10	36
10	3.20	44
11	3.80	36

20. The following table gives the relative values of two variables:

X	42	44	58	50	90	88	60
Y	56	40	50	52	60	70	58

Required: Calculate Karl Pearson's coefficient of correlation.

21. The following data for the dependent variable, y, and the independent variable, x, have been collected using simple random sampling:

x	y
10	120
14	130
16	170
12	150
20	200
18	180
16	190
14	150
16	160
18	200

Required:

- Construct a scatter plot for these data. Based on the scatter plot, how would you describe the relationship between the two variables?
- Compute the correlation coefficient.

22. Calculate Person's coefficient of correlation from the bivariate sample of 50 distributed as below:

X →	30-35	35-40	40-45	45-50	50-55	55-60
↓Y						
80-90	2	3	2	-	-	-
90-100	-	2	5	4	2	-
100-110	-	4	8	5	1	-
110-120	-	-	2	3	1	1
120-130	1	-	-	2	1	1

23. The following table gives the number of candidates obtaining different marks in Economics and Statistics.

Marks in Economics	Marks in Statistics				Total
	30-40	40-50	50-60	60-70	
30-40	3	1	1	-	5
40-50	2	6	1	2	11
50-60	1	2	2	1	6
60-70	-	1	1	1	3
Total	6	10	5	4	25

Required:

Calculate r between marks obtained in the above two subjects.

24. From the following data calculate the rank correlation coefficient 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

25. Calculate the rank coefficient of correlation of the following data:

X	80	78	75	75	68	67	60	59
Y	12	13	14	14	14	16	15	17

26. Twelve pictures submitted in a competition were ranked by two judges with results as shown below:

Pictures	1	2	3	4	5	6	7	8	9	10	11	12
Rank by 1st Judge	5	9	6	7	1	3	4	12	2	11	11	8
Rank by 2nd Judge	5	8	9	11	3	1	2	10	4	12	7	6

Required: Calculate the rank correlation coefficient.

27. Calculate the coefficient of rank correlation from the following data:

S. No. of student	Marks in Statistics	Marks in Accountancy
1	30	50
2	50	60
3	25	30
4	30	40
5	60	70
6	70	50
7	80	90
8	75	40
10	85	80

28. Calculate the coefficient of correlation from the data given below by the method of rank differences.

X	78	89	97	69	59	79	68	57
Y	125	137	156	112	107	136	123	108

29. Calculate 'r' by concurrent deviations from the following table:

Year	X	Y
1	368	22
2	384	21
3	385	24
4	361	20
5	347	22
6	384	26
7	395	26
8	403	29
9	400	28
10	385	27

30. Calculate the coefficient of correlation between X and Y series from the following data:

	Series	
	X	Y
No. of pairs of observations	15	15
Arithmetic mean	25	18
Standard deviation	3.01	3.03
Sum of squares of deviations from mean	136	138
Summation of product deviations of X and Y series from their respective arithmetic means = 122.		

31. The table below shows a firm's ranking of eight workers for performance and leadership potential. Calculate the coefficient of rank correlation between the two characteristics:

Performance Ranking	Leadership Potential Ranking
1	3
2	5
3	1
4	6
5	2
6	8
7	4
8	7

32. Find Karl-Pearson's coefficient of correlation between ages and playing habits of the following students:

Age (in yrs):	15	16	17	18	19	20
No. of Student:	250	200	150	120	100	80
Regular Players:	200	150	90	48	30	12

REGRESSION ANALYSIS



Up until now, we have delved into the idea of a statistical relationship between two variables, including sales and advertising budgets, product prices and supply, and many more. Although it shows the strength and direction of the correlation, the relationship between these variables does not resolve the following issues: Is there an algebraic or functional connection between these two variables? If that is the case, is it possible to predict the likely value of one variable from the other using this relationship? Which statistical method uses a question format to convey the connection between multiple variables? One way to estimate a variable's value is using regression analysis, which uses the value of another variable as a starting point. When applied to data sets, it facilitates pattern recognition, prediction, and identification of important impacting elements. In the domains of business, economics, finance, and the social sciences, among others, regression analysis offers useful insights for decision-making by examining these correlations. To help you evaluate and understand data efficiently, this unit will study the basic ideas, techniques, and applications of regression analysis.

Lesson 1: Introduction to Regression Analysis

Lesson Objectives:

After completing this lesson, you will be able to

- Origin of the concept of regression analysis;
- Define regression analysis;
- Understand the concept of regression line;
- Significance of the Study of regression analysis;
- Explain the significance of regression analysis;
- Difference between correlation and regression;
- Explain the obstacles and constraints of regression analysis.

Introduction

A basic statistical tool used to explore the relationship between variables is regression analysis. It assists in knowing how the dependent variable (outcome) varies in reaction to one or more independent variables (predictors). Widely used in many sectors including business, economics, finance, healthcare, and social sciences, this approach helps to forecast, trend analysis, and decision support.

Origin of the Concept of Regression Analysis

The concept of regression analysis originated in the late 19th century with the work of **Sir Francis Galton**, an English scientist and statistician. Galton was studying heredity and observed a fascinating pattern in the heights of parents and their children. He noticed that tall parents tended to have children who were shorter than them, while short parents had children who were taller. This tendency for extreme traits to move closer to the average in subsequent generations led him to coin the term "**regression to the mean.**"

The concept of regression analysis originated in the late 19th century with the work of Sir Francis Galton, an English scientist and statistician.

The work on Regression analysis was pioneered by Sir Francis Galton towards the end of nineteenth century. He used the word 'regression' for the first time while studying the relationship between the height of about one thousand fathers and sons. His study finally revealed two interesting results. They are:

- (i) Tall fathers tend to have tall sons and short fathers' short sons and
- (ii) The average height of the sons of a group of tall fathers is less than that of the fathers and the average height of the sons of a group of short fathers is greater than that of the fathers.

Galton's observations were later expanded upon by **Karl Pearson**, who formalized the mathematical framework for correlation and regression. Pearson's work helped in quantifying the strength and direction of relationships between variables, making regression analysis a valuable tool in statistics.

The development of regression analysis was further refined by Ronald A. Fisher in the early 20th century, who introduced the least squares method to estimate regression coefficients.

The development of regression analysis was further refined by **Ronald A. Fisher** in the early 20th century, who introduced the **least squares method** to estimate regression coefficients. This method provided a

systematic way to model relationships between variables, laying the foundation for modern regression techniques used today in various fields, including economics, biology, engineering, and artificial intelligence.

Thus, the concept of regression analysis originated from studies in heredity but has since evolved into a fundamental statistical technique for analyzing and predicting data effectively.

Definition of Regression Analysis

Regression analysis is a statistical method used to examine the relationship between two or more variables.

Regression analysis is a statistical method used to examine the relationship between two or more variables. It aims to model and analyze the relationships between a dependent variable (also known as the outcome or target variable) and one or more independent variables (also known as predictor or explanatory variables). The purpose is to understand how the dependent variable changes when one or more of the independent variables are altered, and to create a predictive model for future observations.

Regression analysis aims to model and analyze the relationships between a dependent variable (also known as the outcome or target variable) and one or more independent variables (also known as predictor or explanatory variables).

The main goal of regression analysis is to:

1. **Predict** the value of the dependent variable based on known values of the independent variables.
2. **Identify** the strength and nature of the relationship between variables.

Regression analysis, a cornerstone in statistical modeling, has been defined and refined by numerous scholars over the years. While earlier discussions have touched upon its general purpose and applications, a chronological perspective on its definitions offers deeper insight into its evolution.

Thus let us take an example: Regression permits answers to such questions as:

Does the growth rate influence a country's birth rate?

- (i) If the growth rate increases, by how much might a country's birth rate be expected to fall?
- (ii) Are other variables important in determining the birth rate?

In this example we assert that the direction of causality is from the growth rate (X) to the birth rate (Y) and not vice versa. The growth rate is therefore the **explanatory** variable (also referred to as the **independent** or **exogenous** variable) and the birth rate is the **dependent** variable (also called the **explained** or **endogenous** variable).

Regression analysis describes this causal relationship by fitting a straight line drawn through the data, which best summarizes them. It is sometimes called 'the line of best fit'.

Regression analysis describes this causal relationship by fitting a straight line drawn through the data, which best summarizes them. It is sometimes called '**the line of best fit**' for this reason.

Note that (by convention) the explanatory variable is placed on the horizontal axis, the explained on the vertical. This regression line is downward sloping (its derivation will be explained shortly) for the same reason that the correlation coefficient is negative, i.e. high values of Y are generally associated with low values of X and vice versa.

Since the regression line summarizes knowledge of the relationship between X and Y , it can be used to predict the value of Y given any particular value of X .

Here are the definitions of **regression analysis** as provided by different authors and experts in the field of statistics and research:

Wooldridge (2016):

"Regression analysis is a set of statistical techniques used to estimate the relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to model this relationship and make predictions or test hypotheses about the variables."

Montgomery, Peck, and Vining (2012):

"Regression analysis is the process of fitting a mathematical model to data in order to quantify the relationship between a response variable and one or more explanatory variables. The objective is to predict or explain the variation in the response variable."

Gujarati and Porter (2009):

"Regression analysis is a method for estimating the relationships among variables. It is used to model the relationship between a dependent variable and one or more independent variables, with the aim of making predictions or inferences about the variables."

Tabachnick and Fidell (2007):

"Regression analysis involves identifying the nature of the relationship between variables and providing a model that can be used to predict or explain outcomes based on known input variables."

The definitions differ in the level of detail and the focus (e.g., emphasis on prediction, estimation, or hypothesis testing), but they all point to regression analysis as a key tool for understanding and modeling relationships between variables.

Regression lines

In unit six the subject of correlation was discussed in details. It was there shown that the points giving the heights and weights of students, when plotted on a graph paper, form a path and cluster along a straight line and this line was described as a line of "**best fit**".

It can be stated again that in preceding unit we have noticed that where correlation is present, we can apply a line ($Y=a+bX$) which best fits the distribution of the points plotted on the graph paper. Few, if any, points would lie on it, but the line will go through the cluster of points so that there are an equal number of points on each side of it. This line is known as the line of regression. When correlation is perfect positive, then this line would make an angle of 45° with the positive X -axis. If on the other hand, there was perfect negative correlation, the line of regression would be at right angles to that representing perfect positive correlation. In social sciences we usually do not find perfect correlation but partial correlation.

When the correlation is linear the regression equation is of the form $Y=a+bX$. This equation gives the relationship between two variables X and Y. Thus, in the case of heights and weights of students, given the height of student, it is not possible to say what his weight would be, but is possible to estimate by giving the average 'Weight' of students who are of a given height. The height could be represented by X. It would be necessary to find the value of a and b . Then for any given height; the average weight of students of that height could be calculated. In any particular case the actual weight of a student might be different from the computed value. The sum of the deviations between actual and computed weights would be "nil". This is what is meant by being a "**line of best fit**".

Let us take an example. Suppose the figures of rainfall and agricultural produce in a region are given for a period of four years. The rainfall is 22" which occurs four times. The agricultural produce has been 42, 39, 38 and 41 units in different years .. The average produce, *i.e.*, 40 units will be taken as the figure corresponding to the 22" rainfall. The series of corresponding values thus formed may be plotted on a graph taking the rainfall as X and agricultural produce as Y. The line representing the points in the best way will be recognized as the regression line giving regression of Y on X. Symbolically, $Y=a+bX$.

On other hand, particular amount of produce may be taken as Y and the means of the figures of rainfall corresponding to the amount may be calculated and plotted as X. The line representing the points in this case will give the regression of X on Y. Symbolically $X=a+bY$. The two Lines thus depicting the relationship between X and Y series will be distinct being formed in different ways. As mentioned above, if there is perfect correlation, there shall be only one regression line. If the variates are independent *i.e.*, $r = 0$, then the lines of regression are at right. Angles, *i.e.*, parallel to OX and OY. But while dealing with business and economic data, we come across perfect correlation rarely. **Thus, very often there are two regression lines:**

There are two regression lines:

(i) $X=a+bX \rightarrow$ the regression of Y on X

(ii) $Y=a+bY \rightarrow$ the regression of X on Y

(i) $X=a+bX \rightarrow$ the regression of Y on X

(ii) $Y=a+bY \rightarrow$ the regression of X on Y

The statistical method with the help of which it is possible to estimate (or predict) the unknown values of one variable from another variables is known as regression.

The line describing the tendency in the (ii), *i.e.*, to regress or going back was called by him a 'Regression Line'. In fact, the dictionary meaning of the term 'regression' is the act of returning or going back. The term is still used by many statisticians to describe the line drawn for a group of points to represent the trend s own, but it no longer necessarily carries the original implication that Galton had in mind. The modern statisticians prefer to use the term 'estimating line' instead of 'regression line'.

The modern statisticians prefer to use the term 'estimating line' instead of 'regression line'.

Significance of the Study of Regression Analysis

In the field of business and economics, it is sometimes essential to create projections. Once the connection between price (X) and demand (Y) of a certain item is determined, one may predict the probable value of X for a

given value of Y, or the probable value of Y for a given value of X. Thus, by regression analysis, we may determine the anticipated average change in one variable corresponding to a defined change in another. Regression principles can be utilized in the scientific, physical, and social sciences. Regression analysis is vital in various fields, providing critical insights and techniques for understanding relationships between variables. The regression technique can be extended to include three or more variables; however, this work will concentrate exclusively on two variables, specifically simple regression.

Regression analysis is vital in various fields, providing critical insights and techniques for understanding relationships between variables.

Here are few essential reasons why the study of regression analysis is imperative:

1. Comprehending Interconnections among Variables:

Regression elucidates the relationships among several variables. It aids in assessing whether alterations to an independent variable, such as advertising expenditure, influence a dependent variable, such as sales income.

Relationship Quantification: It provides a method for assessing the direction and magnitude of relationships, often expressed using coefficients. The slope of a fundamental linear regression signifies the extent to which the dependent variable alters when the independent variable rises by one unit.

2. Prognostication and Anticipation:

Predict Future Outcomes: Regression models are commonly utilized to project future values based on historical data. This prediction skill is essential in fields such as economics (estimating GDP growth), marketing (projecting sales), and medicine (predicting patient outcomes).

Risk management: Regression analysis facilitates risk forecasting and portfolio administration in the insurance and financial sectors, allowing organizations to prepare for unexpected events.

3. Decision-Making and Policy Development:

Informed Decisions: Regression analysis enables decision-makers to make more data-driven judgments. Businesses can build marketing strategy by analyzing the correlation between expenditure and revenue.

Informing Public Policy: In economics and social sciences, regression analysis assists policy makers in comprehending the impacts of many variables, including income, education, and public policies, on societal outcomes.

4. Simulating Complex Interconnections:

Multivariable Analysis: In the context of real-world data, regression analysis is particularly advantageous as it allows for the formulation of complex interactions among multiple independent variables. Numerous criteria, such as location, size, and number of bedrooms, influence the determination of property pricing.

Nonlinear Relationships: Complex regression techniques, such as logistic or polynomial regression, can be employed to yield accurate models in scenarios where relationships are nonlinear.

5. Enhancing Research and Development:

Data-Driven Insights: Regression is an essential instrument in scientific and industrial research. It assists researchers in identifying causal linkages, testing hypotheses, and validating experimental outcomes.

Optimization: In engineering, physics, and economics, regression analysis facilitates the optimization of processes and systems. In manufacturing, regression can optimize production efficiency by analyzing various input components.

6. Identification of Anomalies and Errors:

Error Detection: Regression assists in detecting outliers, errors, and anomalies within the data. Analyzing residuals (the discrepancies between observed and predicted values) allows for the evaluation of the model's validity and its adequacy in fitting the data.

Model Enhancement: By pinpointing outliers that poorly align with the regression model, researchers and analysts can enhance their models or explore the reasons for these deviations.

7. Enhancing Predictive Accuracy:

Enhanced Decision Making: Regression analysis increases prediction accuracy by considering the statistical importance of predictors, mitigating bias, and contributing to the reliability of forecasts in business, healthcare, and economics.

Alternative Models: To enhance forecast precision, especially in dynamic fields such as finance or meteorology, regression analysis may be integrated with additional methodologies like time series analysis.

8. Establishing Causal Relationships:

Causal Inference (with Caution): Although correlation does not equate to causation, regression analysis may assist in identifying possible causal correlations. By accounting for confounding variables, regression enables researchers to draw more robust causal inferences.

Policy and Interventions: In domains such as healthcare or education, comprehending causal links is essential for formulating successful interventions. Comprehending the influence of nutrition and exercise on health outcomes helps inform public health strategies.

9. Versatility Across Disciplines:

Multidisciplinary Application: Regression analysis is extensively employed across various fields, including economics, finance, psychology, medicine, engineering, and social sciences. Its adaptability to various data formats renders it an indispensable instrument for researchers and practitioners.

10. Improving Model Generalization:

Generalization to Novel Data: By accurately modeling the interrelationship among variables, regression facilitates the creation of models that may effectively generalize to new, unobserved data. This is essential for guaranteeing that models can generate precise predictions beyond the training dataset.

Summary

Regression analysis is crucial as it facilitates the understanding, prediction, and improvement of relationships among variables. It is an essential element of data analysis and decision-making due to its extensive applicability, spanning business, policy-making, and scientific research. Regression analysis provides a systematic, quantitative methodology across various domains to derive insights and improve outcomes, applicable in forecasting, optimization, or hypothesis testing.

Regression analysis is crucial as it facilitates the understanding, prediction, and improvement of relationships among variables.

Difference between Correlation and Regression

As we saw in the previous unit, correlation is a way to measure how closely related two variables are. But regression will not get the job done. This sentence provides a clear explanation of the link between the two variables: the average predicted variation in one variable as a result of a specified change in the other. Additionally, as mentioned in the prior unit, correlation does not help in identifying which variable is the cause and which is the consequence. The relationship between the two variables can be more easily examined by regression. When doing regression analysis, it is common practice to label one variable as dependent and the other as independent. Below, we will go into the differences between regression analysis and correlation analysis:

Differences between Correlation and Regression Analysis

Feature	Correlation Analysis	Regression Analysis
Definition	Measures the strength and direction of the relationship between two variables.	Establishes a cause-and-effect relationship and predicts the dependent variable based on independent variables.
Purpose	Determines if two variables are related and how strong their relationship is.	Identifies how one variable influences another and makes predictions.
Dependency	Both variables are treated equally, with no assumption of dependency.	One variable (dependent) is influenced by the other (independent).
Mathematical Representation	Expressed using the correlation coefficient (r), which ranges from -1 to +1.	Represented by a regression equation (e.g., $y = a + bX$).
Prediction Capability	Cannot predict values; only shows the degree of association.	Can predict the dependent variable based on independent variable(s).
Causation	Does not imply causation; only measures association.	Implies a causal relationship, showing how one variable affects another.
Graphical Representation	Scatter plot showing the pattern of data points.	Scatter plot with a fitted regression line (trend line).
Types	Pearson's correlation, Spearman's rank correlation, etc.	Simple linear regression, multiple regression, multiple regression, etc.
Outliers Effect	Highly sensitive to outliers, which can distort the correlation coefficient.	Less sensitive to outliers, but extreme values can still affect regression results.

Obstacles and Constraints of Regression Analysis

Regression analysis is a powerful method for elucidating correlations between variables and forecasting outcomes; nonetheless, it presents certain challenges and limitations. Here are many of the significant ones:

1. Breaches of Assumptions:

Regression analysis depends on specific fundamental assumptions, and when these assumptions are breached, the outcomes may become unreliable:

Linearity: Presumes a linear correlation between the independent and dependent variables. If the relationship is nonlinear, the model may inadequately represent the actual pattern.

Independence: Presumes that the residuals (errors) are mutually exclusive. This assumption may be contravened using time series or clustered data.

Homoscedasticity: Presumes that the variance of residuals remains uniform across all levels of the independent variable. Violation of this principle, especially heteroscedasticity, may result in inefficient calculations and compromise hypothesis testing.

Normality of Errors: Presumes that the residuals conform to a normal distribution. Non-normally distributed errors can impact statistical inference.

2. The existence of multicollinearity:

Multicollinearity occurs when two or more independent variables exhibit a significant correlation. Consequently, evaluating the individual impact of each predictor on the dependent variable is difficult. Inflated standard errors can result in erroneous estimates and diminish the efficacy of statistical tests.

3. Overfitting:

An overly complex regression model overfits by seizing data noise rather than the fundamental link. An overfitted model has exceptional performance on training data but poor performance on new data, indicating insufficient generalization. This may occur if the model possesses an excessive number of variables or exhibits exceptional flexibility, as seen with high-degree polynomials.

4. Insufficient Fitting:

Underfitting transpires when the model is excessively simplistic to accurately represent the intrinsic relationship between variables. This may occur when essential predictors are excluded, or the selected model is excessively inflexible, resulting in inadequate forecasts and diminished explanatory efficacy.

5. Issues Pertaining to Data Integrity:

Anomalies: Outliers or extreme values may lead to erroneous or deceptive estimates by disproportionately influencing the regression outcomes.

Data Deficiency: The model's precision may be compromised by absent values. Inadequate handling of absent data, such as mean imputation, might introduce bias and distort the findings.

Measurement error: The regression outcomes will be adversely affected if there is an error in measuring either the independent or dependent variables.

6. Correlation and Causation:

While it cannot establish causation, regression analysis can illustrate the relationship between variables. Two variables are not inherently causative of one another solely due to their association. Regression analysis cannot definitively establish causal relationships without a well-defined causal framework or suitable experimental design.

7. Constrained Predictive Capacity in Complex Systems:

In intricate systems with several interacting variables, a basic regression model may inadequately represent the entirety of the interactions. In disciplines such as biology, economics, or social sciences, where correlations may be affected by non-linear dynamics or latent variables, the regression model may lack the necessary robustness for precise predictions.

8. Interpretability and Complexity of Models:

The model may get increasingly complex and difficult to comprehend as the number of variables increases. The model may grow excessively intricate, hindering the extraction of essential information in high-dimensional contexts, such as those with several potential predictors, including genomic data or extensive marketing research.

9. Challenges of Extrapolation:

Regression models can accurately predict within the confines of observable data; however, their effectiveness significantly diminishes when tasked with extrapolation beyond that scope. Utilizing a model derived from data of one age group to forecast outcomes for a completely other age cohort may produce inaccurate findings.

10. Complexity of Multivariate Regression:

The examination of the relationship between predictors and the dependent variable might become complex when employing multivariate regression with several independent variables. The interplay among predictors may further obfuscate the understanding of the coefficients.

Self-Assessment Questions:

Short Questions

1. Define regression?
2. Discuss the uses and significance of regression analysis.
3. What is the regression line?
4. Discuss the differences between correlation and regression.
5. What do you understand by the term "line of best fit"?
6. State the meaning of regression lines X on Y and Y on X.
7. Discuss the obstacles and constraints of regression analysis.

Multiple Choice Questions:

1. **Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:**

- (i) Regression analysis can be described as _____.
 - a) a statistical hypothesis test in which the test statistic follows a student's t -distribution if the null hypothesis is supported
 - b) a collection of statistical models in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation
 - c) a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true
 - d) a tool for building statistical models that characterize relationships among a dependent variable and one or more independent variables, all of which are numerical
- (ii) Which of the following is a limitation of regression analysis?
 - a) It cannot be used for non-linear relationships
 - b) It requires a large sample size
 - c) It assumes no multicollinearity
 - d) It assumes that the dependent variable is categorical
- (iii) An assumption of regression analysis is homoscedasticity, which states that the
 - a) variation of the dependent variable is the same across all values for the independent variable.
 - b) residuals exhibit no patterns across values for the independent variable.
 - c) residuals exhibit no patterns across values for the dependent variable.
 - d) relationship between the independent and dependent variables is linear.
- (iv) A prediction interval for the independent variable X would specify _____.
 - a) all the possible values of the dependent variable Y
 - b) the probability distribution for the various values of X
 - c) the uncertainty in the dependent variable for a single value of X
 - d) all the possible values of X

- (v) Which of the following is not an assumption made in regression analysis?
- a) The parameters of the regression equation are linear.
 - b) The values of successive observations are correlated.
 - c) The errors for each individual value of x are normally distributed.
 - d) Variation about the regression line is constant for all values of the independent variable.

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) An assumption of regression analysis is that the relationship between the independent and dependent variables is linear.
- (ii) A regression line represents the best fit line for a set of data points in a scatter plot.
- (iii) Regression analysis is used to build statistical models that characterize relationships among a dependent variable and one or more independent variables.
- (iv) Using the regression equation to predict values for the dependent variable beyond the range of the data may provide results that are unreliable.
- (v) The slope of the regression line tells us the change in the dependent variable for a unit change in the independent variable.

Answer:

Multiple-Choice Question:

1. (i) d (ii) a (iii) d (iv) c (v) b

True/False

2. (i)- T (ii)- T (iii)- T (iv)- T (v)- T

Lesson 2: Methods of Obtaining Regression Lines

Lesson Objectives:

After completing this lesson, you will be able to

- Understand the method of obtaining regression lines;
- Define the scatter diagram method and understand its nature and interpretations;
- Understand the method of least square and its uses;
- Merits and limitations of least squares regression method ;
- Define and interpret co-efficient of determination (R^2);
- Define and interpret the standard error of estimate;
- Meaning and use of regression coefficient;
- Relationship between coefficient of correlation and regression coefficients.

Introduction

In the previous unit we developed measures to express the strength and the direction of the relationship between two variables. In this lesson we wish to develop an equation to express the linear (straight line) relationship between two variables. In addition, we want to be able to estimate the value of the dependent variable Y based on a selected value of the independent variable X. The technique used to develop the equation and provide the estimates is called regression analysis.

As the unit title indicates, the study of bivariate analysis is typically thought of as two separate but closely related procedures. In simple words, Regression analysis is a technique for making estimates or predictions. The primary objective of regression analysis is the development of a regression lines to explain the association between two or more variables in the given population. A regression line summarizes the relationship between two variables, but only in a specific setting: one of the variables helps explain or predict the other. That is, regression describes a relationship between an explanatory variable and a response variable.

Regression describes a relationship between an explanatory variable and a response variable.

The first step in any regression problem is to choose the form of the functional model. The appropriate form depends on the type of relationship which exists between the variables.

Methods of Obtaining Regression Lines

There are two methods by which regression lines can be obtained. They are:

1. The Scatter Diagram Method
2. The Method of Least Square.

1. The Scatter Diagram Method: It is possible to find out the actual relationship between two variables with the help of a scatter diagram. In the bivariate case, the type of relationship which exists can frequently be perceived from a scatter diagram.

Scatter diagrams for regression analysis are constructed with the independent variable is taken on horizontal axis (*i.e.*, X-axis) and the

dependent variable on the vertical axis (i.e., Y-axis). Then the paired observations are plotted on the graph paper. The scatter diagram enables us to observe the data graphically and to draw preliminary conclusions about the possible relationship between the variables.

If the points form a straight path giving straight line then there is perfect correlation and the values of one variable can be estimated given the value for the other.

If the points form a straight path giving straight line then there is perfect correlation and the values of one variable can be estimated given the value for the other. But as mentioned earlier, in economic and business problems, perfect correlation is a rarity and so the problem is to draw line on graph in such a way that the dots are best represented by it.

This line is to be drawn by inspection and care must be taken to draw it in such a way as to be the best fit. One must bear in mind the following points while drawing this line:

1. The line should be as close as possible to all the points on the graph.
2. Almost an equal number of points should be there on either side of the line.
3. An attempt should be made to draw the line in such a way that the points on its either side are equidistant from it.

Thus the preparation of a scatter diagram is the first step in the solution of a bivariate problem because it enables the analyst to choose the correct form of the regression line and make it quite clear that the regression lines cannot be applied blindly to any bivariate data.

Let us take an example.

Example 7.1 Given the following pairs of values of variables X and Y.

x	2	3	5	6	8	9
y	6	5	7	8	12	11

Required:

By graphic inspection draw an estimating line.

Solution:

Taking X on the X-axis and Y on the Y-axis, pairs of observations are plotted on a graph paper (Fig. 7.1). Then a free hand estimating line is drawn in such a manner that the sum of the positive and negative deviations on either side of the line is zero.

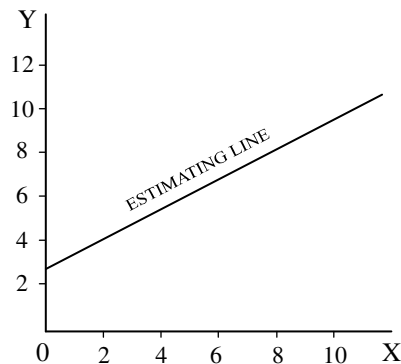


Fig 7.1

Merits and Limitations of the Method

This method is very simple and easy. It does not take much time to draw the estimating line. As it is drawn freehand, different persons may draw different lines for the same data. With practice it is not difficult to draw an approximately correct line.

2. Method of Least Squares: The method of least squares regression was first formalized by **Adrien-Marie Legendre** in 1805, in his work on the **adjustment of observations**. Legendre was a French mathematician and astronomer, and his work focused on fitting data to a mathematical model in the context of astronomy and surveying. He developed the least squares method to minimize the sum of the squared differences (errors) between the observed values and the values predicted by a model. The goal was to find a line or curve that best fits a set of data points by minimizing the error.

The least squares method to minimize the sum of the squared differences (errors) between the observed values and the values predicted by a model.

The least squares method is a procedure for using sample data to find the estimated regression equation. We want to find good estimates of the regression properties of good estimators. Unbiasedness and efficiency are among these properties. A method that will give us good estimates of the regression coefficients is the method of least squares. The method of least squares gives us the best linear unbiased estimators (BLUE) of the regression parameters. These estimators both are unbiased and have the lowest variance of all possible unbiased estimators of the regression parameters.

The method of least squares gives us the best linear unbiased estimators (BLUE) of the regression parameters.

So, the Least Squares Regression is a statistical technique used to estimate the relationship between one or more independent variables (predictors) and a dependent variable (response) by minimizing the sum of the squared differences between the observed values and the predicted values generated by the model.

The method seeks to find the best-fitting line (or curve) in a way that the sum of the squared residuals (the differences between the actual data points and the predicted values) is as small as possible. This approach is widely used in linear regression but can also be applied to more complex models, such as nonlinear regression.

In simple terms, Least Squares Regression finds the line (or curve) that "best fits" the data by reducing the discrepancies between the observed data and the predicted outcomes.

The drawing of a straight line with the help of normal equations was described there.

The Regression of Y on X:

We have considered a case of "given X, what is the value of Y? ", where X is independent. This is also known the regression of Y on X. The slope of the line, b , in the equation is known as the regression coefficient. It shows that Y changes b times as fast as X. Symbolically the regression coefficient of Y on X is b_{yx} .

Least Squares Regression finds the line (or curve) that "best fits" the data by reducing the discrepancies between the observed data and the predicted outcomes.

The Regression of X on Y:

Similarly, we can make Y independent, given the weight of a student; what is his height. The regression equation *in* this case is: $X = a + bY$ and the regression coefficient of X on Y is b_{xy} .

The following example would illustrate the two regression equations

Example 7.2. Given the bivariate data:

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	5

Required:

- (a) Fit the regression line of Y on X and hence predict Y if X=10
 (b) Fit the regression line of X on Y and hence predict X, if Y=2.5

Solution: Both the regression lines are required. It is necessary to find a and b . Then given X, Y can be estimated and *vice versa*. The calculations required are given in following Table no 1:

TABLE No: 1
Computation of Regression Equations

X	Y	X^2	Y^2	XY
1	6	1	36	6
5	1	25	1	5
3	0	9	0	0
2	0	4	0	0
1	1	1	1	1
1	2	1	4	2
7	1	49	1	7
3	5	9	25	15
$\Sigma X = 23$	$\Sigma Y = 16$	$\Sigma X^2 = 99$	$\Sigma Y^2 = 68$	$\Sigma XY = 36$

It will be observed that these are precisely the same values which must be computed to find the coefficient of correlation. As it is necessary to find the values of a and b , therefore, two simultaneous equations are to be solved:

$$\Sigma Y = Na + b\Sigma X \dots\dots\dots (i)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \dots\dots\dots (ii)$$

Substituting the values, we get:

$$16 = 8a + 23b \dots\dots\dots (i)$$

$$36 = 23a + 99b \dots\dots\dots (ii)$$

Multiplying (i) by 23 and (ii) by 8, we get

$$368 = 184a + 529b \dots\dots\dots (iii)$$

$$288 = 184a + 792b \dots\dots\dots (iv)$$

Subtracting (iv) from (iii), we get

$$263b = -80$$

$$b = -0.304 = -0.30 \text{ (approx.)}$$

Substituting $b=-0.30$ in equation (i), we get

$$\begin{aligned} 16 &= 8a + (23)(-0.30) = 8a - 6.9 \\ -8a &= -16 - 6.9 \\ a &= 2.86 \end{aligned}$$

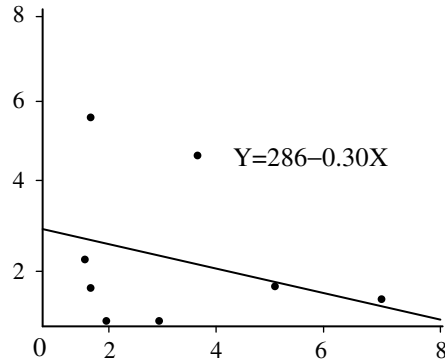


Fig.7.2. Regression of Y on X

Thus, regression of Y on X is

$$Y = 2.86 - 0.30X \text{ (see Fig. 7.2)}$$

Similarly, the second regression equation can be solved with the help of the two simultaneous equations. Thus the regression equation of X on Y is :

$$X = a + bY$$

The two simultaneous equations are:

$$\sum X = Na + b\sum Y \quad \dots\dots\dots(i)$$

$$\sum XY = a\sum Y + b\sum Y^2 \quad \dots\dots\dots(ii)$$

Substituting the values in the equations, we get:

$$23 = 8a + 16b \quad \dots\dots\dots(i)$$

$$36 = 16a + 68b \quad \dots\dots\dots(ii)$$

Multiplying equation (i) by (2),

$$46 = 16a + 32b \quad \dots\dots\dots(iii)$$

Now deducting (iii) from (ii), we get :

$$36b = -10$$

$$b = -0.28$$

Substituting $b = -0.28$ in equation (i), we get

$$23 = 8a + 16(-0.28)$$

$$23 = 8a - 4.48$$

$$a = 3.43$$

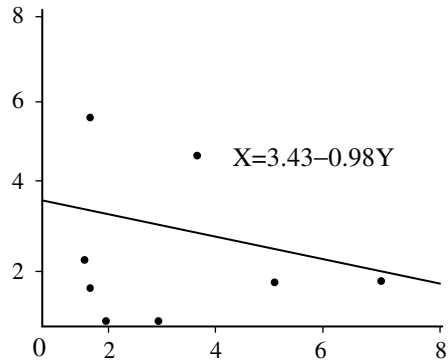


Fig.7.3. Regression of X on Y

The regression of X on Y is

$X = 3.43 - 0.28 Y$ (see Fig.7.3)

The points of two regression lines are arrived in Tables II and III.

TABLE II
 $Y = 2.86 - 0.30 X$

x	Y (Actual)	Y_c Computed from equation	$Y - Y_c$ Deviation
1	6	2.56	+3.44
5	1	1.36	-0.36
3	0	1.96	-1.96
2	0	2.26	-2.26
1	1	2.56	-1.56
1	2	2.56	-0.56
7	1	0.76	+0.24
3	5	1.96	+3.04
			+0.02

TABLE III
 $X = 3.43 - 0.28 Y$

Y	X (Actual)	X_c Computed from equation	$X - X_c$ Deviation
6	1	1.75	-0.75
1	5	3.15	+1.85
0	3	3.43	-.043
0	2	3.43	-1.43
1	1	3.15	-2.15
2	1	2.87	-1.87
1	7	3.15	+3.85
5	3	2.03	+0.97
			+0.04

Alternative formulae for determining the values of a and b :

Alternative formulae for computing a and b have been developed: They are:

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{N \sum X^2 - (\sum X)^2}$$

$$b = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2}$$

These formulae will be applied to find out the values of a and b in earlier Example 7.2.

(1) Regression of X on Y :

The values of constants a and b can be found out with the help of the following formulae:

$$a = \frac{\sum Y^2 \sum X - \sum Y \sum XY}{N \sum Y^2 - (\sum Y)^2}$$

$$b = \frac{N \sum XY - \sum X \sum Y}{N \sum Y^2 - (\sum Y)^2}$$

Substituting the values, we get :

$$a = \frac{68(23) - (16)(36)}{8(68) - (16)^2}$$

$$= \frac{1564 - 576}{544 - 256}$$

$$= \frac{988}{288} = 3.43.$$

$$b = \frac{8(36) - (23)(16)}{8(68) - (16)^2}$$

$$= \frac{288 - 368}{288} = -0.28.$$

Thus the values of a and b are the same as above.

(2) Regression of Y on X :

$$a = \frac{\sum X^2 \sum Y - \sum X \sum XY}{N \sum X^2 - (\sum X)^2}$$

$$b = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2}$$

Substituting the values, we get

$$a = \frac{99(16) - (23)(36)}{8(99) - (23)^2}$$

$$= \frac{1584 - 828}{792 - 529}$$

$$= \frac{756}{263} = 2.86$$

$$b = \frac{8(36) - (23)(16)}{8(99) - (23)^2}$$

$$= \frac{288 - 368}{263} = -0.30.$$

Again we notice that results are the same as we obtained by the first method.

Deviations taken from Actual Arithmetic Means of X and Y:

In the above example we have found out the regression lines directly from the actual data. If we deal with the deviations of X and Y variables from their respective means, calculations can be simplified. Then the equation $Yc=a+bX$ is transformed into

$$Y - \bar{Y} = b(X - \bar{X})$$

Taking $y = (Y - \bar{Y})$ and $x = (X - \bar{X})$

we get; $y = bx$

We are already familiar with the two normal equations, i.e.,

$$\sum Y = Na + b \sum X \quad \dots\dots\dots (i)$$

$$\sum XY = a \sum X + b \sum X^2 \quad \dots\dots\dots (ii)$$

Transforming them in terms of x and y, we get:

$$\sum y = Na + b \sum x \quad \dots\dots\dots (i)$$

$$\sum xy = a \sum x + b \sum x^2 \quad \dots\dots\dots (ii)$$

Now, as we are taking deviations from the actual arithmetic means, therefore $\sum x = \sum y = 0$.

$\therefore$ Equation (i) reduces to:

$$Na = 0, \quad \text{or } a = 0$$

Equation (ii) reduces to

$$\sum xy = b \sum x^2$$

$$b = \frac{\sum xy}{\sum x^2}$$

Thus, the regression equation of Y on X can be written as:

$$y = \frac{\sum xy}{\sum x^2} \cdot x$$

Similarly, the regression equation $Xc=a+bY$ is reduced to $x=by$ where the value of b can be obtained as follows:

$$b = \frac{\sum xy}{\sum y^2}$$

Example 7.3. Given the bivariate data:

X:	1	5	3	2	1	2	7	3
Y:	6	1	0	0	1	2	1	5

Required:

Find regression equations by taking deviations of items from the means of X and Y respectively

Solution:

Regression: Equation of Y on X, i.e., $y = bx$; $b = \frac{\sum xy}{\sum x^2}$

$$\bar{X} = \frac{\sum x}{N} = \frac{24}{8} = 3; \quad \bar{Y} = \frac{\sum y}{N} = \frac{16}{8} = 2;$$

Substituting the values, we get, $y = \frac{-10}{30}x = -0.33x$

X	(X- $\bar{X}$) x		Y	(Y- $\bar{Y}$) y		xy
1	-2	4	6	4	16	-8
5	2	4	1	-1	1	-2
3	0	0	0	-2	4	0
2	-1	1	0	-2	4	2
1	-2	4	1	-1	1	2
2	-1	1	2	0	0	0
7	4	16	1	-1	1	-4
3	0	0	5	3	9	0
24		30	16		36	-10

But $Y = (Y-2)$ and $X = (X-3)$
 $Y-2 = -.033(X-3)$
 $Y-2 = -.033X + 0.99$
 $Y = -.033X + 2.99$
 $Y = 2.99 - .033X$

Regression Equation of X, on Y, i.e., $x = by$; $b = \frac{\sum xy}{\sum y^2}$

Substituting the values, we get, $x = \frac{-10}{36}$; $y = -.028y$

But $x = (X-3)$ and $y = (Y-2)$
 $(X-3) = -.028(Y-2)$
 $(X-3) = -.028Y + .56$
 $X = 3.56 - .028Y$

The limitation of this method is that it becomes difficult to employ in case when the mean is fractions. In such a situation deviations may be taken from the assumed means.

Deviations taken from Assumed Means

The regression equation in such a case would be as follows:

Regression equation of X on Y

$$X - \bar{X} = r^2 b_{xy} (Y - \bar{Y})$$

As mentioned earlier b_{xy} denotes regression coefficient of X on Y.

$$b_{xy} = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_y^2 - \frac{(\sum d_y)^2}{N}}$$

Regression equation of Y on X

$$Y - \bar{Y} = b_{yx} (X - \bar{X})$$

where b_{yx} denotes regression coefficient of Y on X.

$$b_{yx} = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_x^2 - \frac{(\sum d_x)^2}{N}}$$

It may be noted that in the last method, it was required to find out the value of b only. Under this method, the value of the regression coefficients is to be found before solving the regression equation.

Example 7.4. From the data in Example 14.3, obtain regression equations taking deviations from 4 in case of X and 3 in case of Y.

Solution:

X	(X-4) d_x	d_x^2	Y	(Y-3) d_y	d_y^2	$d_x d_y$
1	-3	9	6	3	9	-9
5	1	1	1	-2	4	-2
3	-1	1	0	-3	9	+3
2	-2	4	0	-3	9	+6
1	-2	9	1	-2	4	+6
2	-2	4	2	-1	1	+2
7	3	9	1	-2	4	-6
3	-1	1	5	2	4	-2
24	-8	38	16	-8	44	-2

Regression equation of Y or X :

$$\begin{aligned}
 X - \bar{X} &= b_{xy} (Y - \bar{Y}) \\
 b_{xy} &= \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_y^2 - \frac{(\sum d_y)^2}{N}} \\
 b_{xy} &= \frac{-2 - \frac{(-8)(-8)}{8}}{44 - \frac{(-8)^2}{8}} = \frac{-2 - 8}{44 - 8} \\
 b_{xy} &= \frac{10}{36} = -0.28 \\
 (X-3) &= -0.28(Y-2) \\
 X &= 3.56 - 0.28Y
 \end{aligned}$$

Regression equation of Y on X :

$$\begin{aligned}
 X - \bar{X} &= b_{yx} (X - \bar{X}) \\
 b_{yx} &= \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_x^2 - \frac{(\sum d_x)^2}{N}} \\
 &= \frac{-2 - \frac{(-8)(-8)}{8}}{38 - \frac{(-8)^2}{8}} = \frac{-10}{30} \\
 &= -0.33
 \end{aligned}$$

Thus, the regression equation becomes

$$\begin{aligned}
 Y - 2 &= -0.33 (X - 3) \\
 Y &= 2.99 - 0.33x
 \end{aligned}$$

Merits and Limitations of Least Squares Regression Method

The Least Squares Regression Method is a prevalent technique in statistical modeling, particularly for fitting a line or curve to data. The least squares method provides the best-fitting line through the data by minimizing the sum of the squares of the vertical deviations from each data point to the line. This ensures the most accurate linear relationship possible based on the available data. Below is an analysis of its merits and Limitations:

1. Determining the Best Fit Line:

By minimizing the sum of the squares of the vertical deviations between each data point and the line, the least squares method finds the line that best fits the data. This guarantees the most precise linear relationship conceivable given the data at hand.

2. A Mathematical and Objective Approach:

It determines the line of greatest fit using a precise mathematical technique that removes subjectivity. Strict computations form the basis of the procedure, guaranteeing reliable outcomes.

3. The Ability to Predict:

Based on known values of the independent variable, the regression equation can be used to predict the value of the dependent variable. This is particularly helpful for trend analysis and forecasting.

4. Broadly Relevant:

Anywhere there are linear correlations between variables, including in the social sciences, biology, engineering, economics, and finance, the least squares method can be used.

5. Straightforward and User-Friendly:

With statistical tools or even spreadsheets, the method for linear regression is rather straightforward and easy to compute. It is perfect for both academic and real-world applications because of its accessibility.

6. Aids in Relationship Understanding:

It facilitates better decision-making and strategy creation by helping to quantify the type and intensity of the link between two variables.

7. The Foundation of Statistics:

More intricate statistical models and methods are based on the least squares approach. It is fundamental to statistical inference, machine learning algorithms, and econometrics.

8. Multiple Regression Extension:

Analysis of more intricate associations in data is made possible by the method's easy extension to include several independent variables (multiple linear regression).

Limitations of the Least Squares Regression Method

1. Outlier Sensitivity:

Extreme values or outliers can greatly skew the regression line and produce false results because the approach is extremely sensitive to them.

2. Presupposes a Linear Connection:

In real-world situations, the straight-line (linear) relationship between the independent and dependent variables may not always hold true, as assumed by least squares regression.

3. Needs Numerical Information:

It is not possible to apply the approach directly to categorical data without first transforming it into a numerical format, which may not always be suitable or efficient.

4. The assumption of homoscedasticity:

It makes the assumption that all levels of the independent variable have the same variance in the errors. Inefficient estimates could result from a violation of this assumption (heteroscedasticity).

5. Assumes Errors Are Independent:

It is believed that the errors (residuals) are unrelated to one another. The results may be deceptive if autocorrelation is present, as it is with time series data.

6. Only Captures Association, Not Causation:

The method shows correlation between variables, but not necessarily causation. Misinterpreting correlation as causation can lead to incorrect conclusions.

7. Multicollinearity's influence:

When independent variables in multiple regression have a high degree of correlation with one another (multicollinearity), the model may become unstable and the interpretation of the coefficients may vary.

8. Restricted Application to Non-Linear Data:

The least squares linear regression approach is insufficient for datasets with non-linear variable relationships unless it is modified or substituted with a non-linear model.

Coefficient of determination

Coefficient of determination indicates the proportion of the variance in the dependent variable (or response variable) that is predictable from the independent variables (or predictors).

A widely used measure of fit for regression models is the coefficient of determination, or r^2 . In simple words, it indicates the proportion of the variance in the dependent variable (or response variable) that is predictable from the independent variables (or predictors). The coefficient of determination is the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).

The coefficient of determination ranges from 0 to 1. An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x. An r^2 of 1

means perfect prediction of y by x and that 100% of the variability of y is accounted for by x . Of course, most r^2 values are between the extremes. The researcher must interpret whether a particular r^2 is high or low, depending on the use of the model and the context within which the model was developed.

In exploratory research where the variables are less understood, low values of r^2 are likely to be more acceptable than they are in areas of research where the parameters are more developed and understood. A business researcher who is trying to develop a model to predict the motivation level of employees might be pleased to get an r^2 near .50 in the initial research.

It indicates the proportion of the variance in the dependent variable (or response variable) that is predictable from the independent variables (or predictors).

Interpretations of R^2 :

1. Range:

- R^2 ranges from 0 to 1 in most contexts.
- Sometimes, it can be negative for models that are worse than a horizontal line (e.g., when using models not forced through the origin or with no intercept).

R^2 ranges from 0 to 1 in most contexts.

2. Value Meaning:

- $R^2 = 0$: The model explains none of the variability of the response data.
- $R^2 = 1$: The model explains all the variability.
- $R^2 = 0.8$: 80% of the variance in the dependent variable is predictable from the independent variable(s).

3. High R^2 Doesn't Mean "Good":

- A high R^2 doesn't guarantee that the model is appropriate (it may be overfitting).
- It also doesn't confirm causation, only correlation.

4. Adjusted R^2 :

- Adjusted R^2 accounts for the number of predictors and only increases if the new term improves the model more than would be expected by chance.

Formula of R^2 :

There are a few equivalent formulas, depending on what data we have:

1. Based on Sum of Squares:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- SS_{res} : Residual Sum of Squares

$$\sum (y_i - \hat{y}_i)^2$$

- SS_{tot} : Total Sum of Squares

$$\sum (y_i - \bar{y})^2$$

Based on Correlation (for simple linear regression):

If there is only one independent variable:

$$R^2 = r^2$$

Where r is the Pearson correlation coefficient between the observed and predicted values.

Example:

Let's say you run a regression predicting student test scores from hours of study.

- $SS_{\text{tot}} = 1000$
- $SS_{\text{res}} = 200$

$$R^2 = 1 - \frac{200}{1000} = 0.8$$

Interpretation: 80% of the variation in test scores is explained by hours of study.

In another example; if we find that in a simple linear regression model, if the R^2 value is 0.85, it means that 85% of the variance in the dependent variable can be explained by the independent variable, and the remaining 15% is unexplained or due to randomness.

In summary, the Coefficient of Determination is a key measure to evaluate the effectiveness and explanatory power of regression models, though it should be considered alongside other metrics and validation techniques.

Standard Error of Estimate

We have seen the equation $Y = a + bX$ is used to estimate a theoretical value of Y for a given value of X . If the relationship is not perfect the actual values will not coincide with the theoretical values, because of the scatter of variation about the line: If the scatter is definitely measured the variation may then be allowed for and a range established within which a given proportion of values will fall.

The measure used for this purpose, the standard Error of Estimate, is similar to the Standard Deviation. The standard deviation measures the variation or scatter about the arithmetic mean, while the standard error of estimate is a measure of the variation or scatter about the line of regression.

Symbolically,

$$S_y = \sqrt{\frac{\sum(Y - Y_c)^2}{N}}$$

$$S_x = \sqrt{\frac{\sum(X - X_c)^2}{N}}$$

and

Thus the larger the value of the S_y or S_x , the greater the scatter about the line of regression. In such a case the degree of correlation between series

The standard deviation measures the variation or scatter about the arithmetic mean, while the standard error of estimate is a measure of the variation or scatter about the line of regression.

will be poorer. But the standard error of estimate is an absolute measure and, therefore two standard error of estimates cannot be compared as they will be in different units. Now if the standard error of estimate is divided by the standard deviation, the resulting value will be:

$$\frac{S_y}{\delta_y}$$

This ratio is used for computing coefficient of correlation. Symbolically,

$$r = \sqrt{1 - \frac{S_y^2}{\delta_Y^2}}$$

Also the standard error of estimate can be computed by this as follows:

$$r^2 = 1 - \frac{S_y^2}{\delta_Y^2}$$

$$S_y^2 = \delta_Y^2 (1 - r^2)$$

$$S_y = \delta_Y \sqrt{(1 - r^2)}$$

Example 7.5. From the data in Example 7.2, calculate r .

$$S_y^2 = \sqrt{\frac{\sum(Y - Y_c)^2}{N}} = \frac{(0.01)^2}{8} = 0.00013$$

$$\delta^2 = \frac{\sum Y^2}{N} = \frac{68}{8} = 8.5$$

$$r^2 = 1 - \frac{S_y^2}{\delta_Y^2} = 1 - \frac{0.00013}{8.5}$$

$$= 1 - 0.000016 = 0.999994$$

$$r = +0.999$$

Regression Coefficients

To better understand the impact of predictors on outcomes, regression analysis revolves around the regression coefficient. Economics, business, the social sciences, and healthcare are just a few of the many areas that rely on it to deduce links, forecast outcomes, and direct decision-making. Analysts can determine the direction, relevance, and strength of these interactions by looking at regression coefficients.

The regression coefficient is a statistical measure that represents the relationship between an independent variable (predictor) and a dependent variable (outcome) in a regression model. In simple terms, it indicates how much the dependent variable is expected to change when the independent variable increases by one unit, assuming all other factors remain constant.

The slope of the regression line is known as the regression co-efficient. It is the value of b in the regression equations. It is also known the coefficient of slope and it may have positive or negative value. As mentioned earlier since there are two regression equations, there are also two regression coefficients: b_{yx} and b_{xy} .

The regression coefficient is a statistical measure that represents the relationship between an independent variable (predictor) and a dependent variable (outcome) in a regression model.

The slope of the regression line is known as the regression co-efficient.

Example:

Let's say we're using regression to predict sales based on advertising spend:

- If the regression coefficient for advertising spend is 5, it means for each additional unit of currency spent on advertising, sales are expected to increase by 5 units.
- If the coefficient is -2, it means that for each additional unit of currency spent, sales are expected to decrease by 2 units, indicating an inverse relationship.

Regression coefficient of X on Y

As mentioned above it is represented by b_{xy} . It measures the change in X corresponding to unit change in Y. The regression coefficient of X on Y is represented by

$$b_{xy} = r \frac{\delta_x}{\delta_y}$$

where r denotes coefficient of correlation.

We have seen earlier that where deviations are taken from the arithmetic means of X and Y, the regression coefficient is given by

$$b_{xy} = \frac{\sum xy}{\sum y^2}$$

This result can also be obtained directly from $b_{xy} = r \frac{\delta_x}{\delta_y}$

We have seen in unit 6 that correlation can be found out with the help of a product moment formula, i.e.,

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

Also, we know ;

$$\delta_x = \sqrt{\frac{\sum x^2}{N}} \text{ and } \delta_y = \sqrt{\frac{\sum y^2}{N}}$$

Substituting them in the above formula, we get,

$$\begin{aligned} b_{xy} &= \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}} \times \frac{\sqrt{\frac{\sum x^2}{N}}}{\sqrt{\frac{\sum y^2}{N}}} \\ &= \frac{\sum xy}{\sum y^2} \end{aligned}$$

Also it has been given that where deviations are taken from assumed means, the value of b_{xy} can be obtained as follows :

$$b_{xy} = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_y^2 - \frac{(\sum d_y)^2}{N}}$$

Regression Coefficient of Y on X:

As mentioned above, it is represented by b_{yx} and it measures the change in Y corresponding to a unit change in X. It is given by:

$$b_{yx} = r \frac{\delta y}{\delta x}$$

If deviations are taken from actual arithmetic means of X and Y, then

$$b_{yx} = \frac{\sum xy}{\sum x^2}$$

If deviations are taken from assumed means of X and Y, then

$$b_{xy} = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_x^2 - \frac{(\sum d_x)^2}{N}}$$

Example 7.6. The following results were worked out from scores in Mathematics and English in a certain examination:

	Scores in Mathematics	Scores in English
	(X)	(Y)
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Karl Pearson's correlation coefficient between X and Y = +0.42

Required :

Find both the regression lines. Using these regression, estimate the value of Y for X = 50 and estimate the value of X for Y = 30.

Solution: The likely value of Y corresponding to the given X will be calculated from the regression equation of Y on X, which is given by

$$(Y - \bar{Y}) = r \frac{\delta y}{\delta x} \cdot (X - \bar{X})$$

Substituting the given values in the equation, we get

$$(Y - 47.5) = 0.42 \frac{17.8}{10.8} \cdot (X - 39.5)$$

$$Y = 0.69X - 27.25 + 47.50$$

$$Y = 0.69(50) - 27.25 + 47.50$$

$$Y = 34.50 - 27.25 + 47.50$$

$$Y = 54.75$$

Similarly, the likely value of X corresponding to the given Y will be calculated from the regression equation of X on Y, which is given by

$$(X - \bar{X}) = r \frac{\delta x}{\delta y} \cdot (Y - \bar{Y})$$

Substituting the given values in the equation, we get

$$(X - 39.5) = 0.42 \frac{10.8}{17.8} (Y - 47.5)$$

$$X = 0.42 \frac{10.8}{17.8} (Y - 47.5) + 39.5$$

$$X = 0.26Y - 12.25 + 39.5$$

$$= 0.26(30) + 27.25$$

$$= 7.80 + 27.25$$

$$= 35.05$$

Example 7.7. From the following data, find the slope of the regression equation of Y on X.

X	6	8	9	10	11	12	13	14	16	18
Y	4	7	5	8	6	8	10	8	12	10

Solution:

X	Y	XY	X ²
6	4	24	36
8	7	56	64
9	5	45	81
10	8	80	100
11	6	66	121
12	8	96	144
13	10	130	169
14	8	112	196
16	12	192	256
18	10	180	324
117	78	981	1491

We know that;

$$b = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2}$$

Substituting the values in the above formula,

$$b = \frac{10(981) - (117)(78)}{10(1491) - (117)^2} = \frac{9810 - 9126}{14910 - 13689}$$

$$= \frac{684}{1221}$$

$$= 0.56$$

Example 7.8. Write down the equation of the line of regression of Y on X

X:	78	89	97	69	59	79	68	61
Y:	125	137	156	112	107	136	123	108

(Assume : 69 as the working mean for X and 112 for Y).

Solution :

$$b_{yx} = \frac{\sum d_x d_y - \frac{(\sum d_x)(\sum d_y)}{N}}{\sum d_x^2 - \frac{(\sum d_x)^2}{N}}$$

Substituting the values in the above formula for Regression Coefficient of Y on X

$$\begin{aligned} b_{yx} &= \frac{2160 - \frac{48(108)}{8}}{1530 - \frac{(48)^2}{8}} \\ &= \frac{2160 - 648}{1530 - 288} = \frac{1512}{1242} = 1.22app. \end{aligned}$$

The equation of the line of regression of Y on X :

$$\begin{aligned} Y - \bar{Y} &= b_{yx}(X - \bar{X}) \\ Y - 125.5 &= 1.22(X - 75) \\ Y - 125.5 &= 1.22X - 91.5 \\ Y &= 125.5 - 91.5 + 1.22X \\ Y &= 34 + 1.22X \end{aligned}$$

Example 7.9. Calculate the coefficient of correlation and obtain the lines of regression for the following g data:

X:	1	2	3	4	5	6	7	8	9
Y:	9	8	10	12	11	13	14	16	15

Required:

Obtain an estimate of Y which should correspond on the average to X =6.2.

Solution:

X	$(X - \bar{X})$ d_x	d_x^2	Y	$(Y - \bar{Y})$ d_y	d_y^2	$d_x d_y$
1	-4	16	9	-3	9	12
2	-3	9	8	-4	16	12
3	-2	4	10	-2	4	4
4	-1	1	12	0	0	0
5	0	0	11	-1	1	0
6	1	1	13	+1	1	1
7	2	4	14	+2	4	4
8	3	9	16	+4	16	12
9	4	16	15	+3	9	12
45	0	60	108	0	60	57

$$\begin{aligned}\bar{X} &= \frac{\sum X}{N} = \frac{45}{9} = 5; \bar{Y} = \frac{\sum Y}{N} = \frac{108}{9} = 12 \\ r &= \frac{\sum d_x d_y}{\sqrt{\sum d_x^2 \cdot \sum d_y^2}} = \frac{47}{\sqrt{60(60)}} \\ &= \frac{57}{60} = +0.95 \\ b_{yx} &= \frac{\sqrt{xy}}{\sqrt{y^2}} = \frac{57}{60} = +0.95; b_{xy} = \frac{\sqrt{xy}}{\sqrt{x^2}} = \frac{57}{60} = +0.95;\end{aligned}$$

Regression equation of Y on X :

$$(Y - \bar{Y}) = b_{yx}(X - \bar{X})$$

Substituting the values,

$$\begin{aligned}Y - 12 &= +0.95(X - 5) \\ Y &= 7.25 + 0.95X\end{aligned}$$

Regression equation of X on Y :

$$(X - \bar{X}) = b_{xy}(Y - \bar{Y})$$

Substituting the values,

$$\begin{aligned}X - 5 &= +0.95(Y - 12) \\ X &= 6.4 + 0.95Y\end{aligned}$$

When X=6.2 , the corresponding value of Y will be:

$$\begin{aligned}Y &= 7.25 + 0.65X \\ &= 7.25 + 0.95(6.2) \\ &= 7.25 + 5.89 = 13.14\end{aligned}$$

Use of Regression Coefficient:

The regression coefficient indicates the strength of the association and whether it is positive or negative.

To comprehend and measure the link between variables in a regression study, we employ the regression coefficient. With all other variables held constant, it assists us in estimating the amount that the dependent variable should change when an independent variable changes by one unit. This is especially helpful for trend analysis, outcome prediction, and data-driven decision making. Additionally, the regression coefficient indicates the strength of the association and whether it is positive or negative. It aids in evaluating the distinct effects of every predictor variable in multiple regression. All things considered, it is essential for deciphering models, directing choices, and extracting valuable information from data.

1. Prediction: The primary application of the regression coefficient is in forecasting the dependent variable (outcome). Upon obtaining the coefficients, they can be utilized to forecast future values of yyy based on established values of xxx. In economics, regression coefficients can predict future GDP growth based on variables such as inflation, government expenditure, or consumer confidence.

The sign of the coefficient, whether positive or negative, denotes the direction of the relationship.

2. Comprehending Interconnections Among Variables: The magnitude of the regression coefficient indicates the extent to which the dependent variable alters in response to a one-unit change in the independent variable. A greater absolute value of the coefficient signifies a more robust link. The sign of the coefficient, whether positive or negative, denotes the direction of the relationship. A positive coefficient indicates

that a rise in the independent variable corresponds to an increase in the dependent variable. A negative coefficient indicates that an increase in the independent variable results in a decrease in the dependent variable.

3. Identifying Key Influences: In multiple regressions, each regression coefficient signifies the effect of a certain predictor while accounting for the influences of other predictors. This facilitates the identification of the most significant factors influencing the outcome. In a model forecasting house price, the regression coefficient for the number of bedrooms indicates the price increment associated with each extra bedroom, while controlling for other variables such as square footage.

4. Policy Formulation: Regression coefficients are frequently employed in policy analysis to measure the effects of certain interventions (e.g., educational expenditure, healthcare initiatives) on critical outcomes (e.g., literacy rates, life expectancy). A government may do regression analysis to assess the influence of infrastructure spending on economic growth.

5. Hypothesis Testing: Regression coefficients are utilized in hypothesis testing to evaluate the statistical significance of the relationship between the independent and dependent variables. A regression coefficient that is substantially different from zero, usually assessed by a t-test, indicates that the independent variable exerts a substantial influence on the dependent variable.

6. Enhancing Business Decisions: Businesses employ regression coefficients to refine plans. A corporation may do regression analysis to ascertain the impact of advertising expenditure on sales or the effect of customer pleasure on loyalty. The coefficients assist in formulating data-driven decisions, such as the allocation of funding for marketing or product development.

7. Comprehending Multivariable Interactions: In multiple regression, regression coefficients elucidate the interactions among various independent variables and a dependent variable concurrently. A model may evaluate the combined impact of education level and income on health outcomes, with each coefficient indicating the distinct contribution of one variable while accounting for the others.

Relationship between Coefficient of Correlation and Regression Coefficients

The coefficient of correlation and regression coefficient are both important concepts in statistics and are used to describe the relationship between two variables. However, they serve different purposes and are related in certain ways.

The coefficient of correlation (often denoted as r) measures the strength and direction of the linear relationship between two variables. It ranges from -1 to +1, with the following interpretations:

- $r = 1$: Perfect positive correlation (both variables increase together).
- $r = -1$: Perfect negative correlation (one variable increases as the other decreases).
- $r = 0$: No linear relationship between the variables.

In multiple regression, regression coefficients elucidate the interactions among various independent variables and a dependent variable concurrently.

The coefficient of correlation (often denoted as r) measures the strength and direction of the linear relationship between two variables.

A regression coefficient represents the slope of the regression line in a linear regression model, which describes how the dependent variable (Y) changes in response to changes in the independent variable (X).

A **regression coefficient** represents the **slope** of the regression line in a linear regression model, which describes how the dependent variable (Y) changes in response to changes in the independent variable (X).

For a simple linear regression model one way we can express the regression model as: $Y = a + bX$, the regression coefficient b is calculated as:

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Where:

- $\bar{X}$ and $\bar{Y}$ are the means of the variables X and Y , respectively.
- b represents the change in Y for a one-unit change in X .

The coefficient of correlation (r) and the regression coefficient (b) are related through the following equation:

$$b = r \times \frac{s_Y}{s_X}$$

Where:

- r is the correlation coefficient.
- s_Y is the standard deviation of the dependent variable (Y).
- s_X is the standard deviation of the independent variable (X).

Interpretation of the Relationship:

- The regression coefficient (b) gives you the slope of the regression line, which indicates how much changes when X increases by one unit.
- The correlation coefficient (r) measures the strength and direction of the linear relationship between X and Y .
- When r is large (close to +1 or -1), the regression coefficient b will have a more significant slope, implying a strong relationship between the variables.

In summary, we can conclude that:

- Correlation coefficient measures the strength and direction of a linear relationship.
- Regression coefficient quantifies the change in the dependent variable for a one-unit change in the independent variable.
- They are related by the formula $b = r \times \frac{s_Y}{s_X}$, indicating that the regression slope depends on both the strength of the relationship (correlation) and the standard deviations of the variables involved.
- Denoting Regression coefficient of Y on X by:

$$d_{YX} = r \frac{\sigma_Y}{\sigma_X} \text{ and}$$

- Regression coefficient of X on Y by:

$$d_{XY} = r \frac{\sigma_X}{\sigma_Y}$$

$$d_{XY} \times d_{YX} = r \frac{\sigma_Y}{\sigma_X} \times r \frac{\sigma_X}{\sigma_Y}$$

$$= r^2$$

$$r = \sqrt{d_{XY} \times d_{YX}}$$

Correlation coefficient measures the strength and direction of a linear relationship.

Regression coefficient quantifies the change in the dependent variable for a one-unit change in the independent variable.

It is clear from the relationship that r is the geometric mean between the two regression coefficients. In other words, the under root of the product of the two regression coefficients gives us the value of r .

It is to be noted that as the value of r cannot exceed one, one of the regression coefficients must be less than one. In other words, both the regression coefficients cannot be greater than one. Also both the regression coefficients will have the same sign i.e., they will be either positive or negative. The coefficient of correlation will have the same sign as that of regression coefficients.

The relationship is explained with the help of the following example:

Example 7-10. In a partially destroyed laboratory record of an analysis of correlation data, the following results are legible:

$$\delta_X^2 = 9, \text{ Regression equation } 8X - 10Y + 66 = 0, 40X - 18Y = 214$$

Required:

What were (a) the mean value of X and Y , (b) δ_Y (c) the coefficient of correlation between X and Y ?

Solution:

The regression equations given are:

$$8X - 10Y = -66 \quad \text{.....(i)}$$

$$40X - 18Y = 214 \quad \text{.....(ii)}$$

Since the lines of regression pass through the means $(\bar{X}, \bar{Y})$ of the distribution, we have

$$8X - 10Y = -66 \quad \text{.....(iii)}$$

$$40X - 18Y = 214 \quad \text{.....(iv)}$$

Solving (iii) and (iv), we get

$$\bar{X} = 13 \text{ and } \bar{Y} = 17$$

Assuming regression of, Y on X is given by

$$Y = \frac{8}{10}X + \frac{66}{10}$$

An the regression of X on Y is given by

$$X = \frac{18}{40}Y + \frac{214}{40}$$

$$\therefore \text{Regression Coefficient of } Y \text{ on } X = \frac{8}{10}$$

$$\therefore \text{Regression Coefficient of } X \text{ on } Y = \frac{18}{40}$$

$$\therefore r = \sqrt{\frac{8}{10} \times \frac{18}{40}} = 0.6$$

$$\text{Assuming Regression, coefficient of } Y \text{ on } X = \frac{8}{10}$$

Or

$$\begin{aligned}
 r \frac{\sigma_Y}{\sigma_X} &= \frac{8}{10} \\
 0.6 \left(\frac{\sigma_Y}{3} \right) &= \frac{8}{10} \\
 \frac{\sigma_Y}{2} &= \frac{8}{10} \times \frac{1}{0.6} \\
 \sigma_Y &= \frac{8}{10} \times \frac{10 \times 3}{0.6} = 4
 \end{aligned}$$

Example 7.11. Find the coefficient of correlation when

$$b_{XY} = 0.84 \text{ and } b_{YX} = 0.40$$

Solution:

Coefficient of correlation

$$\begin{aligned}
 &= \sqrt{b_{XY} b_{YX}} \\
 &= \sqrt{0.84 \times 0.40} \\
 &= \sqrt{0.3360} = 0.58
 \end{aligned}$$

Miscellaneous Problems

1. Consider the information given below:

	X Series	Y Series
Mean	18	1000
Standard Deviation	14	20

Coefficient of correlation between X and Y is +.8.

Required:

- Find out the most probable value of Y if X is 70 and most probable value of X if Y is 90.
- If the regression coefficients are 0.8 and 0.6, what would be the value of the coefficient of correlation.

Solution:

(a) The regression equation of Y on X:

$$(Y - \bar{Y}) = r \frac{\sigma_X}{\sigma_Y} (X - \bar{X})$$

Substituting the values, we get

$$\begin{aligned}
 (Y - 100) &= \frac{0.8(20)}{14} (X - 18) \\
 Y &= 79.48 + 1.14X
 \end{aligned}$$

Now when

$$X = 70,$$

$$Y = 79.48 + 1.14(70) = 159.28$$

The regression equation of X on Y:

$$(X - \bar{X}) = r \frac{\sigma_Y}{\sigma_X} (Y - \bar{Y})$$

$$\begin{aligned}
 (X - 18) &= 0.8 \frac{14}{20} (Y - 100) \\
 X &= -38 + 0.56Y
 \end{aligned}$$

Now when

$$Y = 90,$$

$$X = -38 + 0.56(90) = 12.4$$

(b)

$$r = \sqrt{d_{yx} \times d_{xy}}$$

$$r = \sqrt{0.8(0.6)} = \sqrt{0.48} = 0.69 \text{ approx.}$$

2. Find out σ_y and r from the following data:

$$3x = y, 4y = 3x, \text{ and } \sigma_x = 2.$$

Solution:

The regression equation of y on x is

$$y = \frac{3}{4}x \quad [\because 4y = 3x]$$

The regression equation of x on y is

$$x = \frac{1}{3}y \quad [\because 3x = y]$$

$$r \frac{\sigma_y}{\sigma_x} = \frac{3}{4} \text{ and } r \frac{\sigma_x}{\sigma_y} = \frac{1}{3}$$

$$r^2 = r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y}$$

$$r^2 = \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{4}$$

$$r = \sqrt{\frac{1}{4}} = \frac{1}{2} = +0.50$$

Now

$$r \frac{\sigma_y}{\sigma_x} = \frac{3}{4}$$

$$0.5 \frac{\sigma_y}{2} = \frac{3}{4}$$

$$\sigma_y = 3$$

3. Two lines of regression are given by $x+2y=5$ and $x+3y=8$ and $\sigma_y^2 = 12$.

Required:

Calculate the values of $\bar{X}$, $\bar{Y}$, σ_y^2 and r .

Solution: Solving the given equations, we get

$$x+2y=5 \dots (i)$$

$$2x+3y=8 \dots (ii)$$

Multiply (i) by 2 and then subtracting (ii) from it, we get: $y=2$

Putting $y=2$ in (i), we get $x=1$

As the lines of regression pass through the means, therefore the values of $\bar{X}$ and $\bar{Y}$ are 1 and 2 respectively.

$$\text{Now } b_{xy} = -\frac{3}{2} \quad [\because x = -\frac{3}{2}y + 4]$$

$$b_{yx} = -\frac{1}{2} \quad [\because y = -\frac{1}{2}x + \frac{5}{2}]$$

$$r^2 = b_{xy} \times b_{yx}$$

$$= \left(-\frac{3}{2}\right) \left(-\frac{1}{2}\right) = \frac{3}{4}$$

$$r = -\frac{\sqrt{3}}{2} \quad [\text{Negative sign is taken as the } b_{xy} \text{ and } b_{yx} \text{ are negative}]$$

$$= -0.87$$

$$\begin{aligned}\text{Now } r \frac{\sigma_y}{\sigma_x} &= -\frac{1}{2} \\ \sigma_y &= -\frac{1}{2} \times \frac{\sigma_x}{r} \\ &= -\frac{1}{2} \times \left(\frac{\sqrt{12}}{-\sqrt{3}} \right) 2 = 2 \\ \sigma_y^2 &= 4\end{aligned}$$

4. The following table shows the frequency according to age- groups of marks obtained by 65 students in a general knowledge test.

Required:

Measure the following:

- The Regression Equations.
- The mean age in years and mean test marks.
- The Regression coefficients.
- Coefficient of correlation between age and general knowledge.

Test Marks	Age in Years			
	19	20	21	22
200-250	4	4	2	1
250-300	3	5	4	2
300-350	2	6	8	5
350-400	1	4	6	8

Solution:

Let's assume: Age in Years to be X
Test marks to be Y

Calculations to Find out Regression Equations, Coefficient of Correlation, Means, Regression Coefficients

					X											
X→					19	20	21	22								
$X = d_x$					-2	-1	0	+1								
d_x^2					4	1	0	1								
Mid. Points(y-325) by 50																
	Y	Y'	d_x	d_x^2					f_y	fd_y	fd_y^2	$fd_x d_y$				
200–250	225	-100	-2	4	4	16	4	8	2	0	1	-2	11	-22	44	22
250–300	275	-50	-1	1	3	6	5	5	4	0	2	-2	14	-14	14	9
300–350	325	0	0	0	2	0	6	0	8	0	5	0	21	0	0	0
350–400	375	+50	+1	1	1	-2	4	-4	6	0	8	8	19	19	19	2
$f_x =$					10	19	20	16	N=65	$\Sigma fd_y = -17$	$\Sigma fd_y^2 = 77$	$\Sigma fd_x d_y = 33$				
$fd_x =$					-20	-19	0	16	$\Sigma fd_x = -23$							
$fd_x^2 =$					40	19	0	16	$\Sigma fd_x^2 = 75$							
$fd_x d_y =$					20	9	0	4	$\Sigma fd_x d_y = 33$							

Calculations

	For X	For Y
Mean	$\bar{X} = A + \frac{\sum fd_x}{N} \times i$ Where A=21, N=65, $\sum fd_x = -23$ $\therefore X = 21 + \frac{-23}{65} = 21 - 0.35$ $= 20.65$	$\bar{Y} = A + \frac{\sum fd_y}{N} \times i$ $\bar{Y} = 325 + \frac{-17}{65} \times 50$ $= 325 - 13.00 \text{ approx}$ $= 312 \text{ approx.}$
Regression Equations	Reg. Equation of X on Y: $(X - \bar{X}) = b_{xy}(Y - \bar{Y})$ $b_{xy} = \frac{\sum fd_x d_y - \frac{\sum fd_x \sum fd_y}{N}}{\sum fd_y^2 - \frac{(\sum fd_y)^2}{N}} \times \frac{ix}{iy}$ $b_{xy} = \frac{33 - \frac{(-23)(-17)}{65}}{77 - \frac{(-17)^2}{65}} \times \frac{1}{50}$ $\therefore b_{xy} = \frac{33 - 6}{77 - 4.46} \times \frac{1}{50}$ $= \frac{27}{72.55} \times \frac{1}{50} = \frac{27}{3627.5} = 0.0074$ $\therefore$ Reg. Equation of X on Y is $(X - 20.65) = 0.0074(Y - 312)$ $X - 20.65 = 0.0074Y - 2.31$ $X = 0.0074Y + 18.34.$	Reg. Equation of Y on N: $(Y - \bar{Y}) = b_{yx}(X - \bar{X})$ $b_{yx} = \frac{\sum fd_x d_y - \frac{\sum fd_x \sum fd_y}{N}}{\sum fd_x^2 - \frac{(\sum fd_x)^2}{N}} \times \frac{iy}{ix}$ $b_{yx} = \frac{33 - \frac{(-23)(-17)}{65}}{75 - \frac{(-23)^2}{65}} \times \frac{50}{1}$ $\therefore b_{yx} = \frac{33 - 6}{75 - 8.14} \times \frac{50}{1}$ $= \frac{27}{66.86} \times \frac{50}{1} = \frac{1350}{66.86} = 20.19$ $\therefore$ Reg. Equation of Y on X is $(Y - 312) = 20.19(X - 20.65)$ $Y - 312 = 20.19X - 416.9$ $Y = 20.19X - 104.9.$
Regression Coefficients	$b_{xy} = 0.0074$	$b_{yx} = 20.19$
Coefficients Correction	$r = \sqrt{b_{yx} b_{xy}}$	$= \sqrt{20.19 \times 0.0074}$ $= 0.386.$

Self-Assessment Questions:

Multiple Choice Questions:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) To determine whether a linear relationship exists between variables, a _____ can be used.
 - a) bar chart
 - b) pie chart
 - c) scatter chart
 - d) stacked column chart
- (ii) In least-squares regression, the best-fitting line minimizes _____.
 - a) the sum of the squares of the dependent variables
 - b) the squares of the slope and intercept term
 - c) the squares of the mean values of X and Y
 - d) the sum of squares of the observed errors
- (iii) For the least-squares equation $\bar{Y} = 3,698 + 2,538X$, Y represents house prices and X represents number of rooms. Which of the following statements is true?
 - a) Only 36.98% of the variation in house prices can be explained by the number of rooms in the house.
 - b) For every additional room that is added in a house, house prices increase by \$2,538.
 - c) As the number of rooms in the house increase, prices fall by \$3,698.
 - d) 25.38% of the house price is attributed to the number of rooms.
- (iv) Costco sells paperback books in their retail stores and wanted to examine the relationship between price and demand. The price of a particular novel was adjusted each week and the weekly sales were recorded in the table below.

Sales	Price
3	Tk12
4	Tk11
6	Tk10
10	Tk9
8	Tk8
10	Tk7

Management would like to use simple regression analysis to estimate weekly demand for this novel using the price of the novel. The sum of squares regression for this sample is _____.

- a) 24.17 b) 37.20
c) 40.16 d) 46.25

- (v) Costco sells paperback books in their retail stores and wanted to examine the relationship between price and demand. The price of a particular novel was adjusted each week and the weekly sales were recorded in the table below.

Sales	Price
3	\$12
4	\$11
6	\$10
10	\$9
8	\$8
10	\$7

Management would like to use simple regression analysis to estimate weekly demand for this novel using the price of the novel. The coefficient of determination for this sample is ____.

- a) 0.336 b) 0.624
c) 0.830 d) 0.881

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) If the slope of the simple regression equation is equal to zero, the scatter plot for the ordered pairs will be a vertical straight line indicating that there is no relationship between the independent and dependent variables.
- (ii) When the slope of a population regression line equals zero, we conclude that there is a linear relationship between the dependent and independent variables.
- (iii) Given a regression equation of $\hat{y} = 16 + 2.3x$ we would expect that an increase in x of 2.0 would lead to an average increase of y of 4.6.
- (iv) When the relationship between the variables is statistically significant using simple regression analysis, we have enough evidence to state that the independent variable caused the change in the dependent variable.
- (v) In a simple regression model, the slope coefficient represents the average change in the independent variable for a one-unit change in the dependent variable.

Answer:

Multiple-Choice Question:

1. (i) c (ii) d (iii) b (iv) b (v) c

True/False

2. (i)- F (ii)- F (iii)- T (iv)- F (v)- F

Review Questions

- I. Explain the concept of 'Correlation' and 'Regression' and describe the main properties of Karl Pearson's coefficient of correlation.
1. Define Regression. How would interpret the sign and imaginative of a calculated r.
2. Define scatter diagram? How does it help in studying the correlation between two variable, in respect of both its direction and degree?
3. What is spearman's rank correlation coefficient? Interpret the relationship between two variable for which the coefficient of correlation is +1.
4. Define regression. Distinguish between correlation and regression.
5. Explain the concept of regression and point out its usefulness in dealing with business problems.
6. What is a scatter diagram? Indicate by means of suitable scatter diagrams different types of correlation that may exist between the variables in bivariate data. What are regression lines? Write down the main points of distinction between correlation analysis and regression analysis.
7. Distinguish between correlation and regression analysis and indicate the utility of regression analysis in economic activities.
8. What is regression analysis? How does it differ from correlation? Why there are, in general, two regression equations?
9. Comment on the following: "Regression equations are irreversible".
10. Explain by a graphic illustration or otherwise the meaning of the term regression equation.
11. Why there are two equations of regression?
12. Explain the meaning of regression of Y on X and X on Y.
13. Given the following statistical coefficient deduced in the course of an examination of the relationship between yield of gram and the amount of rainfall, calculate (a) the most likely yield when the annual rainfall is 9.2 inches, and (b) the probable annual rainfall for yield of 1400 lbs. per acre:

	Yield in lbs. per acre	Annual rainfall in inches
Mean	995.0	12.8
Standard Deviation	70.1	1.6
Coefficient of correlation between yield and rainfall	+0.52	

14. 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

15. You are given the following sample data for variables y and x :

y	140	120	80	100	130	90	110	120	130	130	100
x	5	3	2	4	5	4	4	5	6	5	4

Required:

- Develop a scatter plot for these data and describe what, if any, relationship exists.
 - Compute the correlation coefficient.
16. The following data, based on 450 candidates, are given for marks Statistics and Accountancy at a certain examination:

Mean marks in Statistics	40
Mean marks in Accountancy	48
S.D. of marks in Statistics'	12
S.D. of marks in Accountancy	16
Sum of the product of deviations of marks from their respective means	42075

Required:

- Give the equations to the two lines of regression, and explain why there are two regression lines.
 - Estimate the mean marks in Accountancy of the candidates who obtained 50 marks in Statistics.
17. From the data given below find:
- The two regression coefficients.
 - The two regression equations.
 - The coefficient of correlation between the marks in Economics and Statistics.
 - The most likely marks in Statistics when marks in Economics are 30.

Marks in Economics:	25	28	35	32	31	36	29	38	34	32
Marks in Statistics :	43	46	49	41	36	32	31	30	33	39

18. Calculate Karl Pearson's coefficient of correlation and the regression from the following data:

Age of husband	18	19	20	21	22	23	24	25	26	27
Age of wife	17	17	18	18	18	19	19	20	21	22

19. Given the following data find what will be the probable yield when the rainfall is 29".

	Rainfall	Production
$\bar{X}$	25"	40 units per acre
σ	3"	6 units

r between rainfall and production = 0.8

20. You are given the following sample data for variables x and y :

x (independent)	y (dependent)
1	16
7	50
3	22
8	59
11	63
5	46
4	43

Required:

- Construct a scatter plot for these data and describe what, if any, relationship appears to exist.
- Compute the regression equation based on these sample data and interpret the regression coefficients.
- Based on the sample data, what percentage of the total variation in the dependent variable can be explained by the independent variable?

21. The following data are given for marks in English and Mathematics in the S.L.C. examination of the U.P. in a certain year.

Mean marks in English	39.5
Mean marks in Mathematics	47.6
σ marks in English	10.8
σ marks in Mathematics	16.9
r between marks in English and Math.	0.42

Required:

From the two lines of regression, calculate the expected average marks in Mathematics of candidates who received 50 marks in English.

22. Consider the following sample data for the variables y and x :

x	30.3	4.8	15.2	24.9	8.6	20.1	9.3	11.2
y	14.6	27.9	17.6	15.3	19.8	13.2	25.6	19.4

Required:

- Calculate the linear regression equation for these data.
- Determine the predicted y value when $x = 10$.
- Estimate the change in the y variable resulting from the increase in the x variable of 10 units.

23. The following table gives the ages and blood pressure of 10 women:

Age in years x	56	42	36	47	49	42	72	63	55	60
Blood pressure y	147	125	118	128	125	140	155	160	149	150

Required:

- Draw a scatter diagram.
- Find correlation coefficient between x and y and comment.

24. The scores of 12 students in their mathematics and physics classes are:

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

Required:

Find the correlation coefficient distribution and interpret it.

25. On the basis of figures recorded below for 'Supply' and 'Price' for nine years, build a regression of 'Price' on 'Supply'. Calculate, from the equations established, the most likely Price, when Supply=90.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
Supply	80	82	86	91	83	85	89	96	98
Price	145	140	130	124	133	127	120	110	116

26. Obtain the straight line of best fit for the following data on the production of mill cloth in India.

Year	1995	2000	2005	2010	2015	2020	2025
Production (1000Yards)	81	48	88	104	134	148	170

Required:

Estimate the production in 2023.

27. In the following data find the regression of yield of straw (Y) on Yield of grain (X) in lbs. from plots of 1/40 acre.

Grain	54	68	57	63	54	62	60	63	62	61	64	60
Straw	17	27	19	19	18	20	26	21	24	25	20	18

28. The following data refer to information about annual sales (Tk.'000) and year of experience of a super store of 8 salesmen:

Salesmen	1	2	3	4	5	6	7	8
Annual sales (Tk.'000)	90	75	78	86	95	110	130	145
Year of experience	7	4	5	6	11	12	13	17

Required:

- Fit two regression lines.
- Estimate sales for year of experience is 10.
- Estimate year of experience for sales 100000.

29. The following figure related to advertisement expenditure and profit:

Profit (Tk. Crore): x	25	28	27	33	31	10	16	16	18	23
Adv. Exp. (Tk. Lakh): y	87	91	92	95	93	52	68	72	78	86

Required:

- Draw a scatter diagram and comment
- Find Karl Pearson's correlation coefficient

30. Two lines of regression are given:

$$X+2Y-5=0 \text{ and}$$

$$2X+3Y-8=0.$$

$$\sigma_x^2 = 12.$$

Required:

Calculate the value of $\bar{X}$, $\bar{Y}$, σ_y and r .

31. Find out the coefficient of correlation between the deaths from the fevers and total deaths given below.

Year	Deaths from fevers	Deaths from other causes	Total Deaths
1	1025	281	1306
2	853	223	1076
3	698	207	1076
4	970	325	1295

Required :

Calculate standard error of this coefficient and the line of regression of the deaths from fevers on total deaths.

32. If two regression coefficients are 0.8 and 0.2 what would be the value of r ?

33. Are the following two statements consistent? Give reasons.

The regression coefficient of X on Y is 3.2 and that of Y on X is 0.8.

34. Two random variables have the least square regression lines with equations: $3X + 2Y = 26$ and $6X + Y = 31$.

Required:

Find the mean values and the coefficient of correlation between X and Y .

35. For 50 students of a class the regression equation of marks in Statistics (X) on the marks in Accountancy (Y) is $3Y - 5X + 180 = 0$. The mean marks of Accountancy is 44 and the variance of marks in Statistics is $9/16$ th of the variance of marks in Accountancy.

Required:

Find the mean marks of Statistics and the coefficient of correlation between marks in two subjects.

36. Given the following data :

Variance of $x = 9$

Regression equations:

$$4x - 5y + 33 = 0$$

$$20x - 9y - 107 = 0$$

Required:

- the mean values of x and y
- the standard deviation of y ,
- the coefficient of correlation between x and y .

37. Two random variables have the regression with equations :

$$3X + 2Y - 26 = 0 \text{ and} \\ 6X + Y - 31 = 0$$

Required:

- (i) Find the mean values and the correlation coefficient between X and Y.
 - (ii) If the variance of X is 25, find from the data given above.
38. An index of production of a particular commodity follows the following trend :

$$Y = 204 + 6X$$

where origin is the year 1976, time unit is 1 year and Y represents totals of yearly production.

Find the monthly trend equation with the origin at July, 1976.

$$[Y = 207 + 0.5X]$$

39. From a sample of 200 pairs of observations the following quantities were calculated:

$$\begin{aligned} \sum X &= 11.34 & \sum Y &= 20.72 & \sum X^2 &= 12.16 \\ \sum Y^2 &= 84.96 & \sum XY &= 22.13 \end{aligned}$$

Required:

From the above data show how to compute the coefficient if the equation

$$Y = a + bX$$

40. The following calculations have been made for closing prices of twelve stocks (X) on the Chattogram Stock Exchange on a certain day, along with the volume of sales in thousands of shares (Y).

Required:

From these calculations, find the regression equation.

$$\begin{aligned} \sum X &= 580 ; & \sum Y &= 370 ; & \sum XY &= 11,494 \\ \sum X^2 &= 41,658, & \sum Y^2 &= 17,206. \end{aligned}$$

INDEX NUMBERS



Numerical characteristics of changes all fluctuations in prices, quantities all values are widely used in business. When they appear as absolute values, such as the value that is added by the manufacturer, consumer products produced, they are generally called business indicators. If they are expressed as relative numbers, that is, as a percent of some base time or place, they are called index numbers. The theory of index numbers is widely applied in business and economics. Index numbers may be classified according to the procedure used in their computations, the weights used or the phenomena which is characterized. Business indexes usually measure changes in prices, quantities, or values, but may be used for other phenomena, such as expenditure for some specified purpose, physical or environmental qualities, and so on. The distinguishing characteristic of all index numbers is that they are expressed relative to some specified or generally understood place, time, or period, called the base. Index numbers are shown as percent, the percent sign (%) is never used with published index numbers. So the concept of index numbers in business originated from the need to measure and compare changes in economic and business activities over time, such as prices, costs, sales, and production levels.

Lesson 1: Introduction to Index Number

Lesson Objectives:

After completing this lesson, you will be able to

- Define Index Number;
- Describe the types of index numbers;
- Understand the uses of index numbers;
- Explain the Problems in construction of index numbers.

Introduction

Let's begin the introduction part to understand the meaning of index number with the help of an example. Assume the following information regarding the prices of a group of food items in the years 2010 and 2020:

TABLE 8.1-Prices of Food Items			
Price per unit (Taka)			
Commodity	Unit	2010	2020
Rice	kg.	40	80
Wheat	kg.	28	50
Fish	kg.	90	200
Bread	lb.	45	60
Milk	liter	40	70

On the basis of these data, answer the following question:

"How does the overall food price in 2020 compare with that in 2010-how many times or what per cent?"

A number which provides an answer to this question will be called an "index number".

Several points invite attention:

- (i) The proposed index number will give a comparison between 'prices' and hence it will be a 'price Index number'.
- (ii) The index number covers a 'group' of related items, here food items. Hence, we shall speak of the 'index number of food prices'. Note that 'food' whose price the index number compares, is not a commodity by itself.
- (iii) The comparison is made between two 'periods of time'. Here the position in 2020 is compared against that in 2010.
- (iv) The index number is an 'average' computed from data given in heterogeneous units Quintal, kg., lb., liter. It compares the overall position of food prices, i.e. an average computed from several items. Again, the prices quoted in the table are some sort of 'average' prices--Rice, for instance, is sold in several qualities or grades whose prices are different.
- (v) Lastly, did you notice that the price of Rice in 2010 cannot remain a content throughout the year. It is an average for the whole year. The index number is thus a "special type of average".

Definition of Index Numbers

The concept of index numbers dates back to the 18th and 19th centuries, when statisticians and economists began developing methods to track changes in economic and social indicators over time. One of the earliest known uses was by William Fleetwood in the early 18th century, who tried to compare prices across time to understand changes in the cost of living.

Later, during the industrial revolution and especially in the 19th century, the need to measure changes in prices, wages, production, and trade grew significantly with expanding economies. Economists like Joseph Lowe (1823) and Étienne Laspeyres (1871) contributed to formalizing index number formulas that are still used today (like the Laspeyres Index).

Index Numbers are numerical figures which indicate the relative position in respect of price, or quantity or value of a group of articles at certain periods of time as compared with another period; called base period. When the comparison is in respect of price, they are called 'Price Index Numbers'; similarly, we have 'Quantity Index Numbers' and 'Value Index Numbers'. Index number for the base period is always taken as 100. Index number for any other period, called current period, shows the overall level of price (or quantity or value) of the group of articles as a percentage of that in the base period.

Index numbers gained prominence with the growth of industrial economies, as they provided a systematic way to track economic performance, price changes, cost of living, and business trends. In business statistics, an index number is a statistical measure that shows changes in a variable or group of related variables over time, location, or other characteristics.

In simple words, "An index number is a tool that shows the relative change in the level of a variable (such as price, quantity, or value) compared to a base period."

An index number is a quantitative tool that reflects the relative change in the level of a phenomenon—such as price, quantity, or value—between two or more periods of time.

In business statistics, an index number is a statistical measure designed to show changes in a variable or group of related variables over time, relative to a base period.

Assume the statement "Index Number of Wholesale Prices in Bangladesh for the year 2017 was 185 (Base: 2010-17 = 100)" signifies that as compared with the wholesale prices prevailing during the period 2010-17, the wholesale price of all articles during 2017 was on an average 185%. In other words, prices increased in 2017 by 85% over the 2010-17 prices.

Types of Index Numbers:

In Business Statistics, index numbers are statistical measures designed to show changes in a variable or group of related variables over time, location, or other characteristics. They are widely used to measure economic indicators like prices, production, and costs. Here are the main types of index numbers:

Index Numbers are numerical figures which indicate the relative position in respect of price, or quantity or value of a group of articles at certain periods of time as compared with another period; called base period.

In business statistics, an index number is a statistical measure that shows changes in a variable or group of related variables over time, location, or other characteristics.

"An index number is a tool that shows the relative change in the level of a variable (such as price, quantity, or value) compared to a base period."

1. Price Index Numbers

These measure changes in the price level of goods and services over time.

Consumer Price Index (CPI): Measures changes in the retail prices of goods and services consumed by households.

Wholesale Price Index (WPI): Measures changes in the price of goods at the wholesale level.

Producer Price Index (PPI): Measures the average changes in prices received by domestic producers.

Measures changes in the retail prices of goods and services consumed by households.

2. Quantity Index Numbers

These measure changes in the quantity or volume of goods produced, consumed, or sold over time.

Example: Index of Industrial Production (IIP) – measures the changes in the volume of production of industrial goods.

Changes in the quantity or volume of goods produced, consumed, or sold over time.

3. Value Index Numbers

These reflect changes in the total value (price $\times$ quantity) of items.

Used when both price and quantity change, like in total revenue or total expenditure comparisons over time.

4. Cost of Living Index (COLI)

This is a specialized price index that measures the relative cost of living over time or between locations, taking into account consumption patterns.

5. Wage Index

This tracks changes in wages over time, helping understand labor market trends.

6. Stock Market Index

Measures the performance of a group of stocks, giving an indication of overall market trends.

Examples: S&P 500, Dow Jones Industrial Average, etc.

7. Special Purpose Index Numbers

These are designed for specific uses, such as:

Agricultural Index – for crop production

Export/Import Price Index – For international trade

Retail Price Index (RPI) – similar to CPI, used for different purposes in some countries

Uses of index numbers

Index numbers are primarily used to measure the relative position of business and economic conditions. There are many different types of index numbers and the use of an index number depends on its type. Index numbers of wholesale prices, retail prices, cost of living, industrial production, quantum of exports and imports, business activity, to name only a few, are useful in their own fields.

'Wholesale price index number' tells us about changes taking place in the value of money.

'Consumer price index number', or 'Cost of living index number' measures changes in the real income of people.

Price index numbers are used for various purposes. 'Wholesale price index number' tells us about changes taking place in the value of money. 'Consumer price index number' or 'Cost of living index number' measures changes in the real income of people. It helps in the calculation of dearness allowance, so that the real wage may not decrease. 'Index numbers of stock prices are used by economists, speculators and bankers in various ways. An economist uses them to measure changes in the purchasing power of money over stocks, a speculator uses them for forecasting the future course of the market, and the insurance company may require the index numbers for estimating future interest rate. Similarly index number of industrial production' reveals the comparative position in productivity and 'index number of business activity' throws light on the progress of business conditions.

Index numbers are also used to measure the comparative position in respect of price in different regions at the same period of time, e.g. for comparing the standards of living in several cities.

Index numbers are vital useful tools in business statistics. They help measure changes over time in economic data, making it easier for businesses and analysts to understand trends and make informed decisions. Here are the key uses of index numbers in business statistics:

1. Assessing Variations in Price Levels (Price Index Numbers):

Index statistics, such as the Consumer Price Index (CPI) and Wholesale Price Index (WPI), are utilized to quantify the average variation in prices of products and services over time. The Consumer Price Index aids enterprises in comprehending inflationary trends. An increase of 5% in the CPI indicates that, on average, consumer prices have risen by 5% during the year. A food manufacturing company may utilize the Consumer Price Index (CPI) to determine whether to increase prices in response to rising production expenses.

2. Assessing Cost of Living:

Index figures facilitate the monitoring of the changes in consumer expenditure required to sustain their standard of living over time. For instance, if the cost-of-living index escalates, a human resources department may advocate for augmenting employee compensation to align with inflation. The dearness allowance (DA) for government employees is frequently correlated with the cost-of-living index.

3. Examination of Business and Economic Trends:

Businesses utilize index numbers to monitor fluctuations in production, sales, inventory, or consumption trends over time. A retail corporation monitors a sales index to assess growth. Should the index demonstrate a consistent rise, the corporation may consider expansion plans. Additionally An automobile company may examine a production index to assess output patterns and adjust inventory levels accordingly.

4. Adjusting Financial Data for Inflation:

Index numbers are employed to eliminate the impact of inflation from economic or financial statistics in order to assess genuine growth. For

instance, a company's income increased from 1 crore to 1.1 crore. However, inflation increased by 10% throughout the same timeframe. Employing a deflator index may indicate a lack of genuine growth in wages. This is crucial for comprehending real GDP or actual profit as opposed to merely nominal statistics.

5. Analyzing Performance Over Time or Among Regions:

Index numbers facilitate the straight forward comparison of data across several time periods or locations by transforming raw data into a relative metric. A municipality corporation with locations in many cities may employ a sales performance index to evaluate the success of each branch. A productivity index can evaluate manufacturing outputs in 2020 and 2024 to ascertain enhancements.

6. Analysis of Stock Markets (Stock Indices):

Stock indices encapsulate the performance of the stock market or specific sectors. Investors utilize index fluctuations to assess market sentiment. An ascending indices signifies overall market optimism. A mutual fund manager may assess a fund's performance by comparing its return.

7. Prognostication and Strategic Planning:

Historical index numbers assist firms in predicting future patterns, facilitating strategic planning. A corporation may utilize historical consumer demand indices to forecast future demand and modify output accordingly. Economic planners utilize price and production indexes to formulate future budgets and plans.

8. Formulating Government Policy:

Governments utilize index statistics to formulate monetary, fiscal, and trade policies in accordance with economic trends. For instance, if the Wholesale Price Index indicates elevated inflation, the central bank may raise interest rates. Subsidies or taxes may be modified according to price or output indices in the agricultural or petroleum sectors.

Problems in construction of index numbers

(1) Definition of Purpose and Scope: Before going to construct an index number, a clear statement as to the purpose and its scope is necessary. All index numbers do not serve the same purpose, and there is no all-purpose index. The selection of items, etc., will depend upon the purpose of construction and the people for whom it is intended. For example, in constructing an index number of wholesale prices, the prices from retailers are unnecessary, just as for a cost of living index number, quotations of cloth price ex-mill or prices of cotton yarn are useless. One must be sure of what the index number is going to measure.

(2) Selection of Items: For reasons of economy and ease of calculation, it is not possible to include all commodities in the construction of an index number. For a price index number, only a few selected items are, therefore, included whose price movements appear to be representative of the whole group of commodities. On the other hand, inclusion of too few items would make the index unrepresentative of the general level. With the passage of time, some items lose importance while some other

Before going to construct an index number, a clear statement as to the purpose and its scope is necessary.

new items appear to be more useful. We should then delete the less important items from the list of commodities and replace them with new ones that align with their relative importance.

(3) Selection of Sources and Collection of Data: For a regular source of index numbers, a systematic collection of prices and quantities should be made at regular intervals of time from prominent business firms or standard retail stores located at different important centers. A large majority of customers should visit the selected shops. Due care must also be taken in selecting the enumerators, who are entrusted with the collection of data, because upon their honesty and intelligence will depend the quality and reliability of index numbers.

The base period should be chosen with much care and be one when no abnormal increase or decrease in price was noticed.

(4) Choice of Base: The base period should be chosen with much care and be one when no abnormal increase or decrease in price was noticed. Selecting a base period that is recent is desirable. The base should not be too long or too short a period. Generally, a year is taken as the base, preferably a year of some economic importance for the country.

(5) System of Weighting: The commodities used in the formulation of an index number do not possess equal significance, as a price alteration in one item does not influence the price level to the same degree as an equivalent modification in another item. The weighting system, especially the distribution of weights among various things, is therefore of paramount importance. Values influence relationships, quantities determine prices, and prices govern both quantities and price relationships. The prices or amounts employed as weights may pertain to either the base era or the current period. Laspeyres' price index employs base period values as weights, whereas Paasche's price index utilizes current period amounts as weights.

6. Form of Average to Use: Price index numbers are sometimes computed by averaging the percentage positions in price of the commodities. Generally, we use either the arithmetic mean or the geometric mean for averaging. Occasionally, we also use the median. The arithmetic mean, due to the simplicity in calculation, is used in a great majority of cases, but since it is highly affected by even a few very large or small values, the geometric mean is preferred in many cases.

Self-Assessment Questions:

Short Questions

1. What is an index number?
2. What are the main uses of index numbers?
3. Why are index numbers important in statistics?
4. What are the different types of index numbers?
5. What is the base year in the context of index numbers?
6. How is a price index calculated?
7. What does the term "weighted index" mean?
8. What are the problems in construction of using index numbers?
9. How does an index number help in measuring economic performance?
10. What is the difference between a simple index and a weighted index?

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) What constitutes an index number?
 - a) A numerical representation denoting absolute prices
 - b) A graphical representation of temporal variation
 - c) A proportion of Gross Domestic Product
 - d) A financial valuation of output
- (ii) What is the constant value of the index number for the base period?
 - a) Zero
 - b) One
 - c) One hundred
 - d) Ten
- (iii) Which of the following is NOT a category of index number?
 - a) Price Index
 - b) Quantity Index
 - c) Volume Index
 - d) Value Index
- (iv) Who was among the pioneers in utilizing index numbers for pricing comparison?
 - a) Adam Smith
 - b) William Fleetwood
 - c) Joseph Lowe
 - d) Étienne Laspeyres
- (v) What is a purpose of index numbers?
 - a) Generating raw data
 - b) Forecasting meteorological conditions
 - c) Monitoring economic changes
 - d) Archiving extensive datasets

2. Write "T" if the statement is true and "F" if the statement is false:

- (i) A value of 100 is always used for the base period index.
- (ii) Quantity index figures show how prices have changed through time.
- (iii) index numbers assist remove the effect of inflation from economic statistics.
- (iv) Index numbers are based on homogeneous units only.
- (v) Cost of Living Index does not consider consumer behavior.

Answer:

Multiple-Choice Question: 1. (i) b (ii) c (iii) c (iv) b (v) c

True/False: 2. (i) T (ii) F (iii) T (iv) F (v) F

Lesson 2: Methods of Construction of Index Numbers

Lesson Objectives:

After completing this lesson, you will be able to

- Understand the methods of construction of index numbers;
- Apply the methods of construction of index numbers;
- Explain the Importance and use of weights in the construction of index numbers;
- Explain the Advantages of geometric mean in the construction of index numbers.

Introduction:

Index numbers, as we discussed in the earlier lesson, are statistical tools that quantify how the magnitude of a group of connected variables changes over time, space, or other variables. Frequently presented as percentages, they function as a comparative metric to describe and compare average shifts in the cost, quantity, value, or other characteristics of products or groups of products. A statistical instrument for tracking changes in variables over time, such as prices, quantities, or values, is an index number. A statistical instrument for tracking variations in a phenomenon's level over time is an index number. Businesses and economists frequently use them to monitor shifts in costs, quantities, or values. Index numbers include, for example, industrial production indices, cost of living indices, and consumer price indices. The creation of index numbers necessitates the proper selection of data and the use of suitable techniques because they are dependent on comparisons between two points in time, the base period and the current period. Choosing a base period and applying certain algorithms to compare it with one or more current periods is the process of creating index numbers. An index number's construction process has a substantial impact on its correctness and usefulness. The type of data, the index's goal, and the degree of accuracy needed all influence the best approach. We will examine the main techniques for creating index numbers in the ensuing sections, as well as their formulas and uses.

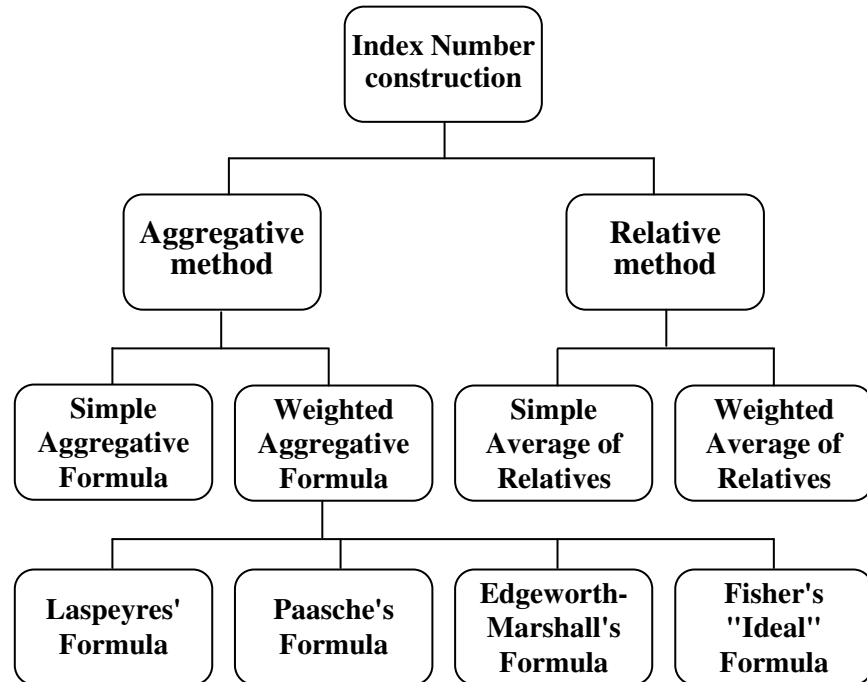
A statistical instrument for tracking changes in variables over time, such as prices, quantities, or values, is an index number.

Methods of construction of index numbers

Index numbers are statistical tools used to measure and compare changes in variables such as prices, quantities, or values over time. Two fundamental approaches to constructing index numbers are the Aggregative Method and the Relative Method. These methods differ in their calculation techniques but serve the same purpose of tracking changes over time.

The Aggregative Method involves adding up the prices (or quantities) of selected items in the current year and the base year, then comparing the totals. It gives an overall index by directly comparing total values.

The Relative Method involves computing the price relatives (or quantity relatives) for each item and then averaging them. A price relative is the ratio of the current price to the base price, expressed as a percentage.



Here, we shall discuss the methods of construction in relation to a "Price Index". The following notations are used:

(1) p_o, p_n denote price per unit in base year and current year respectively.

q_o, q_n „ quantity „ „ „ „ „

However, when data for several years are available, price (p) and quantity (q) are used with subscripts 0, 1, 2, 3, etc. Thus,

(2) $p_o, p_1, p_2, \dots$ denote price in the years 0, 1, 2, ... respectively;

$q_o, q_1, q_2, \dots$ denote quantity

The letter I, with appropriate subscripts, is used to denote "index number".

(3) I_{on} denotes index number for year n with base year 0

I_{no}	"	"	"	0	"	"
I_{23}	"	"	"	3	"	2

(The first subscript indicates the base and the second subscript the current year).

In order to devise formulae for index numbers, let us consider the question raised at the beginning of this unit. Our problem is to find a quantity which will indicate a comparison of the price of food in general in 2020 as against that in 2010. This can be done in two ways: -

(a) We may compare the 'average price per unit of a commodity in 2020 against the same in 2010.

Table 8.2-Construction of Price Index				
Commodity	p_o	p_n	$\frac{p_n}{p_o}$	Price Relative
Rice	80.00	200.00	2.5	250
Wheat	0.50	1.40	2.8	280
Fish	10.00	23.00	2.3	230
Bread	0.60	1.35	2.25	225
Milk	1.50	3.00	2.0	200
Total	92.60	228.75	11.85	1185

Index Number for 2020 (Base 2010 = 100) :

$$\begin{aligned}
 &= \frac{\text{Average price per unit in 1980}}{\text{Average price per unit in 1970}} \times 100 \\
 &= \frac{228.75/5}{92.60/5} \times 100 = \frac{228.75}{92.60} \times 100 = 247
 \end{aligned}$$

The index number is thus a ratio of 'aggregate prices', expressed as a percentage. This method is known as "Aggregative Method".

(b) We may also compare the prices of each commodity individually between the two periods and then find an average. For instance, Rice price is 2.5 times, Wheat price 2.8 times, Fish 2.3 times, Bread 2.25 times, Milk 2 times, so that on an average this comes to $11.85/5 = 2.37$ times, i.e. 237%. With base year 2010 taken as 100, the index number for 2020 is thus 237.

Index Number for 2020 Average of Price Relatives $1185 \div 5 = 237$

(Price Relative "Relative Method". Price ratio $\times 100$). This method is known as the "Relative Method".

It may be noted that

- (i) Aggregative Method shows "Relative of averages (or aggregates)"
- (ii) Relative Method shows "Average of relatives".

The average used may however be either 'simple' or 'weighted'. Thus, we have Simple Aggregative or Weighted Aggregative Index and Simple Average or Weighted Average of Relatives Index.

I. Aggregative Method:

In this method, the aggregate price of all items in the given year is expressed as a percentage of the same in the base year, giving the index number.

$$\text{Index number} = \frac{\text{Aggregate Price in the given year}}{\text{Aggregate Price in the base year}} \times 100$$

If simple aggregates of prices are compared, we get

$$\text{Simple Aggregative Index} \quad I_{on} = \frac{\sum p_n}{\sum p_o} \times 100$$

the summation extending over all items included for the construction of index number.

If, however, weighted aggregates of prices in the two periods are compared, we have

$$\text{Weighted Aggregative Index} \quad I_{on} = \frac{\sum p_n w}{\sum p_0 w} \times 100$$

where w represents the "weight". It should be noted that the same set of weights must be used both for base year as well as for current year.

In the construction of a Price Index, quantities (q) are used as weights. There are several formulae for weighted aggregative index depending on the nature of weights employed:

- (i) If the base year quantity (q_0) is used as weight, i.e. $w = q_0$ we get :

$$\text{Laspeyres' Index} \quad I_{on} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

- (ii) If the current year quantity (q_n) is used as weight, i.e. $w = q_n$

$$\text{we get Paasche's Index} \quad I_{on} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

- (iii) If the sum of quantities in the base year and the current year is used as weight, i.e. $w = (q_0 + q_n)$ we get :

$$\text{Edgeworth-Marshall's Index} \quad I_{on} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100$$

- (iv) The geometric mean (i.e. square-root of the product) of Laspeyres' Index and Paasche's index is of special importance, because of certain properties (Example 17:28), and is known as

Fisher's Ideal Index

$$\begin{aligned} I_{on} &= \sqrt{(\text{Laspeyres' Index}) \times (\text{Paasche's Index})} \\ &= \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} \times 100 \end{aligned}$$

The following index numbers of the weighted aggregative type are also sometimes used: -

- (v) The arithmetic mean of Laspeyres' index and Paasche's index is known as

Bowley's Index (I_{on})

$$= \frac{1}{2} (\text{Laspeyres' index} + \text{Paasche's index})$$

$$= \frac{1}{2} \left[\frac{\sum p_n q_0}{\sum p_0 q_0} + \frac{\sum p_n q_n}{\sum p_0 q_n} \right] \times 100$$

- (vi) If the geometric mean of base year and current year quantities is used as weight, i.e. **Walsh's Index** $w = \sqrt{q_0 q_n}$, we get

$$I_{on} = \frac{\sum p_n \sqrt{q_0 q_n}}{\sum p_0 \sqrt{q_0 q_n}} \times 100$$

- (vii) If the weights used are kept fixed for all periods, i.e. weights are constant quantities (q), without any reference to base or current period, we get :

$$\text{Kelly's Index } I_{on} = \frac{\sum p_n q}{\sum p_0 q} \times 100$$

This is also known as "Aggregative index with fixed weights".

II. Relative Method

In this method, the price of each item in the current year is expressed as a percentage of the price in the base year. This is called Price Relative and is given by the formula

$$\begin{aligned} \text{Price Relative} &= \frac{\text{Price in the given year}}{\text{Price in the base year}} \times 100 \\ &= \frac{P_n}{P_o} \times 100 \end{aligned}$$

The average of price relatives, which shows the average percentage change for the whole group of items, gives the index number.

$$\text{Price Index} = \text{Average of Price Relatives}$$

Usually, A.M. or G.M. is used for averaging the relatives. In special cases, H.M., or median, is also used. Again, the average employed may be either 'simple' or 'weighted'. If a simple average is used, the index number is called Simple Average of Relatives Index. If a weighted average is used, it is known as Weighted Average of Relatives Index. Thus,

$$\text{Simple A. M. of Relatives Index } (I_{on}) = \sum(\text{Price Relatives}) \div k$$

where k is the number of items included.

Simple G. M. of Relatives Index

$$(I_{on}) = \sqrt[k]{\text{Product of price relatives}}$$

Weighted A. M. of Relatives Index

$$(I_{on}) = \frac{\sum (\text{Price Relative}) \times w}{\sum w}$$

The weights (w) employed for averaging price relatives are the values (= price × quantity) of items. In most cases, these values are given not in absolute units, but as percentages of the total value for all the items, i.e. the weights are given as pure numbers [see (8.8.1)]

Aggregative Formulae by Relative Method

It is interesting to note that the weighted average of relatives leads to several index number formulae of the aggregative type, depending on the nature of weights used. Considering price index numbers

- (1) The A.M. of relatives formula weighted by base year values ($p_0 q_0$) gives exactly the same formula as Laspeyres':

$$I_{on} = \frac{\sum \left(\frac{p_n}{p_0} \times 100 \right) p_0 q_0}{\sum p_0 q_0} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100 = \text{Laspeyres' index}$$

- (2) The A.M. of relatives' formula weighted by values of current year quantities at base year prices ($p_0 q_n$) gives Paasche's formula:

$$I_{on} = \frac{\sum \left(\frac{p_n}{p_0} \times 100 \right) p_0 q_n}{\sum p_0 q_n} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 = \text{Paasche's index}$$

- (3) The H.M. of relatives' formula weighted by current year values (pay) gives the same formula as Paasche's:

$$I_{on} = \frac{\sum p_n q_n}{\sum p_n q_n / \left(\frac{p_n}{p_0} \times 100 \right)} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 = \text{Paasche's index}$$

Construction of General Index from Group Indices

In the construction of any index number, the items included are usually classified under some broad categories called Groups, with similar or related items coming under each group. A separate index number is constructed for each group, and is called Group Index. The weighted average (usually A.M.) of group index numbers gives the General Index.

$$\text{General Index} = \frac{\sum IW}{\sum W}$$

where I represents the Group Index and W is the Group Weight.

Example 8:5 Find Index Numbers by the (a) method of aggregates, and (b) method of relatives (using arithmetic mean), from the following:

Commodity	Base Price	Current Price
Rice	35	42
Wheat	30	35
Pulse	40	38
Fish	107	120

Solution: (a) Let p_0, p_n denote the Base Price and Current Price.

Table 8.3-Calculations for Index Number

Commodity	p_0	p_n	Price Relative = $\frac{p_n}{p_0} \times 100$
Rice	35	42	$\frac{42}{35} \times 100 = 120.0$
Wheat	30	35	$\frac{35}{30} \times 100 = 116.7$
Pulse	40	38	$\frac{38}{40} \times 100 = 95.0$
Fish	107	120	$\frac{120}{107} \times 100 = 112.1$
Total	212	235	443.8

- (a) Simple Aggregative Index:

$$I_{on} = \frac{\sum p_n}{\sum p_0} \times 100 = \frac{235}{212} \times 100 = 110.8$$

- (b) Simple A.M. of Price Relatives Index:

$$I_{on} = \frac{\sum \text{Price Relative}}{k} = \frac{443.8}{4} = 111.0$$

It should be remembered that at index number compares current price as a percentage of base price. Since a number of commodities are to be

covered, a measure of this comparison can be given either (i) as the percentage of average prices per commodity (the same as percentage of aggregate prices), or (ii) as the average of percentage prices for each commodity. The former gives Aggregate Index. and the latter gives A.M. of Relatives Index.

Example 8:6 Calculate price index numbers from the following data, using (i) weighted aggregative formula, and (ii) weighted arithmetic mean of price relatives' formula:

Commodity	Unit	Price (Rs.) per unit		Weight
		Base period	Current period	
A	Quintal	80	110	14
B	Kg.	10	15	20
C	Dozen	40	56	35
D	Litre	50	95	15
E	lb.	12	18	16

Solution: Note: Compare this with Example 8: 5, where no weights are given].

Table 8.4: Calculations for Index Numbers

Commodity	Base price (p_0)	Current price (p_n)	Weight (w)	$p_0 w$	$p_n w$	Price* Relative(I)	IW
A	80	110	14	1120	1540	137.5	1925
B	10	15	20	200	300	150	3000
C	40	56	35	1400	1960	140	4900
D	50	95	15	750	1425	190	2850
E	12	18	16	192	288	150	2400
Total	-	-	100	3662	5513	-	15075

[*Note: Price Relative ($\frac{p_n}{p_0}$) $\times 100$. For example, $(110 \div 80) \times 100 = 137.5$ $(15 \div 10) \times 100 = 150$ etc.]

(i) Weighted Aggregative Index :

$$= \frac{\sum P_n W}{\sum P_0 W} \times 100 = 5513 \div 3662 \times 100 = 150.55$$

(ii) Weighted Arithmetic Mean of Price Relatives Index:

$$= \frac{\sum Iw}{\sum w} = \frac{15075}{100} = 150.75$$

Example 8:7 From the following price and quantity data, compute Paasche's price index number for 2020 with 2010 as base:

	Price (Tk . Per kg.)		Quantities Sold (kg.)	
	2010	2020	2010	2020
Commodity A:	4	5	95	120
Commodity B:	60	70	118	130
Commodity C:	35	40	50	70

Solution: Paasche's Price Index is obtained by weighted aggregative formula with current year quantities as weight.

Table 8.5 Calculations for Paasche's Price Index

Commodity	p_0	p_n	q_0	q_n	p_0q_n	p_nq_n
A	4	5	95	120	480	600
B	60	70	118	130	7800	9100
C	35	40	50	70	2450	2800
Total	-	-	-	-	10730	12500

$$\text{Paasche's Price Index} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 = \frac{12500}{10730} \times 100 = 116$$

Example 8:8 Construct Fisher's ideal index number for the following data:

	2010(Base year)		2018 (Current year)	
Commodity	Price	Quantity	Price	Quantity
A	8	6	12	5
B	10	5	11	6
C	7	8	8	5

Solution: Fisher's Ideal Index = $\sqrt{\text{Laspeyres' index} \times \text{Paasche's index}}$

Table 8.6 Calculations for Fisher's Ideal Index

Commodity	p_0	q_0	p_n	q_n	p_0q_0	p_nq_0	p_0q_n	p_nq_n
A	8	6	12	5	48	72	40	60
B	10	5	11	6	50	55	60	66
C	7	8	8	5	56	64	35	40
Total	-	-	-	-	154	191	135	166

$$\text{Laspeyres' index} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100 = \frac{191}{154} \times 100 = 124.0$$

$$\text{Paasche's index} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 = \frac{166}{135} \times 100 = 123.0$$

$$\therefore \text{Fisher's Ideal index} = \sqrt{(124 \times 123)} = 123.5$$

Example 8:9 "Marshall-Edgeworth index number is a good approximation to the Fisher's Ideal Index Number"-Verify the truth of this statement from the following data:

	RICE		WHEAT		JOWAR	
Year	Price	Quantity	Price	Quantity	Price	Quantity
2010	9.3	100	6.4	11	5.1	5
2017	4.5	90	3.7	10	2.7	3

Solution: Let us take 2010 as Base and 2017 as Current year. Marshall-Edgeworth Price Index

$$= \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)} \times 100 = \frac{\sum p_n q_0 + \sum p_n q_n}{\sum p_0 q_0 + \sum p_0 q_n} \times 100$$

$$\text{Fisher's Ideal Price Index} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} \times 100$$

Table 8.7-Marshall-Edgeworth's and Fisher's Price Index

Commodity	p_0	p_n	q_0	q_n	p_0q_0	p_0q_n	p_nq_0	p_nq_n
Rice	9.3	4.5	100	90	930	837	450	405
Wheat	6.4	3.7	11	10	70.4	64	40.7	37
Jowar	5.1	2.7	5	3	25.5	15.3	13.5	8.1
Total	-	-	-	-	1025.9	916.3	504.2	450.1

$$\text{Marshall - Edgeworth Price Index} = \frac{504.2 + 450.1}{1025.9 + 916.3} \times 100 = 49.135.$$

$$\text{Fisher's Ideal Price Index} = \sqrt{\frac{504.2}{1025.9} \times \frac{450.1}{916.3}} \times 100 = 49.134.$$

The two index numbers are very close to each other. The statement is thus verified.

Example 8:10 Calculate the price index number for the year 2008 with 2006 as base using Laspeyres' or Paasche's formula, which-ever will be applicable, on the basis of the following data:

Commodity	Price (in Tk.)		Money value ('000 Tk.)
	2006	2008	2006
A	12.50	14.00	112.50
B	10.50	12.00	126.00
C	15.00	14.00	105.00
D	9.40	11.20	47.00

(Here money value means total value of a commodity).

Solution: We are given p_0, p_n and p_0q_0 (i.e. value in "base" year), from which it is possible to find q_0 by relation

$$q_0 = \frac{p_0q_0}{p_0} = \frac{\text{Money value in 1976 ('000 Tk.)}}{\text{Price in 1976 (Tk.)}}$$

in units of '000. Now using p_0, p_n and q_0 we can only find

$$\text{Laspeyres' Price Index} = \frac{\sum p_nq_0}{\sum p_0q_0} \times 100$$

(Note that it is not possible to find Paasche's Price Index, because the formula involves q_n which is not available from the given data)

Table 8.8 Calculations for Laspeyres' Price Index

Commodity	p_0	p_n	p_0q_0	$q_0 = 4 \div 2$	p_nq_0
(1)	(2)	(3)	(4)	(5)	(6)
A	12.50	14.00	112.50	9	126.00
B	10.50	12.00	126.00	12	144.00
C	15.00	14.00	105.00	7	98.00
D	9.40	11.20	47.00	5	56.00
Total	-	-	390.50	-	424.00

$$\text{Laspeyres' Price Index} = \frac{\sum p_nq_0}{\sum p_0q_0} \times 100 = \frac{424.00}{390.50} \times 100 = 109$$

Example 8:11 With regard to Laspeyres's and Paasche's price index numbers, it is maintained that "if the prices of all the goods change in the same ratio, the two indexes will be equal, for then the weighting system is irrelevant; or, if the quantities of all the goods change in the same ratio, they will be equal, for then the two weighting systems are the same relatively."

Required:

Verify the above statement.

Solution:

$$\text{Laspeyres' Price Index (L)} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

$$\text{Paasche's Price Index (P)} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

- (i) If all prices change in the same ratio, we have $p_n = k \cdot p_0$ where k is a constant. Then :

$$P = \frac{\sum (k \cdot p_0) q_n}{\sum p_0 q_n} \times 100 = \frac{k \sum p_0 q_n}{\sum p_0 q_n} \times 100 = 100k$$

Thus, we find that $L = P$

- (ii) If all quantities change in the same ratio, we put $q_n = k' \cdot q_0$, where k' is a constant. Then :

$$L = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

$$P = \frac{\sum p_n (k' \cdot q_0)}{\sum p_0 (k' \cdot q_0)} \times 100 = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100; (\text{because } k' \text{ cancels})$$

Again we find that $L = P$ the statements are thus verified.

Example 8:12 Given below are the data on prices of some consumer goods and the weights attached to the various items. Compute price index numbers for the year 2019 (Base: 2018 = 100), using (i) simple average, and (ii) weighted average, of price relatives.

Item	Unit	Price (Tk.)		Weight
		2018	2019	
Wheat	Kg.	0.50	0.75	2
Milk	Litre	0.60	0.75	5
Egg	Dozen	2.00	2.40	4
Sugar	Kg.	1.80	2.10	8
Shoes	Pair	8.00	10.00	1

Solution:**Table 8.9-Calculations for Price Relatives Index**

Item	p_0	p_n	Price Relative $I = \frac{p_n}{p_0} \times 100$	Weight (w)	Iw
Wheat	0.50	0.75	150	2	300
Milk	0.60	0.75	125	5	625
Egg	2.00	2.40	120	4	480
Sugar	1.80	2.10	117	8	936
Shoes	8.00	10.00	125	1	125
Total	-	-	637	20	2466

$$\text{Simple Average of P.R. Index} = \frac{\sum(\text{Price Relative})}{\text{No. of items}} = \frac{637}{5} = 127.4$$

$$\begin{aligned} \text{Weighted Average of P R. Index} &= \frac{\sum (\text{Price relative} \times \text{Weight})}{\sum \text{Weight}} \\ &= \frac{2466}{20} = 123.3 \end{aligned}$$

Example 8:13 On the basis of the following data, compute the wholesale price index number for the 5 groups combined:

Group	Weight(w)	Index number for week ending 27.9.69 (Base: 2022-23 = 100)
Food articles	50	241
Liquor and tobacco	2	221
Fuel, power, light and lubricants	3	204
Industrial raw materials.	16	256
Manufactured commodities	29	179

Solution: Using the weighted arithmetic mean of group indices as the method of combination, the General Index is given by the formula:

$$\text{General Index} = \frac{\sum IW}{\sum W}$$

where I = Group Index, and W = Group Weight.

Table 8.10-General Index from Group Indices

Group	Weight (W)	Group Index(I)	IW
1. Food articles	50	241	12,050
2. Liquor & tobacco	2	221	442
3. Fuel, power, light & lubricants	3	204	612
4. Industrial raw materials	16	256	4,096
5. Manufactured commodities	29	179	5,191
Total	100	-	22,391

$$\therefore \text{Index Number of Wholesale Prices} = \frac{22391}{100} = 223.91$$

Example 8:14 Apply the geometric mean to find general index from the following group indices, by assigning the given weights:

Group	A	B	C	D	E	F
Group Index	118	120	97	107	111	93
Weight	4	1	2	6	5	2

Solution: The weighted geometric mean of the group indices will be found by applying logarithms:

$$\log (\text{General Index}) = \frac{\Sigma(\log I) \times W}{\Sigma W}$$

Table 8.11-General Index using G.M.

Group	Group index (I)	Weight (W)	$\log I$	$\Sigma(\log I) \times W$
A	118	4	2.0719	8.2876
B	120	1	2.0792	2.0792
C	97	2	1.9868	3.9736
D	107	6	2.0294	12.1764
E	111	5	2.0453	10.2265
F	93	2	1.9685	3.9370
Total	-	20	-	40.6803

Substituting the values from the table,

$$\log (\text{General Index}) 40.6803 \div 20 = 2.0340$$

$$\therefore \text{General Index} = \text{antilog } 2.0340 = 108.1$$

Importance and use of weights in the construction of index numbers

Weights play a very important part in the construction of index numbers. Index numbers of price are calculated either by taking the average of price relatives or by taking the relative of average prices of the items at two periods of time. In either case, the averaging process is involved, and naturally the question arises whether it should be a simple average or a weighted average. If a simple average is used, it will be assumed that all the items included are equally important. But in almost all cases this cannot be so. All items cannot be considered as equally important in the sense that a change in the price of one of the items does not affect the price level to the same extent as does the same amount of change in the price of another item. For instance, in constructing a wholesale price index number, textiles must have greater weight than tobacco. If we ignore weights, we shall not get an unweighted index but an inappropriately weighted index.

Since index numbers should not depend on the units in which the prices or quantities are reported, price relatives are weighted by 'values' (= price quantity), prices by quantities and quantities by prices. The quantity or value used as weight may relate either to the base period or to the current

Index numbers of price are calculated either by taking the average of price relatives or by taking the relative of average prices of the items at two periods of time.

period or to a combination of both. The weights used in some price index formulae are :-

- (i) *Laspeyres' Index* – Base period quantity ($w = q_0$),
- (ii) *Paasche's Index* Current period quantity ($w = q_n$).
- (iii) *Dgeworth–Marshall's Index* Sum of base and current period quantities ($w = q_0 + q_n$).
- (iv) Simple A.M. of Price Relatives Index-Number of units of the commodity, that can be purchased in the base period by one unit of money ($w = 1/p_0$).
- (v) Weighted A.M. of Price Relatives Index-Usually, the base period values are used as weights ($w = p_0 q_0$) leading to Laspeyres' index. This has the advantage that the same set of weights calculated from base year data can be used for a long period of time. If the values of current period quantities at base year prices are used as weights ($w = p_0 q_0$), Paasche's index is obtained.

Again, the quantity or value used as weight need not necessarily be the actual physical quantities or values produced or consumed, but their relative magnitudes. Weights are, therefore, as a rule expressed as percentages of total, which is taken as 100.

Advantages of geometric mean in the construction of index numbers

Index numbers are designed to measure the 'average' level of any particular factor (e.g. price, price, quantity or value) value) from one period to another. Naturally, the question arises as to which average to use. For reasons of simplicity in calculation, the arithmetic mean is used in a great majority of cases. But the geometric mean (G.M.) has definite advantages from several standpoints:

- (i) The G.M. is useful in averaging ratios, rates and percentages. It is particularly suitable for the construction of index numbers; because index numbers show percentage changes, rather than absolute amounts of change. It also gives equal weight to equal ratios of change.
- (ii) Again since the G.M. is less affected than the arithmetic mean by the presence of extremely large or small values, it is considered all the more appropriate in index number construction. An unusual change in the price of a single commodity should not upset the whole index number.
- (iv) The G.M. also makes index numbers time-reversible. While the arithmetic mean of relatives index does not satisfy time reversal test, the g.m. of relatives index satisfies this test. Laspeyres' and Paasche's index numbers do not satisfy either the time reversal or the factor reversal test, but their G.M., viz. Fisher's ideal index number, satisfies both these tests, and as such is considered "ideal" from theoretical considerations.

Index numbers are designed to measure the 'average' level of any particular factor (e.g. price, price, quantity or value) value) from one period to another.

Self-Assessment Questions:

Short Questions

1. What is an index number?
2. What are the main uses of index numbers?
3. Why are index numbers important in statistics?
4. What are the different types of index numbers?
5. What is the base year in the context of index numbers?
6. How is a price index calculated?
7. What does the term "weighted index" mean?
8. What are the limitations of using index numbers?
9. How does an index number help in measuring economic performance?
10. What is the difference between a simple index and a weighted index?

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which method of index number uses base year quantities as weights?
a) Paasche's Method b) Laspeyres Method
c) Fisher's Method d) Marshall-Edgeworth Method
- (ii) Which index number method uses both base year and current year prices and quantities?
a) Laspeyres Method b) Paasche's Method
c) Fisher's Ideal Method d) Simple Aggregative Method
- (iii) Which of the following is NOT a purpose of index numbers?
a) Measuring inflation
b) Forecasting future trends
c) Comparing economic development across countries
d) Predicting individual behavior
- (iv) Which type of average is generally used in constructing Fisher's Ideal Index?
a) Harmonic Mean b) Geometric Mean
c) Arithmetic Mean d) Weighted Mean
- (v) The main difference between Laspeyres and Paasche methods is in the:
a) Use of averages b) Use of price weights
c) Use of quantity weights d) Formula type

2. Write "T" if the statement is true and "F" if the statement is false:

- (i) Index numbers measure relative changes over time.
- (ii) Laspeyres Index uses current year quantities as weights.
- (iii) Paasche's method uses current year prices and quantities.
- (iv) Fisher's Ideal Index is the arithmetic mean of Laspeyres and Paasche indices.
- (v) An index number of 120 means a 20% increase over the base year.

Answer:

Multiple-Choice Question: 1. (i)- b (ii)- c (iii)- d (iv)- b (v)- c

True/False: 2. (i)- T (ii)- F (iii)- T (iv)- F (v)- T

Lesson 3: Quantity Index Number, Tests of Index Numbers and Other Index Methods

Lesson Objectives:

After completing this lesson, you will be able to;

- Understand the meaning and methods of quantity index numbers;
- Apply the various methods of test of index number;
- Understand the Chain base method and its Advantages and Disadvantages;
- Understand concept of Cost of living index numbers, its Construction procedures and determination of Weights;
- Understand bias in Laspeyres and Paasche's formulae for Cost of Living Index;
- Meaning and application base shifting, Splicing and deflation;
- Understand and explain the errors in Index Numbers.

Introduction:

In business and economics and, Quantity Index Numbers serve as a crucial instrument for assessing variations in the physical volume or quantity of products and services across time. In contrast to price index numbers, which concentrate on fluctuations in commodity prices, quantity index numbers assess the variations in the actual quantity of products produced, consumed, or sold in relation to a base period. These indices facilitate the assessment of fluctuations in production or consumption, irrespective of price variations. They are extensively utilized in industrial output assessment, agricultural production evaluation, foreign trade analysis, and national income accounting. For instance, if a nation seeks to evaluate its agricultural output in the current year relative to a base year, it would employ a quantity index number.

Price index numbers, which concentrate on fluctuations in commodity prices, quantity index numbers assess the variations in the actual quantity of products produced, consumed, or sold in relation to a base period.

In other words, it facilitates the analysis of fluctuations in unit quantity over time, assuming price constancy. It is represented as an index, with:

- The base period often assigned a value of 100.
- An index beyond 100 indicates an increase in amount.
- An index below 100 signifies a drop in quantity.

Calculations of Quantity Index

Just as price index numbers measure and permit comparison of the price of a group of related items, quantity index numbers similarly measure and permit comparison of the physical quantity of goods produced or consumed or marketed or distributed. On the other hand, a Quantity Index Number is a statistical measure used to show how the physical quantity or volume of goods or services has changed over time compared to a base period. It is used to track actual changes in output, sales, consumption, or production, without the influence of price changes.

Index Number is a statistical measure used to show how the physical quantity or volume of goods or services has changed over time compared to a base period.

Quantity index number formulae may be obtained from the corresponding price index number formulae replacing p by q , and q by p .

$$\text{Simple Aggregative Quantity Index} = \frac{\sum q_n}{\sum q_0} \times 100$$

$$\text{Laspeyres' Quantity Index} = \frac{\sum q_n p_0}{\sum q_0 p_0} \times 100$$

$$\text{Paasche's Quantity Index} = \frac{\sum q_n p_n}{\sum q_0 p_n} \times 100$$

$$\text{Edgeworth – Marshall's Index} = \frac{\sum q_n (p_0 + p_n)}{\sum q_0 (p_0 + p_n)} \times 100$$

$$\text{Fisher's Ideal Index} = \sqrt{\frac{\sum q_n p_0}{\sum q_0 p_0} \times \frac{\sum q_n p_n}{\sum q_0 p_n}} \times 100$$

$$\text{Quantity Relative} = \frac{q_n}{q_0} \times 100$$

$$\begin{aligned} \text{Simple A.M. of Quantity Relatives Index} \\ = \sum (\text{Quantity Relatives}) \div k \end{aligned}$$

$$\begin{aligned} \text{Weighted A.M. of Quantity Relatives Index} \\ = \frac{\sum (\text{Quantity relative Weight})}{\sum (\text{Weight})} \end{aligned}$$

Example 8:15 Prepare price and quantity index numbers for 2022 with 2001 as base year from the following data by using (i) Laspeyre's, (ii) Paasche's and (iii) Fisher's method.

Commodity	Unit	2001		2022	
		Quantity	Price (Tk.)	Quantity	Price (Tk.)
A	Kg.	5	2.00	7	4.50
B	Quintal	7	2.50	10	3.20
C	Dozen	6	8.00	6	4.50
D	Kg.	2	1.00	9	1.80

Solution:

Table 8.12-Laspeyres', Paasche's and Fisher's Index

Commodity	q_0	p_0	q_n	p_n	$p_0 q_0$	$p_0 q_n$	$p_n q_0$	$p_n q_n$
A	5	2.00	7	4.50	10.00	22.50	14.00	31.50
B	7	2.50	10	3.20	17.50	22.40	25.00	32.00
C	6	8.00	6	4.50	48.00	27.00	48.00	27.00
D	2	1.00	9	1.80	2.00	3.60	9.00	16.20
Total	-	-	-	-	77.50	75.50	96.00	106.70

(a) Price Index Numbers :-

$$\text{Laspeyres' formula} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100 = \frac{75.50}{77.50} \times 100 = 97$$

$$\text{Paasche's formula} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 = \frac{106.70}{96.00} \times 100 = 111$$

$$\text{Fisher's formula} = \sqrt{\text{Laspeyres'} \times \text{Paasche's}} = \sqrt{97 \times 111} = 104$$

(b) Quantity Index Numbers:-

$$\text{Laspeyres' formula} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100 = \frac{96.00}{77.50} \times 100 = 124$$

$$\text{Paasche's formula 1} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100 = \frac{106.70}{75.50} \times 100 = 141$$

$$\text{Fisher's formula} = \sqrt{\text{Laspeyres'} \times \text{Paasche's}} = \sqrt{(124 \times 141)} = 132$$

Example 8:16 Annual production (in million tons) of four commodities is given below:

Commodity		Production in year			
		2015	2019	2020	Weights
A	---	160	200	216	20
B	---	24	42	45	30
C	---	50	72	68	13
D	---	120	168	156	17

Required:

Calculate quantity index numbers for the years 2015 and 2020 with 2015 as base year, using (i) simple arithmetic mean, and (ii) weighted arithmetic mean, of the relatives.

Solution: [Working Notes:

Quantity Relatives for 2019 (Base 2015)

$$\frac{300}{160} \times 100 = 125; \frac{42}{24} \times 100 = 175; \frac{75}{50} \times 100 = 144; \frac{160}{120} \times 100 = 140$$

Quantity Relatives for 1955 (Base 1950)

$$\frac{216}{160} \times 100 = 135; \frac{45}{24} \times 100 = 187.5; \frac{68}{50} \times 100 = 136; \frac{156}{120} \times 100 = 130]$$

Table 8.13-Calculations for Quantity Index Number

Commodity	Quantity 2010	Relatives 2020	Weight	(Q.R. for 2010) X (weight)	(Q.R. for 2020) X (weight)
A	125	135	20	2500	2700
B	175	187.5	30	5250	5625
C	144	136	13	1872	1768
D	140	130	17	2380	2210
Total	584	588.5	80	12002	12303

Using simple arithmetic mean of quantity relatives,

$$\text{Index Number for 1954} = 584 \div 4 = 146$$

$$\text{Index Number for 1955} = 588.5 \div 4 = 147$$

Using weighted arithmetic mean of quantity relatives,

$$\text{Index Number for 1954} = 12002 \div 80 = 150$$

$$\text{Index Number for 1955} = 12303 \div 80 = 154$$

Example 8.17 Calculate a number which will indicate the percentage change in volume of traffic (Oct. 2020 = 100) from October 2020 to October 2021, when account is taken of the relative values of the different types of traffic.

Type of traffic	Tons ('000)		Receipts (£ '000)
	Oct. 2020	Oct. 2021	Oct. 2020
(a) Merchandise	1246	1206	776
(b) Minerals	1125	981	252
(c) Fuel	4794	4229	562

Solution:

(First Method):

We have to find a quantity index number for Oct. 2021 with base October 2020. Since 'Receipts in 2020 represents the base period values (p_0q_0), the required quantity index may be obtained as the weighted A.M. quantity relatives, using these receipts as weights.

Table 8.14 Calculations for Quantity Index Number

Type of traffic	q_0	q_n	Weight	Quantity Relative (q_n/q_0) $\times$ 100	(5) $\times$ (4)
(1)	(2)	(3)	(4)	(5)	(6)
(a)	1246	1206	776	$(1206 \div 1246) \times 100 = 97$	75272
(b)	1125	981	252	$(981 \div 1125) \times 100 = 87$	21924
(c)	4794	4229	562	$(4229 \div 4794) \times 100 = 88$	49456
Total	-	-	1590	-	146652

$$\text{Quantity Index} = \frac{\sum (\text{Quantity relative} \times \text{Weight})}{\sum (\text{Weight})} = \frac{146,652}{1590} = 92$$

(Second method):

We are given q_0 , q_n and p_0q_0 for each of the three types of traffic, q_n and it is required a quantity index number. However, we can find p_0 by using the relation

$$p_0 = \frac{p_0q_0}{q_0} = \frac{\text{Receipts in Oct. 2020 (£ '000)}}{\text{Volume of traffic (Tons '000)}}$$

Now, using p_0 , q_0 and q_n it is possible to find :

$$\text{Laspeyres' Quantity Index} = \frac{\sum q_n p_0}{\sum q_0 p_0} \times 100$$

Table 8.15-Calculations for Quantity Index Number

Type of traffic	q_0	q_n	p_0q_0	$p_0 = \frac{(4)}{(2)}$	$q_n p_0 = (3) \times (5)$
(1)	(2)	(3)	(4)	(5)	(6)
(a)	1246	1206	776	$776 \div 1246 = 0.623$	$1206 \times 0.623 = 751.33$
(b)	1125	981	252	$252 \div 1125 = 0.224$	$981 \times 0.224 = 219.744$
(c)	4794	4229	562	$562 \div 4794 = 0.117$	$4229 \times 0.117 = 494.793$
Total	-	-	1590	-	1465.875

$$\therefore \text{Quantity Index} = \frac{1465.875}{1590} \times 100 = 92$$

Example 8.18 From the following data calculate Paasche's quantity index number for the year 2022, with 2001 as base:

Commodity	Quantity		Value
	2010	2020	2020
A	54	250	540
B	93	75	825
C	18	56	448
D	6	8	56
E	23	47	141

Solution: (Note: Here, values in current period are given)

$$\text{Paasche's Quantity Index} = \frac{\sum q_n p_n}{\sum q_0 p_n} \times 100$$

Where q_0, q_n denote quantities in the base year (2010) and current year (2020) respectively; and p_n denotes price in the current year. Since, 'Value' denotes the product of 'price per unit' and 'quantity', i.e. Value = ($\text{price} \times \text{quantity}$), hence, 'Value in 2020 as shown in the question must be the product $q_n p_n$. Thus, we are given q_0, q_n and $q_n p_n$ for each commodity, and it is required to find Paasche's Quantity Index by combining the data. However, we can find the values of p_n using the $p_n = \frac{q_n p_n}{q_n} = \frac{\text{Value in 1969}}{\text{Quantity in 1969}}$. Utilizing the values of q_0, q_n and p_n the index can be calculated, where values in base period are given.

Table 8.16-Calculations for Paasche's Quantity Index

Commodity	q_0	q_n	$p_n q_n$	$p_n = \frac{(4)}{(3)}$	$q_0 p_n = (2) \times (5)$
(1)	(2)	(3)	(4)	(5)	(6)
A	54	250	540	$540 \div 250 = 2.16$	$54 \times 2.16 = 117$
B	93	75	825	$825 \div 75 = 11.00$	$93 \times 11.00 = 1023$
C	18	56	448	$448 \div 56 = 8.00$	$18 \times 8.00 = 144$
D	6	8	56	$56 \div 8 = 7.00$	$6 \times 7 = 42$
E	23	47	141	$141 \div 47 = 3.00$	$23 \times 3.00 = 69$
Total	-	-	2010	-	1395

$$\text{Paasche's Quantity Index} = \frac{2010}{1395} \times 100 = 144$$

Note that, If values in the base year ($p_0 q_0$) were given, the weighted arithmetic mean of relatives, using those values as, would lead to Laspeyres' index.

$$\frac{\sum (p_0 q_0) \left(\frac{q_n}{q_0} \times 100 \right)}{\sum p_0 q_0} = \frac{\sum p_0 q_n}{\sum p_0 q_0} \times 100 = \text{Laspeyres' Quantity Index.}$$

However, since values in the current year ($p_n q_n$) are given, this method will not be applicable. Here, the weighted harmonic mean of relatives is the appropriate index, using the current year values as weights. This leads to the same formula as Paasche's Quantity Index.

$$\text{Paasche's Quantity Index} = \frac{\sum p_n q_n}{\sum \frac{p_n q_n}{\left(\frac{q_n}{q_0} \times 100\right)}} = \frac{\sum p_n q_n}{\sum q_0 p_n}$$

In the solution given above, Paasche's Quantity index formula has been applied directly without using the harmonic mean of quantity relatives.

Tests of index numbers

Tests of index numbers are crucial in evaluating the reliability, accuracy, and consistency of index number formulas used in economic and business analysis.

Tests of index numbers are crucial in evaluating the reliability, accuracy, and consistency of index number formulas used in economic and business analysis. These tests help determine whether an index number method is theoretically sound and practically useful. In order to judge the efficiency of an index number formula as a measure of the level of a phenomenon from one period to another, the noted economist Irving Fisher suggested certain tests. The three most important tests of index numbers are:

- (1) Time Reversal test,
- (2) Factor Reversal test, and
- (3) Circular test. These tests are based on the analogy that what is true for an individual item should also hold for a group of items.

The necessity of applying tests to index numbers lies in ensuring that the index numbers serve their purpose accurately, consistently, and objectively in economic analysis and business decision-making.

(1) Time Reversal Test:

Time reversal test is satisfied by simple aggregative formula, Marshall-Edgeworth's formula, Fisher's ideal index formula, and simple geometric mean of relative's formula.

According to this test, a good index number formula should work both ways, forward and backward, with respect to time. In other words, we should get the same picture of change between two points of time, no matter which of the two is taken as base. Consequently, the index number (I_{on}) for period n with base period o should be the reciprocal of the index number (I_{no}) for period o with base period n (omitting the factor 100 from each index). Symbolically,

$$I_{on} \times I_{no} = 1$$

An index number formula which obeys this relation is said to satisfy the time reversal test.

The Time Reversal Test is a mathematical test used to check the consistency of an index number formula when the time periods (base year and current year) are reversed.

Time reversal test is satisfied by simple aggregative formula, Marshall-Edgeworth's formula, Fisher's ideal index formula, and simple geometric mean of relative's formula. Weighted aggregative formula and weighted geometric mean of relative's formula also satisfy this test, if constant weights are used which do not depend upon the base or current period. So the Time Reversal Test is a mathematical test used to check the consistency of an index number formula when the time periods (base year and current year) are reversed.

Time reversal test is based on the following analogy: If the price of a commodity changes from Tk. 4 per unit in 2001 to Tk. 8 in 2010, the price in 2010 is 200% of (i.e. 2 times) the price in 2001, and the price in 2001 is 50% of (i.e. 0.50 times) the price in 2010. The product of the two price ratios is $2 \times 0.50 = 1$. This is true for each commodity and time

reversal test ensures that the same principle holds for an index number, which embraces a group of commodities.

(2) Factor Reversal Test:

An index number formula is said to satisfy the factor reversal test, if the product of Price Index (P_{on}) and Quantity Index (Q_{on}) gives the true Value Ratio (omitting the factor 100 from each index). In other words, a good index number formula should be such that the price ratio multiplied by the quantity ratio between two points of time gives the ratio of total values. Symbolically,

$$P_{on} \times Q_{on} = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

Fisher's ideal index is the only formula which satisfies this test.

Factor reversal test is based on the following analogy: If the price per unit of a commodity changes from Tk. 4 in 2010 to Tk. 8 in 2020, and the quantity of consumption changes from 60 units to 90 units during the same period, then the price and quantity in 2020 are 200% and 150% respectively of the corresponding factors in 2010. The values (price quantity) of consumption were Tk. 240 in 2010 and Tk. 720 in 2020, so that the value ratio is $720/240 = 3$. Thus we find that the product of price ratio and quantity ratio equals the value ratio: $2 \times 1.50 = 3$. Factor reversal test ensures that the principle which holds for a single commodity should apply to the index number as a whole.

So the Factor Reversal Test checks whether an index number formula maintains consistency when the roles of prices and quantities are reversed. It is used to verify whether the price and quantity indices together reflect the change in total value over time.

(3) Circular Test

The Circular Test is a consistency test used in the theory of index numbers. It checks whether an index formula maintains logical consistency when measuring price (or quantity) changes over three or more time periods in a circular fashion. The Circular Test states that if we calculate the index number from period 0 to 1, then from 1 to 2, and finally from 2 back to 0, the product of these three index numbers should be equal to 1 (or 100 if using percentage form).

This is an extension of time reversal test. An index number formula is said to satisfy the circular test, if the time reversal test is satisfied through a number of intermediate years. Symbolically,

$$I_{01} \times I_{12} \times I_{23} \times \dots \times I_{(n-1)n} \times I_{n0} = 1$$

This means that the relation is satisfied in a circular fashion through several years, 0 to 1, 1 to 2, 2 to 3, ... (n-1) to n, and finally from a back to 0. Simple aggregative formula and the simple geometric mean of relatives formula satisfy this test. Weighted aggregative formula and weighted geometric mean of relatives formula satisfy this test, if constant weights are used for all time periods.

An index number formula is said to satisfy the factor reversal test, if the product of Price Index (P_{on}) and Quantity Index (Q_{on}) gives the true Value Ratio.

The circular test checks whether an index formula maintains logical consistency when measuring price (or quantity) changes over three or more time periods in a circular fashion.

Example 8:19 Using the following data, verify that Laspeyres' formula does not satisfy Time Reversal Test:

	1979		1980	
Commodity	Price	Quantity	Price	Quantity
Rice	32	50	30	50
Barley	30	35	25	40
Maize	16	55	18	50

Solution: Using Laspeyres' Price Index formula and omitting the factor 100,

$$\text{Index Number for 1980 with base 1979 } (I_{0n}) = \frac{\sum p_n q_0}{\sum p_0 q_0}$$

Interchanging the suffixes 0 and n

$$\text{Index Number for 1979 with base 1980 } (I_{n0}) = \frac{\sum p_0 q_n}{\sum p_n q_n}$$

[Note that we have to calculate all 4 combinations p and q , viz. $p_0 q_0, p_n q_0, p_0 q_n, p_n q_n$.]

Table 8.17-Calculations for Laspeyres' Index

Commodity	p_0	q_0	p_n	q_n	$p_0 q_0$	$p_0 q_n$	$p_n q_0$	$p_n q_n$
Rice	32	50	30	50	1600	1600	1500	1500
Barley	30	35	25	40	1050	1200	875	1000
Maize	16	55	18	50	880	800	990	900
Total	-	-	-	-	3530	3600	3365	3400

Substituting the values,

$$I_{0n} = \frac{3365}{3530}, \quad I_{n0} = \frac{3600}{3400}$$

$$\text{We find that } I_{0n} \cdot I_{n0} = \frac{3365}{3530} \times \frac{3600}{3400} \neq 1$$

This verifies that Laspeyres' formula does not satisfy Time Reversal Test.

Example 8:20 With the help of the data of Example 8:23, calculate Price Index number using Fisher's formula and show that it satisfies the Time Reversal Test.

Solution: (Calculations are shown in Table 8.17 above).

(1) Price Index Number for year 1980 with base 1990:

$$\text{Laspeyres' Index} = \frac{\sum p_n q_0}{\sum q_0 p_0} = \frac{3365}{3530}$$

$$\text{Paasche's Index} = \frac{\sum p_n q_n}{\sum p_0 q_n} = \frac{3400}{3600}$$

$$\text{Fisher's Ideal Index } (I_{0n}) = \sqrt{\text{Laspeyres' Index} \times \text{Paasche's Index}}$$

$$= \sqrt{\frac{3365}{3530} \times \frac{3400}{3600}} \quad \dots \dots \dots (i)$$

(II) Interchanging the suffixes o and n in the above formulae, Price Index Number for year 1990 with base 2010:

$$\text{Laspeyres' Index} = \frac{\sum p_0 q_n}{\sum p_n q_n} = \frac{3600}{3400}$$

$$\text{Paasche's Index} = \frac{\sum p_0 q_0}{\sum p_n q_0} = \frac{3530}{3365}$$

$$\begin{aligned} \therefore \text{Fisher's Ideal Index } (I_{no}) &= \sqrt{\text{Laspeyres' Index} \times \text{Paasche's Index}} \\ &= \sqrt{\left(\frac{3600}{3400} \times \frac{3530}{3365}\right)} \quad \dots \dots \dots (ii) \end{aligned}$$

Multiplying (i) and (ii),

$$\begin{aligned} &\text{Fisher's Index } (I_{on}) \times \text{Fisher's Index } (I_{no}) \\ &= \sqrt{(3365/3530 \times 3400/3600 \times 3600/3400 \times 3530/3365)} = \sqrt{1} = 1 \end{aligned}$$

Using Fisher's formula, we find that $I_{on} \cdot I_{no} = 1$. This verifies that Fisher's formula satisfies Time Reversal Test.

Example 8:21 Calculate the quantity index number using Fisher's formula for the following data and show that it satisfies the Time Reversal Test.

	2013		2014	
Commodity	Price	Quantity	Price	Quantity
A	6	70	8	120
B	8	90	10	100
C	12	140	16	280

Solution: Let us take 2013 as base year and 2014 as current year.

Table 8.18-Calculations for Quantity Index

Commodity	p_0	q_0	p_n	q_n	$p_0 q_0$	$q_0 p_n$	$q_n p_0$	$p_n q_n$
A	6	70	8	120	420	560	720	960
B	8	90	10	100	720	900	800	1000
C	12	140	16	280	1680	2240	3360	4480
Total	-	-	-	-	2820	3700	4880	6440

1. Quantity Index Number for year 2014 with 2013 as base:

$$\text{Laspeyres' Index} = \frac{\sum q_n p_0}{\sum q_0 p_0} = \frac{4880}{2820}$$

$$\text{Paasche's Index} = \frac{\sum q_n p_n}{\sum q_0 p_n} = \frac{6440}{3700}$$

$$\text{Fisher's Ideal Index } (I_{on}) = \sqrt{\frac{4880}{2820} \times \frac{6440}{3700}} \dots \dots \dots (i)$$

(II) Interchanging the suffixes 0 and n, Quantity Index Number for year 2013 with 2014 as base:

$$\text{Laspeyres' Index} = \frac{\sum q_0 p_n}{\sum q_n p_n} = \frac{3700}{6440}$$

$$\text{Paasche's Index} = \frac{\sum q_0 p_0}{\sum q_n p_0} = \frac{2820}{4880}$$

$$\text{Fisher's Ideal Index } (I_{no}) = \sqrt{\frac{3709 \times 2820}{6440 \times 4880}} \dots \dots \dots (ii)$$

Multiplying (i) and (ii),

$$I_{on} \cdot I_{no} = \sqrt{\frac{4880 \times 6440}{2820 \times 3700}} \times \frac{3700 \times 2820}{6440 \times 4880} = 1$$

This verifies index number formula satisfies Time Reversal Test.

It should be remembered that, examples 8.20 and 8.21 both verify that Fisher's formula satisfies Time Reversal test. The former uses Price Index, and the latter uses Quantity Index.

Example 8.22 Consider the following data:

Commodity	Price (Tk per unit)		Number of units	
	Base period	Current period	Base period	Current period
A	6	10	50	56
B	2	2	100	120
C	4	6	60	60
D	10	12	30	24
E	8	12	40	36

Required:

- Using the following data, show that Paasche's formula does not satisfy the Factor Reversal Test.
- Also use the same data to show that Fisher's Ideal formula satisfies the Factor Reversal Test.

Solution: The factor reversal test may be represented in symbols as $P_{on} \cdot Q_{on} = \text{Value Ratio}$.

Table 8.19-Calculations for Factor Reversal Test

Commodity	p_0	p_n	q_0	q_n	$p_0 q_0$	$p_0 q_n$	$p_n q_0$	$p_n q_n$
A	6	10	50	56	300	336	500	560
B	2	2	100	120	200	240	200	240
C	4	6	60	60	240	240	360	360
D	10	12	30	24	300	240	360	288
E	8	12	40	36	320	288	480	432
Total	-	-	-	-	1360	1344	1900	1880

(a) Using Paasche's formula and omitting the factor 100,

$$\text{Price Index } (P_{on}) = \frac{\sum p_n q_n}{\sum p_0 q_n} = \frac{1880}{1344}$$

Interchanging p and q,

$$\text{Quantity Index, } (Q_{on}) = \frac{\sum q_n p_n}{\sum q_0 p_n} = \frac{\sum p_n q_n}{\sum p_n q_0} = \frac{1880}{1900}$$

$$\text{Also, Value Ratio} = \frac{\sum p_n q_n}{\sum p_0 q_0} = \frac{1880}{1360}$$

We find that $P_{on} \cdot Q_{on} \neq \text{Value Ratio}$. This shows that Paasche's formula does not the Factor Reversal test.

(b) Using Fisher's Ideal formula and omitting the factor 100,

$$\text{Price Index } (P_{on}) = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} = \sqrt{\frac{1900 \times 1880}{1360 \times 1344}}$$

Interchanging p and q,

$$\text{Quantity Index } (Q_{on}) = \sqrt{\frac{\sum q_n p_0}{\sum q_0 p_0} \times \frac{\sum p_n q_n}{\sum q_0 p_n}} = \sqrt{\frac{1344 \times 1880}{1360 \times 1900}}$$

$$\text{Value Ratio} = \frac{\sum p_n q_n}{\sum p_0 q_n} = \frac{1880}{1360}$$

$$\text{Now, } P_{on} \times Q_{on} = \sqrt{\frac{1900 \times 1880}{1360 \times 1344}} \times \sqrt{\frac{1344 \times 1880}{1360 \times 1900}}$$

Cancelling the common factors under the square-root sign and simplifying,

$$P_{on} \times Q_{on} = \frac{1880}{1360} = \text{Value Ratio}$$

This shows that Fisher's Ideal formula satisfies Factor Reversal Test.

Example 8.23 Show that neither Laspeyres' formula nor Paasche's formula obeys time reversal or factor reversal tests of index numbers.

Solution:

(I) Time Reversal Test may be symbolically expressed as: $I_{on} \times I_{no} = 1$.

(a) Using Laspeyres' Price Index formula and omitting the factor 100,

$$\text{Index Number for year n with base year o } (I_{on}) = \frac{\sum p_n q_0}{\sum p_0 q_0}$$

Interchanging the suffixes o and n,

$$\text{Index Number for year o with base year n } (I_{no}) = \frac{\sum p_0 q_n}{\sum p_n q_n}$$

$$\therefore I_{on} \times I_{no} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_n}{\sum p_n q_n} \neq 1.$$

Thus, Laspeyres' formula does not obey Time Reversal Test.

- (b) Using Paasche's Price Index formula and omitting the factor 100,

$$I_{on} = \frac{\sum p_n q_n}{\sum p_0 q_n}$$

Interchanging the suffixes o and n,

$$I_{no} = \frac{\sum p_0 q_0}{\sum p_n q_0}$$

$$I_{on} \times I_{no} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0} \neq 1$$

Thus, Paasche's formula also does not obey Time Reversal Test.

$$(II) \text{ Factor Reversal Test may be expressed as } P_{on} \cdot Q_{on} = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

- (c) Using Laspeyres formula, and omitting the factor 100,

$$\text{Price Index for year n with base year o } (P_{on}) = \frac{\sum p_n q_0}{\sum p_0 q_0}$$

Interchanging p and q, Quantity Index for base year n with base year o

$$Q_{on} = \frac{\sum q_n p_0}{\sum q_0 p_0} = \frac{\sum p_0 q_n}{\sum p_0 q_0}$$

Multiplying, we have formula by Laspeyres formula

$$P_{on} \times Q_{on} = \frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_n}{\sum p_0 q_0} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$$

$$\text{i. e. } P_{on} \times Q_{on} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$$

This proves that Laspeyres formula does not satisfy Factor Reversal Test.

- (d) Applying Paasche's formula, it will be found that

$$P_{on} \times Q_{on} = \frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum p_n q_n}{\sum p_n q_0} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$$

This proves that paasche's formula does not satisfy factor reversal test.

Example 8.24 Examine whether Fisher's ideal index formula satisfies the Time reversal and Factor reversal tests.

Solution: Using Fisher's "ideal" index formula, Price Index for year n with base year o is given by (omitting the factor 100)

$$I_{on} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}}$$

Interchanging the suffixes o and n, Price Index for year o with base year n is

$$I_{no} = \sqrt{\frac{\sum p_0 q_n}{\sum p_n q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0}}$$

Multiplying we get

$$\begin{aligned} I_{on} \times I_{no} &= \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}} \times \sqrt{\frac{\sum p_0 q_n}{\sum p_n q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0}} \\ &= \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum p_0 q_n}{\sum p_n q_n} \times \frac{\sum p_0 q_0}{\sum p_n q_0}} = \sqrt{1} = 1 \end{aligned}$$

(Since all terms cancel one another)

This shows that Fisher's ideal formula obeys Time Reversal test.

In order to apply Factor Reversal test, we see that Price index by Fisher's ideal formula is $P_{on} = \sqrt{\frac{\sum p_n q_0}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}}$

Interchanging p and q, Quantity Index is given by

$$Q_{on} = \sqrt{\frac{\sum q_n p_0}{\sum q_0 p_0} \times \frac{\sum q_n p_n}{\sum q_0 p_n}}$$

$$P_{on} = \sqrt{\frac{\sum p_0 q_n}{\sum p_0 q_0} \times \frac{\sum p_n q_n}{\sum p_n q_0}}$$

(Rearranging the factors p, q)

Multiplying P_{on} and Q_{on} we have

$$\begin{aligned} P_{on} \times Q_{on} &= \sqrt{\frac{\sum p_0 q_n}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_n q_0}} \sqrt{\frac{\sum q_n p_0}{\sum q_0 p_0} \cdot \frac{\sum q_n p_n}{\sum q_0 p_n}} \\ &= \sqrt{\frac{\sum p_0 q_n}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_n q_0} \cdot \frac{\sum q_n p_0}{\sum q_0 p_0} \cdot \frac{\sum q_n p_n}{\sum q_0 p_n}} \\ &= \sqrt{\frac{\sum p_n q_n}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_0 q_0}} = \frac{\sum p_n q_n}{\sum p_0 q_0} \\ \text{i.e. } P_{on} \times Q_{on} &= \frac{\sum p_n q_n}{\sum p_0 q_0} \end{aligned}$$

This shows that Fisher's ideal index formula obeys Factor Reversal test.

Example 8.25 Show that the index number obtained by averaging the unweighted price relatives does not satisfy "time reversal test"

Solution: The price relative for year n with base year o is given by the formula (omitting the factor 100)

$$\text{Price Relative} = P_n / P_0$$

If there are k items in the series, the unweighted A.M. of Price Relatives Index is given by

$$I_{on} = \sum (\text{Price Relatives}) \div k = \sum (P_n / P_0) \div k$$

Interchanging the, suffixes o and n,

$$\begin{aligned} I_{no} &= \sum (P_0 / P_n) \div k \\ I_{on} \times I_{no} &= \frac{1}{k^2} \sum (P_n / P_0) \cdot \sum (P_0 / P_n) \\ &= \frac{1}{k^2} \left(\frac{P'_n}{P_0} + \frac{P''_n}{P_0} + \dots + \frac{P^k_n}{P_0} \right) \left(\frac{P'_n}{P_n} + \frac{P''_n}{P_n} + \dots + \frac{P^k_n}{P_n} \right) \neq 1 \end{aligned}$$

where, $p'_0, p''_0, \dots, p^k_0$ represent the prices of items 1, 2, ..., k respectively in the year 0, and $P'_n, P''_n, \dots, P^k_n$ represent the corresponding prices in the year n. Thus;

$$I_{on} \cdot I_{no} \neq 1$$

This result shows that the index number obtained by averaging the unweighted price relatives does not satisfy the time reversal test.

Chain base method and its Advantages and Disadvantages

There are two methods of construction of index numbers depending on the nature of base period employed: (i) Fixed Base method, and (ii) Chain Base method. Most of the index numbers in common use are of the fixed base type, where a fixed period is chosen as base and the index number for any given year is calculated by direct reference to this fixed base period. The fixed base index for any year is not, therefore, affected by changes in price or quantity in any other year. It is however considered that the net changes in any given year are the result of gradual changes that have taken place during the past years. This idea is reflected in "Chain Base Index" numbers.

The fixed base index for any year is not, therefore, affected by changes in price or quantity in any other year.

For the construction of index numbers by the chain base method, using an appropriate index number formula (say, Laspeyres' formula), it is first necessary to compute index numbers for all the years, always using the preceding year as base. These are known as Link Index.

It is first necessary to compute index numbers for all the years, always using the preceding year as base.

Link Index = Index Number with preceding period as base.

For example, using Laspeyres' formula,

$$\text{Link index for year 1 } (I_{01}) = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100;$$

$$\text{Link index for year 2 } (I_{12}) = \frac{\sum p_2 q_1}{\sum p_1 q_1} \times 100;$$

$$\text{Link index for year 3 } (I_{23}) = \frac{\sum p_3 q_2}{\sum p_2 q_2} \times 100;$$

$$\text{Link index for year 4 } (I_{34}) = \frac{\sum p_4 q_3}{\sum p_3 q_3} \times 100; \text{ etc.}$$

The link indices I_{01} , I_{12} , I_{23} , I_{34} ----- are then multiplied successively (called chaining process) in order to relate them to a common base. The progressive products, expressed as percentages, give the required index numbers by the chain base method. These are called Chain Index Numbers or Chain Base Index Numbers. Thus, a chain index number is the product of several index numbers, each calculated with the preceding period as base.

The chain index numbers with reference to year 0 are (omitting the factor 100 from each index)

$$I'_{01} = I_{01}$$

$$I'_{02} = I_{01} \times I_{12}$$

$$I'_{03} = I_{01} \times I_{12} \times I_{23}$$

$$I'_{04} = I_{01} \times I_{12} \times I_{23} \times I_{34}$$

(Here I' is used for Chain Index and I for index of the fixed base type).

The chain index number I'_{0n} , will not in general be equal to the corresponding fixed base index number I_{0n} unless the formula employed satisfies the circular test of index numbers.

Advantages:

- (1) The chain base index is more realistic in nature than the fixed base index, since the effects of all intermediate years are taken into consideration.
- (2) The chain base method enables comparison between two adjacent time periods through the link indices. This is far more useful in business and commerce than the indirect comparison through a remote fixed base.
- (3) The method also makes it possible the dropping of obsolete items and inclusion of new ones. The necessity of substituting certain items in the existing list is frequently felt when computing a series of index numbers over a long period of time, because of the changing habits of people and new commodities coming into use. If the fixed base method were used, the entire series of index numbers will have to be recalculated, when the list of commodities is altered.

The chain base index is more realistic in nature than the fixed base index, since the effects of all intermediate years are taken into consideration.

Disadvantages:

- (1) The significance of index numbers calculated by the chain base method is difficult to understand.
- (2) The calculations are heavier in the chain base method.
- (3) If an error is committed in the calculation of any link index number the entire series of chain base index numbers will be wrong. Also, if data for even one year are missing, the subsequent chain index numbers can not be calculated.
- (4) Chain base index numbers are really suitable for short periods only. If changes in the list of items are frequent, the index may in the later years reflect quite different movements than the figures in the earlier periods.

Chain base index numbers are really suitable for short periods only.

Example 8.26 Given the following information, construct chain index numbers (Base 2012= 100) for the years 2013-2017:

Year	2013	2014	2015	2016	2017
Link Index	103	98	105	112	108

Solution:**Table 8.20 Calculations for Chain Index**

Year	Link Index	Chain Index (Base 1962 = 100)	
$Y_0 = 1962$	100	100	
$Y_1 = 1963$	$I_{01} = 103$	$I'_{01} = 100 \times 1.03$	=103
$Y_2 = 1964$	$I_{12} = 98$	$I'_{02} = 100 \times (1.03 \times 0.98)$	=101
$Y_3 = 1965$	$I_{23} = 105$	$I'_{03} = 100 \times (1.03 \times 0.98 \times 1.05)$	=106
$Y_4 = 1966$	$I_{34} = 112$	$I'_{04} = 100 \times (1.03 \times .98 \times 1.05 \times 1.12)$	=119
$Y_5 = 1967$	$I_{45} = 108$	$I'_{05} = 100 \times (1.03 \times .98 \times 1.05 \times 1.12 \times 1.08)$	=128

Cost of living index numbers, Construction procedures and Determination of Weights

Cost of Living Index numbers are special-purpose index numbers which are designed to measure the relative change in the cost level for maintaining similar standard of living in two different situations. These are generally intended to represent the average changes in prices over

Cost of Living Index numbers are special-purpose index numbers which are designed to measure the relative change in the cost level for maintaining similar standard of living in two different situations.

time, paid by the ultimate consumer for a specified group of goods and services; and hence are also called Consumer Price Index numbers. Generally, the consumption pattern varies with the class of people and the geographical area covered. Hence cost of living index (C.L.I.) numbers must always relate to a specified class of people and a specified geographical area.

The steps in the construction of a Cost of Living Index are as follows:

- (1) The first step is to decide on the class of people for whom the index number is intended. It is extremely important to define this in clear terms.
- (2) The next step is to conduct a '*family budget enquiry*' in the base period relating to the class of people concerned, by the process of random sampling. This would give us information regarding the nature and quality of goods consumed by an average family' and also enable determination of weights for computing the index. Only important items among those which are used by the majority of the class of people are included in the construction of a cost of living index.
- (3) The items of expenditure are classified in certain major Groups, e.g. (i) Food, (ii) Clothing, (iii) Fuel & light, (iv) Housing, and (v) Miscellaneous. These major groups are further divided into smaller groups and sub-groups, so that the items are individually mentioned.
- (4) Arrangements should be made to collect retail prices of the items at regular intervals of time from important local markets. Price quotations are taken at least once a week.
- (5) For each item there will be a number of price quotations covering different qualities and markets. The simple average of price relatives of the different quotations is taken as the price relative for the particular item.
- (6) A separate index number is then computed for each Group. Using Laspeyres' formula in the form of weighted average of price relatives.

$$\text{Group Index (I)} = \frac{\sum w \left(\frac{p_n}{p_0} \times 100 \right)}{100}, \quad \text{where } w = \frac{p_0 q_0}{\sum p_0 q_0} \times 100$$

Thus, in the construction of a Group Index, the weight (w) of an item is the percentage expenditure of an 'average family' on that item in relation to the total expenditure in the Group, as obtained from the family budget enquiry.

- (7) The weighted average of group index numbers gives the final Cost of Living Index number.

$$\text{Cost of Living Index} = \frac{\sum IW}{100}$$

The weight (W) of a group index is the percentage of total expenditure of an average family spent on that group, as shown by the family budget enquiry.

- (8) Cost of living index numbers are generally constructed for each week. The average of the weekly index numbers is taken as the index number for a month. The average of monthly index numbers gives the cost of living index for the whole year.

Cost of living index numbers are generally constructed for each week. The average of the weekly index numbers is taken as the index number for a month. The average of monthly index numbers gives the cost of living index for the whole year.

Determination of weights:

- (i) Cost of Living Index numbers are primarily used for the calculation of dearness allowance, so that the same standard of living as in the base year can be maintained.
- (ii) The reciprocal of C.L.I. may be used to measure the purchasing power of money.
- (iii) Cost of living index numbers are also used to find "real wages" by the process of 'deflation'.

Bias in Laspeyres and Paasche's formulae for Cost of Living Index (C L I):

The various tests of index numbers, viz. time reversal test, factor reversal test and circular test, are not the sole criteria for determining the suitability of a formula, and practical considerations often influence the choice of one formula in preference to another. It has been shown that none of Laspeyres' and Paasche's formulae obeys any of the tests of index numbers. However, Laspeyres' formula has the superior advantage that it uses the base period quantities (for price index) as weights; so that the same set of weights can be used over a long period of time, until it becomes necessary to change the base. On the other hand, Paasche's formula, which uses the current period quantities, necessitates determination of weights every time the index number is calculated. It is for this reason that Laspeyres' formula is by far the most widely used in the construction of index numbers, especially in the form of weighted A.M. of price relatives.

Practical considerations often influence the choice of one formula in preference to another.

$$\begin{aligned} \text{Laspeyres Price Index} &= \frac{\sum p_n q_o}{\sum p_o q_o} \times 100 = \frac{\sum \left(\frac{p_n}{p_o} \times 100 \right) \times p_o q_o}{\sum p_o q_o} \\ &= \frac{\sum \left(\frac{p_n}{p_o} \times 100 \right) \times w}{100} = \frac{\sum (\text{Price relative}) \times w}{100} \end{aligned}$$

where $w = \frac{p_o q_o}{\sum p_o q_o} \times 100$ is a pure number showing the value ($\sum p_o q_o$) of each item in the base period expressed as a percentage of total value ($\sum p_o q_o$) and $\sum w = 100$

Cost of Living Index number measures the ratio of money values required to maintain equal satisfaction for a particular class of people in two different situations. Let p_o, p_n denote the price per unit of a set of goods and services consumed and q_o, q_n denote the actual quantities (number of units) consumed in the base and current periods respectively. If $q^{(n)}$ denotes the current period quantities which will produce equal satisfaction relative to base period quantities of consumption q_o , then the total money value of consumption is $\sum p_o q_o$ in the base period and the amount necessary to produce the same satisfaction in the current period is $\sum p_n q^{(n)}$ [Note the distinction between q_n and $q^{(n)}$]. The true cost of living index is then

$$I = \frac{\sum p_n q^{(n)}}{\sum p_o q_o} \times 100$$

The practical difficulty of calculating this index lies in determining the exact quantities $q^{(n)}$, which yield equal satisfaction as in the base period,

Cost of Living Index number measures the ratio of money values required to maintain equal satisfaction for a particular class of people in two different situations.

especially because the consumption pattern varies with change in the real income level in the two periods.

$$\text{Laspeyres formula } L = \frac{\sum p_n q_0}{\sum p_0 q_0} \times 100$$

measures roughly the cost of maintaining the base period rate of consumption at current period prices, compared with base period cost.

$$\text{Paasche's formula } P = \frac{\sum p_n q_n}{\sum p_0 q_n} \times 100$$

shows a comparison of the cost in the current period relative to what it would have cost if current period quantities were consumed in the base period. None of these two formulae measures the true index. They may only be used as approximations.

It is a common experience that when prices increase, relatively smaller quantities are consumed, and cheaper articles are used in larger quantities. It is for this reason that in Laspeyres' formula (L) the numerator is slightly larger than that in the true index I , making L larger than I . Similarly, the denominator in Paasche's formula would be relatively larger than that in the true index I . Thus Laspeyres' formula has a positive bias and Paasche's formula has a negative bias. The concept can be clearly explained with the help of weighted co-efficient of correlation.

Example 8.27 With the help of the concept of weighted co-efficient of correlation, show that Laspeyres' formula has an upward bias.

Solution: In a bivariate distribution with variables x and y if the pairs of observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ have weights $f_1, f_2, \dots, f_n$ respectively, the coefficient of correlation between the variables is

$$r = \frac{\text{Covariance}}{\sigma_x \sigma_y}$$

where

$$\text{Covariance} = \frac{\sum fxy}{N} - \left(\frac{\sum fx}{N} \right) \left(\frac{\sum fy}{N} \right), \quad N = \sum f,$$

and σ_x, σ_y denote the standard deviations of x and y . Let the variable x price relative, y represent quantity relative, and the weight f represent value in the base period, i.e. (omitting the factors 100 from each relative)

$$x = \frac{p_n}{p_0}, \quad y = \frac{q_n}{q_0}, \quad f = p_0 q_0$$

Then

$$\begin{aligned} \text{Covariance } y &= \frac{\sum (p_0 q_0 \cdot \frac{p_n}{p_0} \cdot \frac{q_n}{q_0})}{\sum p_0 q_0} - \frac{\sum (p_0 q_0 \cdot \frac{p_n}{p_0})}{\sum p_0 q_0} \cdot \frac{\sum (p_0 q_0 \cdot \frac{q_n}{q_0})}{\sum p_0 q_0} \\ &= \frac{\sum p_n q_n}{\sum p_0 q_0} - \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum p_0 q_n}{\sum p_0 q_0} \\ &= \frac{\sum p_n q_n}{\sum p_0 q_0} - \frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum p_0 q_n}{\sum p_0 q_0} \\ &= P_p \cdot L_q - L_p \cdot L_q = L_q (P_p - L_p) \end{aligned}$$

Where L_p = Laspeyres' Price index, L_q = Laspeyres' Quantity index, and P_p = Paasche's Price index (the factor 100 from each index) we have shown that:

Covariance between Price Relative and Quantity Relative

$$= L_q(P_p - L_p)$$

Since the sign of correlation coefficient is the same as that of covariance (because in the denominator σ_x and σ_y are always positive), the covariance will be positive or negative according as there is positive or negative correlation between the variables. In the present case, the variables x and y are price relative and quantity relative respectively, and the nature of correlation between them can be determined from the following considerations. It is a common knowledge that for any commodity the quantity consumed will be relatively less when price is higher, i.e. they move in opposite directions; hence, price relative and quantity relative will be negatively correlated. Therefore, the covariance between them, viz. $L_q(P_p - L_p)$ will be negative. Since L_q is always positive, the other factor $(P_p - L_p)$ must be negative, and consequently Laspeyres' Price index L_p than Paasche's Price index P_p . In other words, Laspeyres' formula, has an upward bias.

Base shifting, Splicing and Deflation

Base Shifting refers to the technique of changing the given base period of a series of index numbers and recasting them to form a new series with reference to a new base period. Base shifting is used in the following situations:

- (i) When the base period is too old, it is necessary to shift the base to a more recent period in order that the data are more meaningful.
- (ii) Base shifting is also necessary for comparing two or more series of index numbers with different base periods. If all the series are expressed with a common base, the comparison is easier and quicker.

Strictly speaking, base shifting will involve recomputation of the entire series of index numbers by applying the formula already employed. However, this is a difficult job. A relatively simple but approximate method consists in assuming the index number of the new base period as 100 and expressing the old series of index numbers as percentages of the index number for the new base.

$$\begin{aligned} & \text{Index Number for any year (with new base)} \\ &= \frac{\text{Old Index Number for the year}}{\text{Old index Number for the new base year}} \times 100 \end{aligned}$$

Theoretically, this method will give exact results if the formula employed for index number obeys the Circular Test.

Example 8.28 The following table shows the Index Number of Industrial Production in India:

Year :	2013	2014	2015	2016	2017
Index Number (Base 2000=100)	112	114	119	132	139

Shift the base to the year 2013 and recast the data.

Base Shifting refers to the technique of changing the given base period of a series of index numbers and recasting them to form a new series with reference to a new base period.

Solution:**Table 8.21:-**Base Shifting from 2000 to 2013

Year	Index Number (2000 = 100)	Index Number (2013 = 100)
1973	112	100
1974	114	$\frac{114}{112} \times 100 = 102$
1975	119	$\frac{119}{112} \times 100 = 106$
1976	132	$\frac{1132}{112} \times 100 = 118$
1977	139	$\frac{139}{112} \times 100 = 124$

Example 8.29 Assume that an index number is 100 in 2018; it rises 3% in 2019, falls 1% in 2020 and rises 2% in 2021 and 3% in 2022; rise and fall begin with respect to the previous year. Calculate the index for the five years, using 2022 as the base year.

Solution:**Table 8.22:-**Base Shifting from 2018 to 2022

Year	Index Number (Base 2018=100)	Index Number (Base 2022=100)
2018	100	$\frac{100}{107} \times 100 = 93$
2019	$\frac{103}{100} \times 100 = 103$	$\frac{103}{107} \times 100 = 96$
2020	$\frac{99}{100} \times 103 = 102$	$\frac{102}{107} \times 100 = 95$
2021	$\frac{102}{100} \times 102 = 104$	$\frac{104}{107} \times 100 = 97$
2022	$\frac{103}{100} \times 104 = 107$	100

Splicing:

Splicing is the technique of combining two or more overlapping series of index numbers with different base periods to obtain a single continuous series of index numbers with a common base period.

This is the technique of combining two or more overlapping series of index numbers with different base periods to obtain a single continuous series of index numbers with a common base period. In effect, this is equivalent to shifting the bases of the different series to one fixed base period. Splicing helps comparison among the different years by means of a single continuous series of index numbers. Like base shifting, the technique of splicing will give accurate results only when the formula employed satisfies the Circular Test.

Example 8.30 Two series of index numbers are given below:

'A' series

Year	2010	2011	2012	2013	2014	2015	2016	2017
Index (2009 =100)	120	130	200	300	350	370	380	400

'B' series

Year	2017	2018	2019	2020	2021	2022	2023	2024
Index (2017=100)	100	110	90	98	101	110	98	96

For purposes of continuity of records, you are required to construct a combined continuous series 'C', say, on 2009=100 basis, and covering records upto 2014. What is this technique called? Suggest any alternative continuous series which can serve the main purpose.

Solution:

The technique is called "Splicing". We shall first shift the base of B-series to the year 2010, so that the index numbers for all the years from 2010 to 2024 will be comparable. This combined continuous series is shown in col. (4) of Table 8.22. The alternative continuous series is obtained by shifting the base of the A-series to the year 2017. This is shown in col. (5) of the Table.

Table 8.23 Splicing Two Series of Index Numbers

Year	A-series index (2009=100)	B-series index (2017=100)	New Continuous C-series Index	
			(2009= 100)	(2017 =100)
(1)	(2)	(3)	(4)	(5)
2010	120	-	120	30
2011	130	-	130	32.5
2012	200	-	200	50
2013	300	-	300	75
2014	350	-	350	87.5
2015	370	-	370	92.5
2016	380	-	380	95
2017	400	100	400	100
2018	-	110	440	110
2019	-	90	360	90
2020	-	98	392	98
2021	-	101	404	101
2022	-	110	440	110
2023	-	98	392	98
2024	-	96	384	96

[Note: In col. (4), the figures for 2018 to 2024 are obtained on multiplying the B-series by $400/100=4$. In column (5), the figures for 2010 to 2016 are obtained on multiplying the A-series by $100/400 = \frac{1}{4}$]

Deflation:

This refers to the technique of adjusting the values of a series after making allowance for changes in price level. When prices increase, we can purchase less with the same amount of money spent in earlier years; i.e. the 'purchasing power' of money diminishes. Consequently a rise in the 'money income' will not really means the same percentage increase in the standard of living; i.e. the 'real income' will be less than the 'money income'. For a true comparison, it becomes necessary to adjust the money income for changes in the cost of living index numbers. Thus, the real income is determined after 'deflating' the money income by cost of living index.

Deflation refers to the technique of adjusting the values of a series after making allowance for changes in price level.

$$\text{Real Income} = \frac{\text{Money Income}}{\text{Cost of Living Index}} \times 100$$

The technique of deflation is used extensively to find 'real wages' and also to deflate value series, taka sales, etc. by the corresponding price index numbers.

Example 8.31 Deflate the per capita income shown in the following table on the basis of the rise in the cost of living index and comment on your results:

Year	2015	2016	2017	2018	2019	2020	2021	2022
Cost of Living Index	100	110	120	130	150	200	250	350
Per Capita Income (Rs.)	65	70	75	80	90	100	110	130

Solution:

Table 8.24 Deflating Per Capita Income

Year	Cost of Living Index (2015=100)	Actual Per Capita Income (Tk.)	"Real Income" (Tk.)
(1)	(2)	(3)	(4)
2015	100	65	65.00
2016	110	70	$\frac{70}{110} \times 100 = 63.64$
2017	120	75	$\frac{75}{120} \times 100 = 62.50$
2018	130	80	$\frac{80}{130} \times 100 = 61.54$
2019	150	90	$\frac{90}{150} \times 100 = 60.00$
2020	200	100	$\frac{100}{200} \times 100 = 50.00$
2021	250	110	$\frac{110}{250} \times 100 = 44.00$
2022	350	130	$\frac{130}{350} \times 100 = 37.14$

Comments: It is observed from col. (2) that although actual income has gradually increased from Tk. 65 in 2015 to double its value in 2022, the "real income" has considerably gone down. This indicates that people of the particular category have been hard hit by the substantial rise in the cost of living index.

8.10 Errors in Index Numbers:

All index numbers are affected by mainly three types of errors:-

- (1) Formula error,
- (2) Sampling Error,
- (3) Homogeneity error.

Formula error: There is no index number formula which can measure the price changes exactly. Each formula in common use has its own defect, and consequently some error is inherent in each index number formula. This is known as 'formula error'. This error can never be eliminated.

Sampling error: All index numbers are computed on the basis of price and quantity of some selected commodities. It is expected that this sample of commodities will give a fair picture of the level of price or quantity. However, since many other commodities cannot be taken into consideration, the calculated index number can never represent the changes in the phenomenon accurately. The error thus introduced by selecting a sample of commodities is known as the 'sampling error'. Naturally, the sampling error diminishes with increase in the number of commodities.

Since many other commodities cannot be taken into consideration, the calculated index number can never represent the changes in the phenomenon accurately.

Homogeneity error: This error arises due to the fact that index numbers are constructed from such commodities which are marketed approximately in the same quality both in the base and the current periods. With the passage of time, new commodities replace many of the old commodities and hence the homogeneity in composition of the commodities cannot be strictly maintained. Consequently, the error increases as the gap between the base and current periods increases.

The error increases as the gap between the base and current periods increases.

Self-Assessment Questions:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The Time Reversal Test is satisfied when:
 - a) $P_{01} \times P_{10} = 0$
 - b) $P_{01} \times P_{10} = 1$
 - c) $P_{01} = P_{10}$
 - d) $P_{01} + P_{10} = 100$
- (ii) Factor Reversal Test states that:
 - a) Price index $\times$ Quantity index = Value index
 - b) Price index + Quantity index = Value index
 - c) Price index – Quantity index = Value index
 - d) Price index / Quantity index = Value index
- (iii) Which index passes both the Time Reversal Test and the Factor Reversal Test?
 - a) Laspeyres
 - b) Paasche
 - c) Fisher's Ideal Index
 - d) Edgeworth Marshall
- (iv) To find the index for a year in a chain index, do the following:
 - a) Adding up fixed base values
 - b) Making changes from one year to the next
 - c) Multiplying link relatives by the score from the previous year
 - d) Utilizing only the base year and present year information
- (v) This is what the Cost of Living Index is:
 - a) Index of Quantity
 - b) The Value Index
 - c) Price List
 - d) Chain Index

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The Quantity Index measures pricing fluctuations over time.
- (ii) The Time Reversal Test determines if the index number remains consistent across time periods.
- (iii) Laspeyres Index always passes the Time Reversal Test.
- (iv) Fisher's Ideal Index passes the Factor Reversal Test.
- (v) Chain-based index numbers are calculated by linking many short-term indexes.

Answer:

Multiple-Choice Question:

1. (i) b (ii) a (iii) c (iv) c (v) c

True/False

2. (i)- F (ii)- T (iii)- F (iv)- T (v)- T

Review Questions

1. Explain with example what you mean by a price index number and write down its uses.
2. Write down the correct answer-
If now the prices of all commodities in a place have been decreased by 35% over the base period prices, then the index number of prices for the place is (index number of prices of base period = 100): (i) 100, (ii) 135, (iii) 65, (iv) 35, (v) None of these
4. "Index Numbers are economic barometers" Explain.
5. Discuss the different steps that have to be taken in the construction of a price index number.
6. Write down the well-known formulae for comparing price levels in two time periods, explaining every symbol used. Give two interpretations of Laspeyres' Price Index number.
7. Explain the terms: Price Relative, Quantity Relative, and Value Relative-with reference to a single commodity and deduce the Factor Reversal property.
In 2017 the price of a commodity increased by 50% over that in 1970 while the production of the quantity decreased by 30%. By what percentage did the total rupee value of the commodity in 2017 increase or decrease with respect to the 1970 value?
8. Explain and give the expressions for Time Reversal test and Factor Reversal test.
9. Show that both Laspeyres' and Paasche's price index numbers may be regarded as weighted averages of price relatives.
10. What are the tests to be satisfied by a good index number? Examine how far they are met by Fisher's ideal index number.
11. Explain briefly: Time Reversal and Factor Reversal Tests of index number. Indicate whether the following index numbers satisfy one or other of these tests: Laspeyres', Paasche's, Marshall-Edgeworth's and Fisher's Ideal index numbers.
12. Show that the simple aggregative type of index number satisfies the time reversal and circular tests, but does not satisfy the factor reversal test.
13. What is the chain base method of construction of index numbers and how does it differ from the fixed base method? Explain.
14. What do you mean a link index? Discuss the relative merits and demerits of chain base and fixed base index numbers.
15. Briefly describe the various steps involved in constructing the cost of living index number.
16. An enquiry into the budgets of the middle class families of a certain city revealed that on an average the percentage expenses on the different groups were-Food 45, Rent 15, Clothing 12, Fuel and Light 8, and Miscellaneous 20. The group index numbers for the current year as compared with a fixed base period were respectively 410, 150, 343, 248 and 285. Calculate the consumer price index number for the current year. Mr. X was getting Tk. 24000 in the base period

and Tk. 4300 in the current year. State how much he ought to have received as extra allowance to maintain his former standard of living.

17. During a certain period the cost of living index number goes up from 110 to 200 and the salary of a worker is also raised from Tk. 325 to Tk. 500. Does the worker really gain, and if so, by how much in real terms?
18. Explain what is precisely meant by saying that Laspeyres' formula has an upward bias while Paasche's has a downward bias.
19. What is meant by (i) base shifting, (ii) splicing, and (iii) deflating of index numbers? Explain with illustrations.
20. Discuss the different types of errors that affect a price index number.
21. Find the Simple Aggregative index number from the following data:

Commodity	Base Price	Current Price
Rice	140	180
Sugar	100	300
Oil	400	550
Wheat	125	150
Pulse	160	200

22. Find by the weighted aggregative method, the index number of the following data:

Commodity	Base Price	Current Price	Weight
Rice	140	180	10
Oil	400	550	7
Sugar	100	250	6
Wheat	125	150	8
Fish	200	300	4

23. Calculate the price index numbers by (a) Paasche's method, (b) Laspeyre's method, (c) Bowley's method, (d) Fisher's ideal formula.

Commodities	2019		2020	
	Price (Tk.)	Quantity (Kgs.)	Price (Tk.)	Quantity (Kgs.)
A	20	8	40	6
B	50	10	60	5
C	40	15	50	10
D	20	20	20	15

24. Prepare price index numbers for 2017 with 2015 as base year from the following data, using (i) Laspeyres', (ii) Paasche's, (iii) Fisher's method.

Commodity	Unit	Quantity	Price (Tk.)	Quantity	Price (Tk.)
A	Kg.	5	2.00	7	4.50
B	Quintal	7	2.50	10	3.20
C	Dozen	6	3.00	6	4.50
D	Kg.	2	1.00	9	1.80

25. Using the data given below, calculate price index numbers for the year 2018 by (i) Laspeyres' formula, (ii) Paasche's formula, (iii) Fisher's formula, with the year 2009 as base:

Commodity	Price (Tk.)		Quantity ('000 kg.)	
	2009	2018	2009	2018
Rice	9.3	4.5	100	90
Wheat	6.4	3.7	11	10
Pulses	5.1	2.7	5	3

State with reasons one advantage of the Laspeyres' index over the Paasche index in case revisions of an index number are to be made from year to year.

26. Given the following data, calculate price index numbers by (i) Laspeyres' formula (ii) Paasche's formula, and (iii) Fisher's formula, with 2017 as base:

Year	Rice		Wheat		Jowar	
	Price	Qty.	Price	Qty.	Price	Qty.
2017	9.3	100	6.4	11	5.1	5
2024	4.5	90	3.7	10	2.7	3

27. Calculate the price index number for 2020 with 2017 as base year by the aggregative method, using (a) base year quantities as weights, and (b) given year quantities as weights, from the following data:

Commodity	2017		2020	
	Quantity ('000 tons)	Price per ton (Tk.)	Quantity ('000 tons)	Price per ton (Tk.)
A	350	100	400	120
B	200	130	180	200
C	140	50	200	100
D	80	125	100	140

28. The following table gives the change in the price and consumption of three commodities in the workers' consumption basket. Compute Fisher's ideal index number from the data given in the table.

Commodity	2010		2020	
	Quantity ('000 tons)	Consumption (units)	Quantity ('000 tons)	Consumption (units)
Wheat	100	10	110	6
Rice	150	15	170	18
Cloth	5	50	4	30

29. From the data given below, calculate Fisher's Ideal Index number of prices for 2023 with reference to 2020 as base period:

Commodity	Price (Tk.)		Quantity ('000 kg.)	
	2020	2023	2020	2023
A	4.3	5.2	20	16
B	2.1	3.9	5	4
C	0.8	1.6	11	8
D	3.2	4.8	8	6

30. Find by Arithmetic Mean method the index number from the following data:

Commodity	Base Price	Current Price
Rice	140	180
Sugar	100	300
Oil	400	550
Wheat	125	150
Pulse	160	200

31. Calculate a suitable index number from the data given below:

Commodity	Price Relative	Weight
A	125	5
B	67	2
C	250	3

32. Explain the term 'Price Relative'. Find by Arithmetic Mean method the index number from the following:

Commodity	Base Price	Current Price	Weight
Rice	30	52	8
Wheat	25	30	6
Fish	130	150	3
Potato	35	49	5
Oil	70	105	7

33. Using Paasche's formula, compute the quantity index and the price index numbers for 2020 with 2016 as base year:

Commodity	Quantity Units		Value Rs.	
	2016	2020	2016	2020
A	100	150	500	900
B	80	100	320	500
C	60	72	150	360
D	30	33	360	297

34. Using Fisher's 'ideal' formula, calculate the quantity index number from the following data:

Commodity	Base year Price (Tk.)	Base year Quantity (Kg.)	Current year Price (Tk.)	Current year Quantity (Kg.)
A	5	50	10	56
B	3	100	4	120
C	4	60	6	60
D	11	30	14	24
E	7	40	10	36

35. Annual production in million tons of three commodities are given:

Commodity	Production in year		Weights
	2015	2020	
A	160	200	13
B	10	12	21
C	80	100	35

Calculate quantity index number for the year 2020 with 2015 as base year, using simple arithmetic mean and weighted arithmetic mean of the relatives.

36. Using the following data, show that Laspeyres' price index formula does not satisfy the time reversal test:

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

37. Compute chain index numbers with 2010 prices as base, from the following table giving the average wholesale prices for the years 2010-2014.

Commodity	Average Wholesale Prices (Tk)				
	2010	2011	2012	2013	2014
A	20	16	28	35	21
B	25	30	24	36	45
C	20	258	30	24	30

38. From the table of group index numbers and group expenditures given below calculate the cost of living index number:

Group	Index Number	Percentage of Total Expenditure
Food	428	45
Clothing	250	15
Fuel & light	220	8
House rent	125	20
Others	175	12

39. The following are the group index numbers and the group weights of an average working class family's budget. Construct the cost of living index number.

Groups	Food	Fuel & lighting	Clothing	Rent	Miscellaneous
Index No.	352	220	230	160	190
Weight	48	10	8	12	15

40. The percentage increase in price in 2021 over 2010, in the following groups for middle class people in Dhaka and the percentage of total expenditure spent on those groups are shown below. Calculate the cost of living index number for 2021 with 2010 as base.

Group	Percentage increase in price	Percentage of total expenditure
Food	125	45
Clothing	66	6
Fuel & Lighting	112	5
House Rent & Tax	90	10
Miscellaneous	105	34

41. Determine the relative importance for the food group, given that the cost of living index number for 2015 with 2010 as base is 175 from the following figures:

Group	% increase in expenditure	Weight
Food	65	-
Clothing	90	12
Fuel etc	20	18
Miscellaneous	70	10
Rent etc	150	20

42. The following are index numbers of prices (1979 = 100):

Year	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Index	100	120	180	207	243	270	300	360	400	420

Shift the base from 1979 to 1985 and recast the index numbers.

43. The following table shows the Index Number of Wholesale Prices in Bangladesh (Revised Series) with base 2010-2011.

Year	2011	2013	2014	2015	2016	2017
Index Number	105	132	169	176	172	185

Find the index numbers for these years with base 2013 = 100.

44. Given below are two series of index numbers, one with 2011 as base and the other with 2020 as base:

(a) Year	Index	(b) Year	Index
2015	180	1970	100
2016	192	1971	108
2017	208	1972	112
2018	220	1973	125
2019	232	1974	130
2020	250	1972	150

The index number series (a) was discontinued in 2021 Splice the series (a) to the series (b) with 2020 as base.

45. Given below are the average wages in rupees per hour of unskilled workers of a factory during the years 2015-2020. Also shown is Consumer Price index for these years (taking 2015 as base year with Price Index 100). Determine the real wages of the workers during 2015-2020 compared with their wages in 1975.

Year	2015	2016	2017	2018	2019	2020
Consumer Price Index	100	120.2	121.7	125.9	129.3	140
Average Wage (Tk ./hour)	1.19	1.94	2.13	2.28	2.45	3.10

How much is the worth of one taka of 2015 in subsequent years?

PROBABILITY AND THE THREE IMPORTANT DISTRIBUTIONS



The theoretical aspects of probability have been first scented by an Italian Mathematician G-Cardano (1501-1576). The modern probability theory has been developed by Chebyshev, Markov etc. on this unit we have discussed definition of probability and its uses.

Lesson 1: The Concept of Probability

Lesson Objectives:

After studying this lesson you will be able to explain:

- Concept of probability;
- Properties of probability.

Introduction

The dictionary meaning of probable is likely and this meaning serves the purpose for the general conversational language. But such a vague meaning is insufficient for the purpose of scientific method and we would be able to describe a precise definition to the term for our purpose. In our everyday life we often come across comments like ‘the chance of raining today is high’, ‘there is little likelihood of rain tomorrow’, ‘his chance of success is very little’, etc. In all these cases the commentator has an idea of probability in his mind. So the idea of probability has got its place in common language.

Probability

One may state the probabilities in numerical terms such as “the chances are two in three that it will rain tomorrow” or “I have three chances out of ten of passing this test”. There is no mathematical basis for these statements. Such statements may be termed as subjective or personalistic interpretation of the probability. In this approach, probability is interpreted as a measure of degree of common belief or a quantitative judgment or common sense of an individual. The statisticians are not satisfied with such vague statement of probability but they have developed the theory of probability in numerical terms based on logical interpretation.

Foundation of mathematical probability was laid by Pascal and Fermat in the seventeenth century in solving the problems connected with games of chance. The famous Swiss Mathematician Jacob Bernoulli (1654-1705) and French Mathematician P.S.Laplace (1749-1827) have developed the theory of probability using the concept that different outcomes of a game or experiment are equally likely. This principle has several characteristics one of which is that it assumes symmetry of events. Thus we have a fair coin, a fair die or a fair deck of cards. A second characteristic is that it is based on abstract reasoning and does not depend on experience. The probability computed on the basis of such reasoning is known as a priori probability.

In this approach the probability of an event is the ratio of the number of cases favorable of that event to the total number of cases where all the cases are equally likely and mutually exclusive. The events are called mutually exclusive if they cannot occur simultaneously. For example a card cannot be both a spade and a club. So these two events (spade and club) are mutually exclusive. But one card can be both spade and ace. So these two events are not mutually exclusive. If there are n mutually exclusive equally likely cases and out of them m cases are favorable to a

The probability of an event is the ratio of the number of cases favorable of that event to the total number of cases where all the cases are equally likely and mutually exclusive.

If there are n mutually exclusive equally likely cases and out of them m cases are favourable to a particular event A then the probability of occurring the event A is $\frac{m}{n}$ and the probability of not occurring the event A is $\frac{n-m}{n}$.

particular event A then the probability of occurring the event A is $\frac{m}{n}$ and probability of not occurring the event A is $\frac{n-m}{n}$.

For example, if we toss a coin we may get either flower or otherside. If the coin is unbiased the chance of obtaining either flower or otherside are equal and, therefore, both the cases are equally probable or equally likely. They are also mutually exclusive because both the cases cannot occur simultaneously. Now out of these two mutually exclusive equally likely cases one is favorable to the event of showing flower. The probability of occurring flower is, therefore, $\frac{1}{2}$ similarly the probability of showing other side is $\frac{1}{2}$. A dice has six sides and if it is a perfect cube and made of the same metal probability of showing 3 on the top, when thrown, is $\frac{1}{6}$ because it has only one side with number 3.

Illustration 9.1

A box contains 5 white balls and 8 red balls, all of which are of equal size. A ball is drawn from the box at random. What is the probability that it is a white ball?

As there are 13 balls in the box any one of which may occur in the draw, we have 13 equally likely cases. Out of those, 5 cases are favorable to the event of a white ball. So the probability that the drawn ball is white is given by.

$$P(\text{White ball}) = \frac{\text{Favourable cases}}{\text{Total equally likely cases}} = \frac{5}{13}$$

Illustration 9.2

A card is drawn from a full packet of cards at random. What is the probability that it is (i) an ace, (ii) a red card.

(i) There are 52 cards in the packet any one of which may occur in the draw. So we have 52 total numbers of cases. As there are 4 aces in the packet out of these 52 cards, 4 cases are favorable to an ace. So the probability of drawing an ace is

$$P(\text{ACC}) = \frac{4}{52} = \frac{1}{13}$$

(ii) There are 26 red cards in the packet. So 26 cases out of 52 total cases are favorable to a red card.

$$\therefore \text{The probability of drawing a red card } P(\text{Red card}) = \frac{26}{52} = \frac{1}{2}$$

Properties of Probability

According to the definition the numerical measure of probability varies from zero to one, zero indicating impossibility and one meaning certainty. All other values between these two limits indicate doubtfulness. Probability that a man will go to the sun is zero indicating

impossibility and the probability that a man will die is one i.e. it is certain that the man will have to die one day. But the statement that it will rain today is doubtful and we can say like this only with certain degree of confidence or probability but with no absolute certainty.

Now let us see what the practical significance of this probability is. The statement that the probability of getting a flower in a coin toss is $\frac{1}{2}$ means that if we toss the coin a large number of times we will get close to 50% flower and 50% other side in the long run. This does not mean that in 20 tosses we will get exactly 10 flower and 10 other side but the proportion of flowers will approach the figure as we increase the number of tosses indefinitely. This notion may be applied in cases of economic and social phenomena also. Probability that the price of rice will rise in the month of July is 0.80 means that the price of rice increases in the month of July is 80% of the cases. The proportion of times that an event occurs actually is called its relative frequency. The concept of probability in terms of relative frequency was first formulated and proved by J. Bernoulli. His theorem goes as follows:

The numerical measure of probability varies from zero to one, zero indicating impossibility and one meaning certainty. All other values between these two limits indicate doubtfulness.

If the probability of occurrence of an event 'A' is P and if n trials are made independently and under the same conditions, then the probability that the relative frequency of A differs from P by an amount, however small, approaches zero as the number of trials tend to infinity. Symbolically the theorem is stated like as n tends to infinity, m/n tends to $P(A)$

In many of the statistical studies of economic and social phenomena the probability is estimated by using the relative frequency.

The probability may be obtained by using the past records. For example, we can say that the probability that a particular shop will succeed is 0.75 if we see that out of a large number of shops under similar conditions in the past 75% shops succeeded and 25% shops failed. In many of the statistical studies of economic and social phenomena the probability is estimated by using the relative frequency.

Theorem of Total Probability

The theorem of total probability is stated as: The probability of either of the two events A and B is equal to the probability of occurring the event A plus the probability of occurring the event B minus the probability of occurring the two events simultaneously. Symbolically

$$P(A \text{ or } B) = P(A) + P(B) - P(AB)$$

Where $P(AB)$ is the probability of their simultaneous occurrence.

If the two events are mutually exclusive, probability of their simultaneous occurrence is zero. So the probability of occurring either of the two mutually exclusive events is equal to the sum of the probabilities of individual events i.e. $P(A \text{ or } B) = P(A) + P(B)$. In general the probability that an event will occur in any one of the several mutually exclusive ways is the sum of probabilities of the various ways of occurrences. For example if we toss a coin we can get either a head or a tail. Probability of getting a flower is $\frac{1}{2}$ and that of getting other side is also $\frac{1}{2}$. Then the probability of getting either a flower or otherside is according to the theorem is $(\frac{1}{2} + \frac{1}{2}) = 1$.

If the two events are mutually exclusive, probability of their simultaneous occurrence is zero. So the probability of occurring either of the two mutually exclusive events is equal to the sum of the probabilities of individual events.

Illustration 9.3

In a dice throw what is the probability of getting less than 4?

We may get either 1 or 2 or 3.

$$\text{Probability of getting 1} = \frac{1}{6}$$

$$\text{Probability of getting 2} = \frac{1}{6}$$

$$\text{Probability of getting 3} = \frac{1}{6}$$

So the probability of getting 1 or 2 or 3 is the total $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

Illustration 9.4

A box contains 6 white ball, 7 red balls and 9 black balls. One ball is drawn from the box at random. What is the probability that it is a white or red ball?

$$\therefore \text{Probability of getting a white ball} = \frac{6}{22} = \frac{3}{11}$$

$$\text{Probability of getting a red ball} = \frac{7}{22}$$

$$\text{Probability of getting a white or a red ball is the total} = \frac{6}{22} + \frac{7}{22} = \frac{13}{22}$$

Theorem of Compound Probability

The theorem of compound probability is stated as, “the probability of simultaneous occurrence of two events A and B is given by the product of the unconditional probability of one event, say A , by the conditional probability of the other event i.e. B , supposing that A actually occurred.” Symbolically $P(AB) = P(A)P(B/A)$. If the two events are independent then $P(AB) = P(A)P(B)$.

Here $P(B/A)$ is called conditional probability of B .

Self-Assessment Questions:

Short Question

1. Define probability.
2. What do you mean by mutually exclusive?
3. Can you explain compound probability?
4. Explain total probability?
5. Define events.

Multiple-Choice Question:

1. **Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:**
 - (i) Who point out mathematical foundation of probability?
(a) Fisher (b) Pascal and Fermal (c) Fox (d) Pearson
 - (ii) The sum of all probability is equal to
(a) Zero (b) One (c) Ten (d) negative
2. **Write “T” if the statement is true and “F” if the statement is false:**
 - (i) The numerical measures of probability is $0 \leq P(A) \leq 1$.
 - (ii) $\text{Prob}[A \text{ Or } B] \neq P[A] + P[B] - P[AB]$

Answer:

Multiple Choice Questions

- (i)- b (ii)- b

True/False

- (i) T (ii) F

Lesson 2: Random Variables, Probability Distribution, the Binomial Distribution

Lesson Objectives:

After completing this lesson you will be able to explain

- Random variable;
- Probability distribution;
- Bi-nominal distribution;
- Different problem.

Introduction

Some time the out causes of experiment are quantitative then outcomes are expressed by numerical figure assigning real numbers of different sample points in a sample space. Now we discuss Random variable, probability distribution etc in this unit.

Random Variable

The variable is a characteristic of the population, the values of which vary from unit to unit. When each value of a variable has a probability associated to it then the variable is called a random variable. Let X be the outcome of a throw of a die. X has 6 possible outcomes such as 1,2,3,4,5,6 and to each outcome there is an associated probability of $1/6$. So X is a random variable.

There are two types of random variables - discrete and continuous.

Discrete Random Variable

In the dice throw example the variable X can assume only a finite number of values which can be counted and usually are integers. Such a variable is called a discrete random variable.

Continuous Random Variable

A random variable X is said to be continuous if it can assume any value in an interval limited by two values. For example, age of a person and height of a person can assume any value within certain range.

Probability Distribution

If $(A_1 A_2 \dots A_n)$ is a set of events which are mutually exclusive and exhaustive, then the corresponding set of probabilities, $P(A_1)$, $P(A_2)$, $P(A_3)$, ... $P(A_n)$ is called a probability distribution. For example if a die is thrown the probability distribution of the outcome X is given below:

Face coming up with X dots	1	2	3	4	5	6	Total
$P(X)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	1

In the above example, each value of the random variable has the same probability but this may not be the case always. The random variable r can be defined as the number of flowers (r) occurring in two flips of a fair coin. In this case probability distribution is:

Number of flowers r	0	1	2	Total
Probability of r	$1/4$	$1/2$	$1/4$	1

The variable is a characteristic of the population, the values of which vary from unit to unit. When each value of a variable has a probability associated to it then the variable is called a random variable.

A random variable X is said to be continuous if it can assume any value in an interval limited by two values.

The Binomial Distribution

P is used to denote the probability of a success and q to denote the probability of a failure in a single trial.

The binomial distribution is based on the assumption that there are two possible outcomes of an experiment or trial. As an example in a coin tossing experiment we can get either a flower or otherside. In the real world a population may consists of two categories of observations - one having a particular attribute and the other lacking the same. Occurrence of a particular observation of first category i.e., possession of the attribute is called a success and the occurrence of an observation of second category is called a failure. In coin tossing experiment occurrence of flower may be taken as 'success' and that of otherside as a 'failure'.

P is used to denote the probability of a success and q to denote the probability of a failure in a single trial.

Assumptions:

- 1) Events in a number of trials are independent
- 2) Probability of success and failure (p and q) remains constant throughout the experiment
- 3) $p + q = 1$

Derivation of the Probability Mass Function for Binomial Distribution

Let the experiment be repeated n times out of which r trials resulted successes and n-r trials gave failure. Then the probability of r successes and n-r failures (regardless of the order in which successes and failures occur) will be

$$\frac{p \text{ terms} \times q \text{ terms}}{(n-r) \text{ terms}} = p^r q^{n-r}$$

If we consider different order or arrangements of r success and (n-r) failures, we find that there may be n_{Cr} different arrangements. So the probability of r successes and n-r failures in n trials is given by $p(r) = n_{Cr} p^r q^{n-r}$

$$\text{Where, } n_{Cr} = \frac{n!}{r!(n-r)!}$$

In a coin tossing experiment 2 flowers can be obtained out of 5 coin tossing in 5_{C2} ways. So the probability of getting 2 flowers in 5 tosses is,

$$p = 5_{C2} (1/2)^2 (1/2)^3 = 10/32$$

Properties of Binomial Distribution

1. Binomial distribution is a discrete probability distribution and the total probability of all the events is one i.e.,

$$\sum_r^n P(r) = 1$$

2. Arithmetic mean of r is np and standard deviation is $\sqrt{npq}$ where $q = 1 - p$
3. The distribution is symmetrical when $p = q$
4. It tends to normal distribution when n is large.
5. Skewness is measured by $q - p$

It tends to normal distribution when n is large.

Problem 1 : 5 fair coins are tossed. Find the probability distribution of the number of flower (r).

Solution 1:

Number of flowers (r)	Probability of r
0	$5c^0 (1/2)^0 (1/2)^5 = 1/32$
1	$5c^1 (1/2)^1 (1/2)^4 = 5/32$
2	$5c^2 (1/2)^2 (1/2)^3 = 10/32$
3	$5c^3 (1/2)^3 (1/2)^2 = 10/32$
4	$5c^4 (1/2)^4 (1/2)^1 = 5/32$
5	$5c^5 (1/2)^5 (1/2)^0 = 1/32$
Total	1

Problem 2: A set of 5 coins are tossed 64 times and the number of coins showing up flower each time is recorded. How many times you expect 3 flowers? Find the expected mean and standard deviation of the number of flowers.

Solution 2:

Probability of getting 3 flowers out of 5 coins tossed is,

$$P = 5C_3 (1/2)^3 (1/2)^2 = 10/32$$

So, the expected frequency of 3 flowers in 64 trials = $64 \times (10/32) = 20$

$$\text{Mean} = np = 5 \times (1/2) = 2.5$$

$$\text{Standard deviation} = \sqrt{npq} = \sqrt{5 \times (1/2) \times (1/2)} = \sqrt{5/4} = 1.12$$

Use Binomial distribution is used in economical, business and industrial experiments.

Derivation of the Mean and Variance of the Binomial Distribution:

The expected frequency of the binomial variate r is given by the formula,

$$f = NP = N nCr p^r q^{n-r}, \text{ where } N = \text{Total frequency}$$

So the mean of f is given by

$$\bar{f} = \sum_{r=0}^n rnCr p^r q^{n-r}$$

[Here relative frequency is taken total of which is one.]

$$= \sum \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

$$= \sum \frac{n(n-1)!}{r!(r-1)!(n-r)!} p^r q^{n-r}$$

$$= np \sum \frac{(n-1)!}{(r-1)!(n-r)!} p^{r-1} q^{n-r}$$

$$= np (p + q)^{n-1}$$

$$= np$$

because $(p + q) = 1$.

Variance of r

$$r = \sum (r - np)^2 nCr p^r q^{n-r}$$

$$= \sum (r^2 - 2npr + n^2 p^2) \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

$$= \sum \{ r(r-1) + r(1-2np) + n^2 p^2 \} \frac{n!}{r!(n-r)!} p^r q^{n-r}$$

$$= \sum \frac{r(r-1)n(n-1)(n-2)!}{r(r-1)(r-2)!(n-r)!} p^r q^{n-r} + \sum \frac{rn(n-1)!}{r(r-1)!(n-r)!} p^r q^{n-r}$$

$$- 2nmp \frac{n(n-1)}{r(r-1)!(n-r)!} p^r q^{n-r} + n^2 p^2 \sum nCr p^r q^{n-r}$$

$$= p^2 n(n-1)(p+q)^{n-2} + np \sum (p+q)^{n-1} - 2n^2 p^2 (p+q)^{n-1} + n^2 p^2$$

$$= n^2 p^2 - np^2 + np - 2n^2 p^2 + n^2 p^2$$

$$= np - np^2 = np(1-p) = npq$$

Self-Assessment Questions:

Short Question

1. What do you mean by probability?
2. Define random variable?
3. Distinguish between discrete and continuous random variable.
4. Write down a binomial density function.

Multiple-Choice Question:

1. Write “T” if the statement is true and “F” if the statement is false:

- (i) The mean of binomial distribution is npq .
- (ii) The sum of probability is one
- (iii) The distribution is symmetric when $p=q$

2. Fill up

- (i) Binomial distribution tends to normal distribution when n _____
- (ii) The variance of binomial distribution is _____

Answer:

True/False

- (i) F (ii) T (iii) T

Fill up

- (i) in large
- (ii) npq

Lesson 3: Normal Distribution

Lesson Objectives:

After complete this lesson you will be able to explain

- Normal distribution;
- Normal curve;
- Solve the different types of problem.

Introduction

The normal distribution is a continuous distribution. De-Moivre (1667-1754) first discovered the normal distribution in 1733 as a limit case of the binomial distribution. Gauss describes successfully in the distribution of errors of measurement.

Normal Distribution

Normal distribution is a continuous, symmetrical and mesokurtic probability distribution. Its probability function is given by

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where μ is the mean and σ is the standard deviation of the distribution. Probability that the value of x lies between any two values (say u and v) is given by,

Properties of Normal Distribution

$$P(u \leq x \leq v) = \int_u^v \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

1. Total probability is one i.e.,
2. Mean of the normal variate x is μ and standard deviation is σ .
3. The distribution is symmetrical i.e., mean, median and mode are equal and $B_1=0$.
4. All odd moments are zero
5. The distribution is mesokurtic and $B_2=3$.
6. (i) 99.7% of the values lie between $x - 3\sigma$ and $x + 3\sigma$. These two limits are known as 3σ limits of normal distribution.
 (ii) 95.4% values lie between 2σ limits i.e., between $x - 2\sigma$ and $x + 2\sigma$.
 (iii) 68% values lie between 1σ limits i.e., between $x - \sigma$ and $x + \sigma$.

Normal Curve

The frequency curve representing the normal distribution is known as normal curve. It is a bell shaped, symmetrical, mesokurtic curve. The area under the curve represents the probability; therefore the total area under the curve is taken as one. 68% of the total area are contained between the two ordinates drawn at $x - \sigma$ and $x + \sigma$, 95.4% area between the ordinates drawn at $x - 2\sigma$ and $x + 2\sigma$ and 99.7% area between the ordinates drawn at $x - 3\sigma$ and $x + 3\sigma$.

The frequency curve representing the normal distribution is known as normal curve. It is a bell shaped, symmetrical, mesokurtic curve. The area under the curve represents the probability; therefore the total area under the curve is taken as one.

The curve drawn below shows these limits.

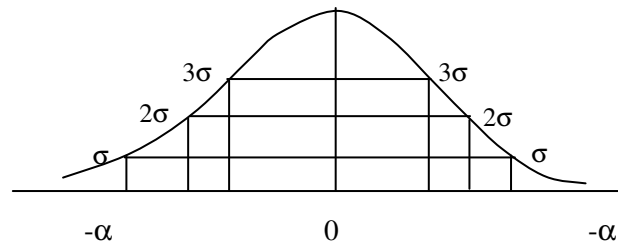


Fig 9.1: normal curve

Standard Normal Distribution

A particular and important special case occurs when $\mu = 0$ and $\sigma = 1$. In this case mean is zero and standard deviation is unity; the probability density function is given by,

We can transform any normal variate x into a standard normal variate by

$$\frac{x-\mu}{\sigma} = Z$$

using the simple transformation rule:

For example if x is a normal variate with mean = 40 and standard

$$Z = \frac{x-40}{8}$$

variation = 8, then

Since we look up the probabilities in terms of Z and all Z 's have mean zero and a standard deviation of one, we need only one table of probabilities. Probability of Z below a particular value (i.e., $p(z \leq a)$) is given in this table.

Problem 1; Find the probabilities for a normally distributed random variable x of mean 6 and standard deviation 2.

- (i) $P(x > 8)$, (ii) $P(x \leq 8)$, (iii) $P(8 \leq x \leq 12)$

Solution: Corresponding values of Z

- (i) When $X = 8$, $Z = (8 - 6)/2 = 1$

So, $P(X > 8) = P(Z > 1) = 1 - P(Z \leq 1) = 1 - 0.8413 = 0.1587$ (from normal integral table)

- (ii) $P(X \leq 8) = P(Z \leq 1) = 0.8413$

- (iii) $P(8 \leq X \leq 12) = P(1 \leq Z \leq 3)$
 $= P(Z \leq 3) - P(Z \leq 1) = 0.9987 - 0.8413 = 0.1574.$

Normal Approximation to Binomial

When number of trials n in a binomial distribution is large and p is moderate the distribution tends approximately to a normal distribution. Let r denotes the number of successes then,

$$\frac{r-np}{\sqrt{npq}} = Z$$

Problem 2 : What is the probability of getting number of heads below 15 in 30 throws of a fair coin?

Solution 2: Here, $p = 1/2$, $q = 1/2$, $n = 30$, $r = 15$.

So, $np = 30 \times 1/2 = 15$, $npq = 30 \times 1/2 \times 1/2 = 7.5$

$Z = (r - 15) / \sqrt{7.5} = (15 - 15) / \sqrt{7.5} = 0$ (when $r = 15$)

The required probability, $p(r \leq 15) = p(Z \leq 0) = 0.5$

Derivation of Mean and Variance of Normal Distribution:

$$\begin{aligned} \text{Mean} &= \int_{-\infty}^{\infty} \frac{(X - m)}{\sqrt{2\pi\sigma}} e^{-\frac{(X-m)^2}{2\sigma^2}} dX + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} m e^{-\frac{(X-m)^2}{2\sigma^2}} dX \\ &= \frac{\sigma^2}{\sqrt{2\pi\sigma}} \left[e^{-\frac{(X-m)^2}{2\sigma^2}} \right]_{-\infty}^{\infty} + m = m. \end{aligned}$$

$$\begin{aligned} \text{Variance} &= \int_{-\infty}^{\infty} \frac{(X - m)^2}{\sqrt{2\pi\sigma}} e^{-\frac{(X-m)^2}{2\sigma^2}} dX \\ &= -\frac{\sigma^2}{\sqrt{2\pi\sigma}} \left[\left((X - m) e^{-\frac{(X-m)^2}{2\sigma^2}} \right) \right]_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-\frac{(X-m)^2}{2\sigma^2}} dX \\ &= 0 + \sigma^2 = \sigma^2. \end{aligned}$$

Use of Normal Distribution

1. Normal distribution is used to test on t, F, X^2 etc.
2. If it and in business, Economic field of experiment.

Example

The tea consumption in an area is assumed to follow normal distribution with mean consumption of 200 kg/day and with S.D = 50 kg/day. Find the probability that, in a day

- (i) The consumption will be more than or equal to 250 kg
- (ii) The consumption will be within the limit 250 kg to 300 kg

Solve:

Let $X \sim N(200, 250)$

$$(i) Z = \frac{X - M}{\sigma} = \frac{250 - 200}{50} = 1 \text{ Now}$$

$$\text{Prob}[Z \geq 1] = 1 - P[Z \leq 1] = 1 - .84134 = .15866$$

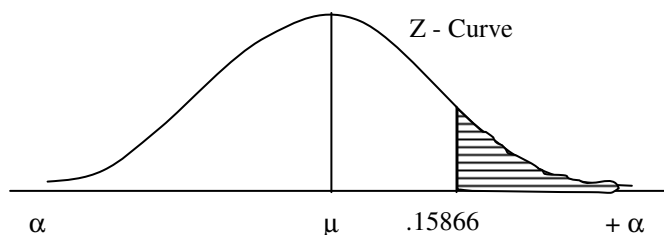


Fig 9.2:

$$(ii) \text{ Again } Z_1 = \frac{250 - 200}{50} = 1$$

$$(iii) \quad Z_2 = \frac{30 - 200}{50} = 2$$

$$\begin{aligned} \text{We have prob } [1 \leq Z \leq 2] &= \text{Prob } [Z \leq 2] - \text{Prob } [Z \leq 1] \\ &= .97225 - .80134 \\ &= 0.1309 \end{aligned}$$

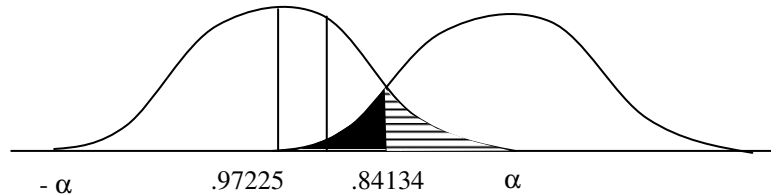


Fig :

Self-Assessment Questions:

Short Question

1. What do you mean by normal distribution
2. Define the dewrite function of normal distribution.
3. Write down the mean and variance of normal distribution

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The mean of the normal distribution is
(a) np (b) μ (c) $n\mu$ (d) μ^2
- (ii) The variance of the normal distribution is
(a) 3σ (b) npq (c) σ^2 (d) μ^2
- (iii) The normal distribution is point out as a limiting binomial distribution in
(a) 1993 (b) 1733 (c) 1833 (d) 1639

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) All odd moments of normal distribution are Zero.
- (ii) The distribution is symmetric of the normal distribution
- (iii) The total probability of normal distribution is “0”

3. Fill up :

- (i) _____ values lies between $x - 3\sigma$ and $x + 3\sigma$
- (ii) _____ value lies between 2σ limit
- (iii) _____ value lies between 1σ limit.

Answer:

Multiple Choice Questions: (i)- b (ii)- c (iii)- b

True/False: (i) T (ii) T (iii) F

Fill up: (i) 99.7% (ii) 95.4% (iii) 68%

Lesson 4: Poisson Distribution

Lesson Objectives:

After complete the lesson you will be able to explain

- Poisson distribution;
- Mean and variance of the Poisson distribution;
- Properties of the Poisson Distribute.

Introduction

Devis Poisson (1781-1840) discovered Poisson distribution. This distribution is also related to the experimental result express as success and failure under some limiting condition.

Poisson Distribution

When the probability of a success is very small and the number of trials n is very large the binomial distribution

$P(r) = {}^nC_r p^r q^{n-r}$ can be reduced to the form

$$P(r) = \frac{m^r e^{-m}}{r!}$$

Where $m=np$, r = The number of successes, e is the constant and
 $r! = 1 \times 2 \times 3 \times \dots \times r$.

Number of Successes (r)	Probabilities
0	e^{-m}
1	me^{-m}
2	$m^2 e^{-m}/1 \times 2$
3	$m^3 e^{-m}/1 \times 2 \times 3$
4	$m^4 e^{-m}/1 \times 2 \times 3 \times 4$
5	$m^5 e^{-m}/1 \times 2 \times 3 \times 4 \times 5$
r	$\frac{m^{r-m}}{r} e$

Poisson Distribution has got applicability where the probability of success is very small. For example, the number of accidents happened in Dhaka city roads per day will follow the Poisson Distribution.

Properties of Poisson Distribution

Poisson distribution has got certain properties. These are:

- (1) The sum of the probabilities of all successes is one

$$\sum_{r=0}^{\infty} \frac{m^r e^{-m}}{r!} = 1$$

- (2) The arithmetic mean of the Poisson distribution is m .
 (3) Variance of the Poisson distribution is m .
 (4) The Poisson distribution is skewed and measure of skewness is, $\sqrt{1/m}$
 (5) The distribution is platykurtic and

Derivation of Poisson Distribution is Shown Below.

In the binomial distribution of p is indefinitely small and n is increased sufficiently so that np is finite, the probability of r takes the form of where

$$m = np.$$

Proof:

The binomial form $nCr p^r q^{n-r}$ may be written as,

$$\begin{aligned} \frac{n!}{(n-r)!r!} p^r q^{n-r} &= \frac{n!}{r!(n-r)!} \left(\frac{m}{n}\right)^r \left(1 - \frac{m}{n}\right)^{n-r} \\ &= \frac{m^r}{r!} \left(1 - \frac{m}{n}\right)^n \times \frac{n!}{(n-r)!n^r \left(1 - \frac{m}{n}\right)^r} \end{aligned}$$

[since, $m=np$, $p=m/n$ and $q=1-p=1-(m/n)$]

Now as n tends to infinity the value of $(1 - m/n)^n$ becomes e^{-m} in the limit. By applying Stirling's approximation for factorial the limiting value of $n!$ as n tends to infinity is

$$\sqrt{(2\pi n)} n^n e^{-n}$$

we have,

$$\begin{aligned} &= \frac{2\pi n n^n e^{-n}}{2\pi(n-r)(n-r)^{n-r} e^{-(n-r)} n^r \left(1 - \frac{m}{n}\right)^r} \\ &= \frac{2\pi n^{n+\frac{1}{2}} e^{-n}}{2\pi(n-r)(n-r)^{n-r} e^{-(n-r)} n^r \left(1 - \frac{m}{n}\right)^r} \\ &= \frac{1}{\frac{(n-r)^{n-r+\frac{1}{2}}}{n^{n-r+\frac{1}{2}}} e^r \left(1 - \frac{m}{n}\right)^r} \\ &= \frac{1}{\left(1 - \frac{r}{n}\right)^{n-r+\frac{1}{2}} \left(1 - \frac{m}{n}\right)^r} \end{aligned}$$

The limiting value of which is, $\frac{1}{e^r - e^{-r}} = 1$

Because as n tends to infinity both $(1-r/n)^{r+1/2}$ and $(1-m/n)^r$ tend to unity.

Hence, $nCr p^r q^{n-r} = (m^r e^{-m})/r!$ for indefinitely small values of p .

In the binomial distribution of p is indefinitely small and n is increased sufficiently so that np is finite, the probability of r takes the form of where $m=np$.

Mean of Poisson distribution,

$$r = \sum_{r=0}^{\infty} \frac{rm^r e^{-m}}{r!} = \sum_{r=1}^{\infty} \frac{rm^{r-1} e^{-m}}{(r-1)!} = m \sum_{r=1}^{\infty} \frac{m^{r-1} e^{-m}}{(r-1)!} = m$$

$$\begin{aligned} \text{Variance of Poisson Distribution} &= \sum (r-m)^2 \frac{m^r e^{-m}}{r!} = \sum (r^2 - 2rm + m^2) \frac{m^r e^{-m}}{r!} \\ &= \sum \{r(r-1) + r - 2rm + m^2\} \frac{m^r e^{-m}}{r!} \\ &= \sum \frac{r(r-1)m^2 m^{r-2} e^{-m}}{r(r-1)(r-2)!} + (1-2m) \sum \frac{rm^r e^{-m}}{r!} + m^2 \sum \frac{m^r e^{-m}}{r!} \\ &= m^2 \sum \frac{m^{r-2} e^{-m}}{(r-2)!} + (1-2m)m + m^2 = m^2 + m - 2m^2 + m^2 = m \end{aligned}$$

Standard Deviation = $\sqrt{m}$

Use an Application of Poisson Distribution

Poisson distribution is suitable for the following area:

1. Number of associated at a crossing per hour during the busy true of the day.
2. Number of plants fail to work during full production process in a large industry.
3. Number of wrong connection received at telephone exchanges.
4. Number of complimentary copies of a book in large packet of bank.
5. Number of firstly bulbs in a packed of 100 bulbs etc.

Self-Assessment Question:

Short Question

1. What do you mean by Poisson distribution
2. Explain the density function of Poisson distribution
3. Find out the mean and variance of Poisson distribution

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The mean of the Poisson distribution is
(a) np (b) npq (c) m (d) 3m
- (ii) The variance of the Poisson distribution is
(a) nσ (b) σ² (c) npq (d) m

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The Poisson distribution is play-to-caustic
- (ii) The Poisson distribution is not skewed
- (ii) The measures of skewness is $\sqrt{\frac{1}{m}}$

3. Fill up

- (i) The _____ of Poisson distribution is m
- (ii) The _____ of the Poisson distribution is m.

Answer:

Multiple-Choice Questions:

- (i)- c (ii)- d

True/False

- (i) T (ii) F (iii) T

Fill up

- (i) mean
- (ii) variance

Exercise

1. Define probability. Give an example of probability cases.
2. (a) State theorem of the total probability;
(b) In a box there are 8 red and 4 yellow balls. Two balls are drawn one after another without replacement. Find the probability that
 - (i) both ball will be red
 - (ii) one ball red and another ball yellow.
3. Mention some cases which are suitable for binomial probability model. Find mean and variance of binomial distribution.
4. What is Poisson distribution? Find the mean and variance of poisson distribution.
5. Derive poisson distribution as a limiting form of binomial distribution.
6. Define normal distribution. Write down the chief characteristics of normal distribution.
7. Let $X \sim N(50,10)$ Find
 - (i) $P[X \geq 45]$
 - (ii) $P[45 < X < 55]$
 - (iii) $P[X < 50]$

TEST OF HYPOTHESIS



If we consider, a particular hypothesis is true but find that the results observed in a random sample differ markedly from the results expected under the hypothesis, then we would say that the observed differences are significant and would these be inclined to reject the hypotheses. For example, If we get 20 heads in 50 tosses of a coin, we would be inclined to reject the hypothesis that the coin is fair, although it is conceivable that we might be wrong.

Lesson 1: Hypothesis Testing

Lesson Objectives:

After studying this lesson, you will be able to explain

- The definition of statistical hypothesis;
- Null hypothesis and alternative hypothesis;
- Level of significance.

Introduction

Procedures that enable us to determine whether observed samples differ significantly from the result expected, and thus help our decision whether to accept or reject the hypothesis, are called hypothesis testing,

Statistical Hypothesis

When we attempt to reach a decision, it is useful to make assumptions about the population. Such assumptions, may or may not be true, are called statistical hypothesis which is generally implies about the probability distributions of the population.

Statistical inference can help answer two somewhat similar, commonly posed questions about a population.

Case-1. Does a particular sample of observations belong to a hypothesized population of observation?

Case-2. Do observations on two groups of subjects differ from one another? i.e. do the sets of observations represent samples from identical population or from different population.

The two examples are given below for case-I and Case-II to answer the above question posed a procedure known as hypothesis testing.

Case-1. Suppose, for example, that Bangladesh Open University wanted to estimate the mean reading level of high score in SSC program i.e. the entire population shown in fig. 10.1

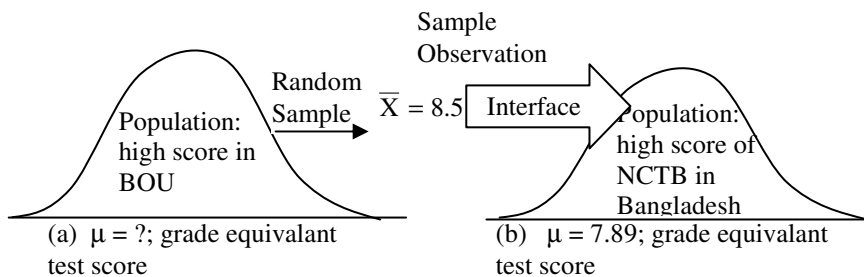


Fig. 10.1: Schematic representation of the steps involved in statistical inference from sample data to the population.

Since testing the high level score of all of the tutorial center in different R.R.C at one time is difficult and costly, a random sample of student is tested in the 18th year. From this sample of test scores, the mean of the

When we attempt to reach a decisions, it is useful to make assumptions about the population. Such assumptions, may or may not be true, are called statistical hypothesis.

sample distribution is calculated see in Fig. 8.2. This statistic is used to estimate the mean of high care of the entire population for SSC program in fig: 10.2.

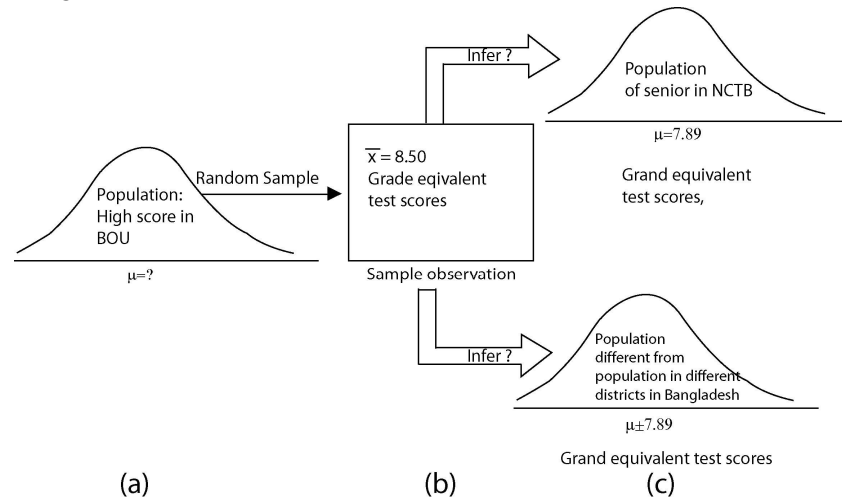


Fig:8.2 Schematic representation of the steps in inference-case-1

The next question, the BOU might ask about the mean reading level of its SSC program from where it differs from NCTB in the entire Bangladesh. Suppose that from national test norms, the mean reading level of SSC Examination is known to be 7.89 on a grade-equivalent scale. The question of interested then is: how likely are we to observe a sample mean of 8.50 or greater when, in facts' the sample was drawn from a population with a mean of 7.89?

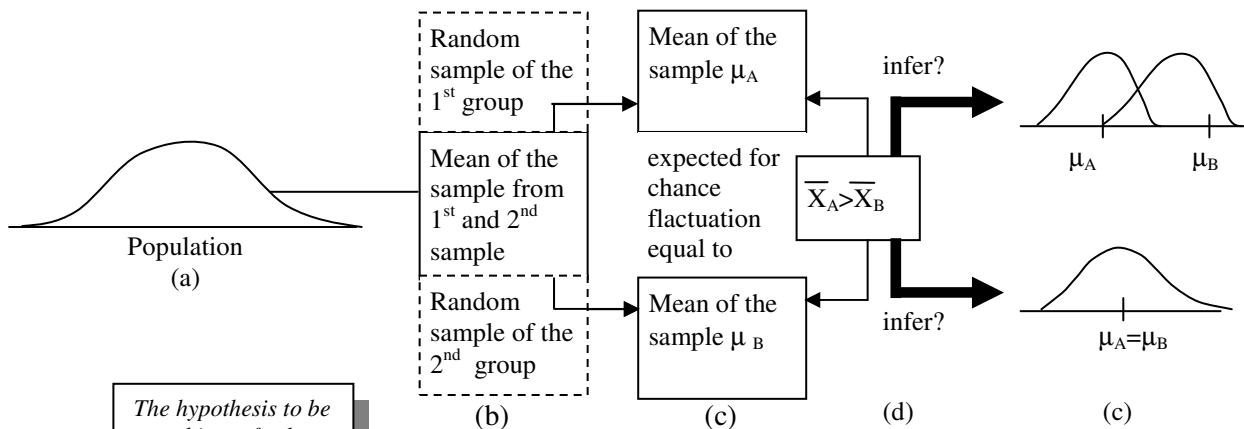


Fig : 10.3 Schematic representation with different means.

The answer to this question will help the university decide whether the high score level in BOU are similar in reading level of NCTB in Bangladesh. Clearly, if a sample mean of 8.50 or greater is likely to occur when random sample are drawn from a population with $\mu= 7.89$, the conclusion is that the high score level education of BOU is similar to those in the country as a whole seems to be warranted. However, if a sample mean of 8.50 or greater is quite unlikely to occur when the population mean is 7.89, the conclusion that high score level differs in

reading level from high score in other district seems reasonable. In short, should we infer that the sample data were drawn from a population with $\mu=7.89$ or, should we conclude that they were drawn from a different population with $\mu>7.89$? This problem of inference is given in fig 10.3.

For making a decision on whether a sample mean differs from some known or expected population mean, we use the following strategy:

The hypothesis to be tested is set forth. This hypothesis is called the null hypothesis, denoted as H_0 . For reading study, the null hypothesis would be :

$$H_0 : \mu = 7.89$$

An alternative hypothesis, denoted as H_A may be taken one of the three forms:

- The population mean is not equal to some specified value i.e. $H_A : \mu \neq 7.89$. This is called a non directional alternative hypothesis.
- The population mean is less than some specified value i.e. $H_A : \mu < 7.8$
- The population mean is greater than some specified value i.e. $H_A : \mu > 7.89$; the two form (b) and (c) are called directional alternative hypothesis.

The following discussion enumerates the steps in hypothesis testing:

- Assume that the sample was drawn from the known and expected population. This assumption is to be tested, is known as the null hypothesis H_0 , defined as $H_0: \mu = \text{some specific value}$, where
 H denoted as “hypothesis”
 0 denoted as “null”
 μ denoted as the population mean.
- Test the assumption of the null hypothesis against an alternative hypothesis denoted as H_A . The alternative hypothesis asserts that the sample different from the population specified in the null hypothesis i.e. it asserts that the sample is drawn from a different population than the one specified in the null hypothesis. The alternative hypothesis is defined as—
 $H_A : \mu \text{ some specified value; the sign } \neq \text{ denote “not equal”}. \text{ and}$
 $H_A: \mu > \text{some specified value; the sign } > \text{ denotes and “greater than”}.$
i.e. $H_A: \mu < \text{some specified value; the sign } < \text{ denoted as “less than”},$
- In order to test the null hypothesis, draw a random sample of subjects from the population of interest.
- From the sample mean, decide whether or not to reject the null hypothesis.

Test the assumption of the null hypothesis against an alternative hypothesis denoted as H_A

If the sample information leads us to reject H_0 , then we accept the alternative hypothesis.

Null hypothesis: The null hypothesis asserts that there is no true difference in the sample statistic and population parameter under consideration. The null hypothesis is denoted as H_0 .

Alternative hypothesis: The hypothesis that is different from null hypothesis, H_0 is called alternative hypothesis. If the sample information leads us to reject H_0 , then we accept the alternative hypothesis. The alternative hypothesis is denoted as H_A .

Level of Significance: In testing a given hypothesis, the maximum probability with which we would be willing to risk i.e. reject the null hypothesis when it should be accepted is called the level of significance or significance level of the test. This probability, often denoted by α , is generally specified before any sample is drawn so that the results obtained will not influence our choice.

Case-II. Two groups are drawn from populations with equal means:

Let us consider, there are two groups of experiment. The steps in reaching a decision about whether the sample difference represents differences in population. The steps in hypothesis testing are given below:

1. Assume that the samples represent populations with the mean then the null hypothesis H_0 , represent that there is no treatment effect and the null hypothesis is denoted as–

$H_0: \mu_A = \mu_B$; where

H denotes the hypothesis

O denotes the null

μ_A denote the mean of the first group

μ_B denote the mean of the second group.

2. Test the null hypothesis against an alternative hypothesis H_A . The alternative hypothesis asserts that the null hypothesis is incorrect. The alternative hypothesis is denoted as–

$H_A: \mu_A \neq \mu_B$ or

$H_A: \mu_A > \mu_B$ or

$H_A: \mu_A < \mu_B$

Thus, H_A specifies a difference between the mean of the population.

3. To test these hypothesis, draw a random sample of some known population from different groups.
4. Perform the experiment and observe the outcome in each group.
5. From the sample data, decide whether or not to reject the null hypothesis, i.e. the observed mean between the groups' means are equal. If we decide to reject the null hypothesis i.e. the group means of the groups are not equal.

Errors in Statistical Inference:

In making inferences about is true of the population or populations, we can never be certain that our decision to reject or not to reject the null hypothesis will be correct. The possibility of making the wrong decision is always there. An error can be made in two ways, which is represented in Table-10.4

Decision	True state of affairs in the population	
	Identical population	Different population
Identical population (Not Reject H_0)	correct (decession)	Error
Different populations (Reject H_0)	Error	Correct (Decesion)

Fig.-10.4: Decesion matrix

Fig. 10.3: Schematic representation of the steps involved in statistical inference from an experiment.

Self-Assessment Questions:

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The hypothesis is denoted as
 - (a) "A" (b) H_y
 - (c) H (d) H
- (ii) The null hypothesis is denoted as
 - (a) H_A (b) H_0
 - (c) H_y (d) H_y
- (iii) The alternative hypothesis is denoted as
 - (a) H_A (b) H_0
 - (c) H_A (d) H_0
- (iv) Prob ability [rejecting null hypothesis when it is accepted] is called
 - (a) null hypothesis (b) alternative hypothesis
 - (c) level of significance (d) hypothesis.

2. Write "T" if the statement is true and "F" if the statement is false:

- (i) The null hypothesis is denoted as H_0
- (ii) Prob ability [Rejecting H_0 | H_0 is accepted] is called level of significance
- (iii) The alternative hypothesis is denoted by H_0 .

Answer:

Multiple Choice Questions

(i)- a (ii)- b (iii)- a (iv)- c

True/False

(i) T (ii) T (iii) F

Lesson 2: Types of Decisions and Error

Lesson Objectives:

After studying this lesson, you will be able to explain

- Type one error and type two error;
- One tailed test and two tailed test with examples.

Introduction

We use statistical inference to make decisions about the validity of computing statistical hypothesis by our test, (H_0). Note that our initial decision always deals with H_0 , the statistical hypothesis tested. Induced, then we use we statistical inference in order to come to a decision as to which of the statements is correct. However, in any decision process, one must alone for making correct or incorrect decisions or error. In this lesson you are able to explain how the error might have to come into a decision and also known about type one error and type two error.

Decision and Error:

Let us consider that the School of Business of Bangladesh Open University has reason to believe that the diploma in management students is not of average scholastic aptitude. If this is true, it would have important implication for both the objectives and curriculum of the system. In order to put his research hypothesis into a form capable of being tested, the Dean decides that scholastic aptitude will be measured by National Aptitude Test (NAT) and that the competing statistical hypothesis will be:

$$H_0 : \mu = \mu_{NAT}$$

$$H_A : \mu \neq \mu_{NAT}$$

Wherever,

μ = the mean corresponding to the Dean's population of interest

μ_{NAT} = the mean for the National Aptitude Test

At the conclusion of his study, the Dean will decide either the students are of average aptitude (H_0 is true) or that the students are not of average aptitude (H_0 is false). We must keep in mind that while H_0 is either true or false, the Dean decides either to accept H_0 (he believes H_0 to be true) or to reject H_0 (he believes H_0 to be false). In making his decision, the Dean never knows for certain whether he has made a correct or an incorrect choice. Since he never knows whether H_0 is true or false (if he did, he wouldn't have to collect data to help him decide in the first place). We represent all possible contingencies as follows:

- i. H_0 is true : Dean declares that H_0 is true.
- ii. H_0 is true: Dean declares that H_0 is false
- iii. H_0 is false: Dean declares that H_0 is true.
- iv. H_0 is false: Dean declares that H_0 is false.

H_0 is true : dean declares that H_0 is true.

Note that in above (i) and (iv), the Dean would make correct decision and in (ii) and (iii) he would make incorrect division or error. In general the situation might be represented in table-(A)

Table-(A):**Correct and incorrect decisions in hypothesis testing**

Decisions	Ho false	Ho true
reject Ho	Correct decision	type I error
accept Ho	type II error	Correct decision

The error represented by rejecting a true Ho differs in kind from an error represented by accepting a false Ho.

It is important to note that the error represented by rejecting a true Ho differs in kind from an error represented by accepting a false Ho. We designated the former as errors of the first kind or type I error and also designate the later as errors of the second kind or type II error.

In our example, for instant, the dean would have made a type I error if he declared that the students were not of average aptitude when indeed they were. He would have committed a type II error if he decided that they were of average aptitude when indeed they were not.

Let α represents the probability of rejecting the null hypothesis given that the null hypothesis is true

In classical statistics, we present the probability of a type I error, and then attempt to minimize the probability of type II error. Let α represents the probability of rejecting the null hypothesis given that the null hypothesis is true i.e.

$$\alpha = \text{Prob. \{ rejecting null hypothesis | null hypothesis is true \}}$$

$$\text{or, } \alpha = \text{Prob. [rejecting Ho | Ho is true]}$$

$$\text{or, } \alpha = \text{Prob. [type I error]}$$

and β represents the probability of accepting the null hypothesis given that the null hypothesis is false i.e.

$$\beta = \text{Prob. \{ accepting the null hypothesis | null hypothesis is false \}}$$

$$\text{or } \beta = \text{Prob [accepting Ho | Ho is false]}$$

$$\text{or } \beta = \text{Prob [type II error]}$$

The probabilities of making correct or incorrect decision in hypothesis testing are given in table - (B)

Table- (B):**Probabilities of making correct or incorrect decisions in hypothesis testing.**

Decisions	Ho false	Ho true
reject Ho	$1 - \beta$	α
accept Ho	β	$1 - \alpha$

Rejecting Ho, when Ho is true, we will make correct decision $100(1-\alpha)\%$ of the level.

Here α is called the level of significance. In making decision of rejecting Ho, when Ho is true, we will make correct decision at confidence level $100(1-\alpha)\%$.

From this, one can estimate the power of a statistical test either before or after concluding the experiment.

Power = Probability of rejecting a False null hypothesis. i.e. Power = P [rejecting a False H_0] = $1 - \beta$.

$$\text{Power} = P[\text{rejecting a false } H_0] = 1 - \beta.$$

General Overview of Statistical Hypothesis Test: We partition statistical hypothesis testing into following steps:

- A. Selection of the computing hypothesis
- B. Consideration of assumption.
- C. Selection of appropriate test statistic and determination of sampling distribution.
- D. Selection of size of α and determine $\alpha\%$ level of significance.
- E. Calculation of data, calculation of observed value of statistic.
- F. Observed value of statistic entered into decision rule, with resulting conclusion about null hypothesis (H_0).
- A. The selection of computing hypothesis: We usually have H_A correspond to H_0 representing the complement.

Null hypothesis - H_0 : not our research hypothesis

Alternative hypothesis - H_A : our research hypothesis.

B. Assumptions:

- i. normal population
- ii. sample that are independent random sample from the population.

- C. Test Statistic and sampling distribution: Taking into consideration that null hypothesis and the assumptions, we specify the appropriate test statistics and its corresponding sampling distribution.

mean of x : $\mu_x = \mu$ (μ = population mean)

standard deviation of x , $S_x = \frac{\sigma}{\sqrt{n}}$ (σ = population standard deviation and n = number of units in the sample)

- D. Selection of size α and determine $\alpha\%$ level of significance: we typically present at some particular value $\alpha = .05$ or $\alpha = .01$ and the size of n , by a consideration of the size of β that we can tolerate.
- E. Collection of data and observed value of statistic: estimate mean and variance from the data for the statistic.
- F. Decision: If the sample value falls in rejection region R then we reject H_0 otherwise we accept H_0 .

- i. normal population
- ii. sample that are independent random sample from the population

Example: Let us assume that the density function corresponding to the population is normal with variance equal to 270 and that we will select a random sample of size n from the population. And also assume that mean, μ equals either 100 or 107. Make a decision about the information the calculated mean, $\bar{x} = 106$ when $\alpha = .05$ and $n=30$.

Answer: Let us proceed through the steps as follows:

- A. Computing hypothesis: we have null hypothesis $H_0 : \mu=100$ against
alternative hypothesis $H_A : \mu =107$
- B. Assumptions: We assume that we have-
i) a random sample of size n from
ii) a population which is normal ie. $N(\mu, 270)$
- C. Statistic and Sampling distribution: Under H_0 we have $\mu=100$, Since the population is $N(\mu, 270)$ with mean, $\mu=100$, we know that $\bar{x} \sim N(100, \frac{270}{n})$
- D. Selection of α , n and decision rule: Assume that we have $\alpha =.05$ and $n = 30$ then if $\bar{x} > 104.95$, reject H_0 : other wise accept H_0

Now, $\text{Prob} [\bar{x} > 104.95 | \bar{x} \sim N(100, \frac{270}{30})] = .05$
- E. Collection of data and calculate of statistic: we have, the calculated mean, $\bar{x} = 106$.
- F. Decision: Since $\bar{x} = 106$ we have $\bar{x} > 104.95$ and $\bar{x} < R$ Hence, reject the null hypothesis. The rejection region R is shown in figure 8.5

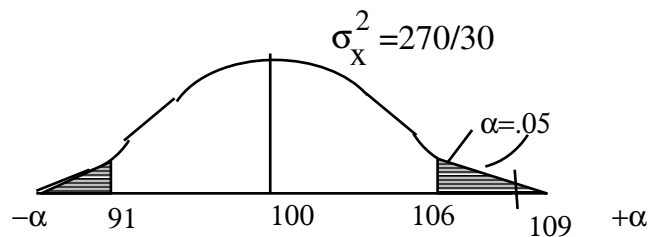


Fig. 10.5: Sampling distribution of $\bar{x}$ under H_0 .

One Tailed and Two Tailed Test:

There are two types of test of hypothesis:

- i. One tailed - test.
- ii. Two tailed- test.

- (i) **One Tailed Test :** When the hypothesis about population mean is rejected only for the value of estimated population mean falling into one of the tails of the sampling distribution, then it is known as one tailed test. If the tail present right side called right tailed test (see fig. 10.6) and if the tailed present left side called left tailed test (see fig. 10.7)

When the hypothesis about population mean is rejected only for the value of estimated population mean falling into one of the tails of the sampling distribution.

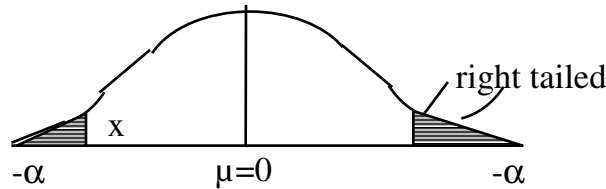


Fig. 10.6: right tailed test

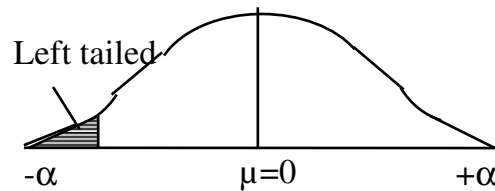


Fig. 10.7: left tailed test.

- (ii) **Two Tailed Test:** When the hypothesis about population mean is rejected for the value of estimated population mean falling into either of the tail of right side or of the tail of left side test of the sampling distribution i.e. when the hypothesis are specific either population mean is greater then hypothesis mean or population mean is less than hypothesized mean, then the test is called two-tailed test.(see fig. 10.8)

When the hypothesis about population mean is rejected for the value of estimated population mean falling into either of the tail of right side or of the tail of left side test of the sampling distribution

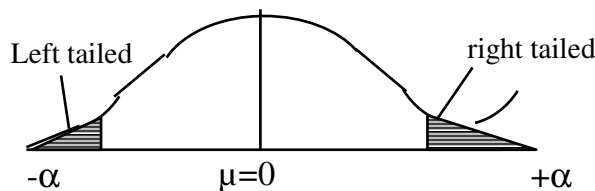
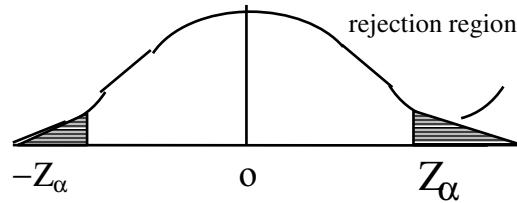


Fig.10.8: Two Tailed Test.

Procedure: procedure for performing a hypothesis test:

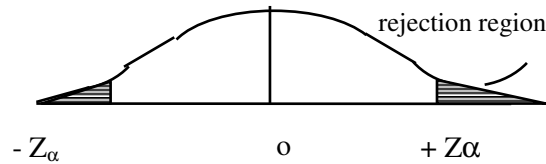
1. Null hypothesis $H_0 : \mu = \mu_0$, against
alternative hypothesis $H_A : \mu > \mu_0$ or $\mu < \mu_0$
2. Assumption: sample size is at least 30 ($n > 30$)
3. Significance level $\alpha = .05$ or $.01$ or $.10$



4. The critical value 'C' can be obtained -

$C = \pm z_{\alpha}$; when $H_A : \mu > \mu_0$ then $c = + z_{\alpha}$

$H_A : \mu < \mu_0$ then $c = -z_{\alpha}$



5. Compute the value $z = \pm \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$; μ_0 is the value for μ given in the null hypothesis

6. If $z > c$ (critical value), reject the null hypothesis.

Example: A tire dealer decides that he will test 40(10 scats) of the new brand of tire. He want to perform the hypothesis test:

$H_0 : \mu = 35,000$ mi

$H_A : \mu > 35,000$ mi. or

or, $H_A : \mu < 35,000$ mi.

Where μ is the mean tire life of the new brand tire. After test of the sample, he shows $\bar{x} = 36720.36$ mi. With standard deviation of the life of tire $s = 2390$ mi. Then standard deviation of the life of tire $s = 2390$ mi. Justify the dealer opinion.

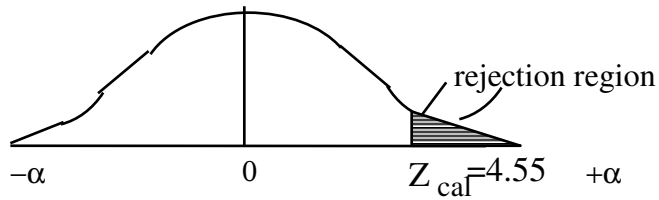
Answer: There are two case:

A. When $H_A : \mu > 35,000$ mi. (Right hand tailed test)

B. $H_A : \mu < 35,000$ mi. (Left hand tailed test)

A. 1. For $H_A : \mu > 35,000$ mi.

$$2. \quad z = \frac{36720.36 - 35000}{2390/\sqrt{40}} = 4.55;$$



Where, $\bar{x} = 36720.40$

$$\mu_0 = 35000$$

$$\sigma = 2390$$

$$n = 40$$

3. The critical region $C > Z$; the dealer reject the null hypothesis and will purchase the new tire.

B. 1. For

$$H_0 : \mu = 35000 \text{ mil.}$$

$$H_A : \mu < 35000 \text{ mile.}$$

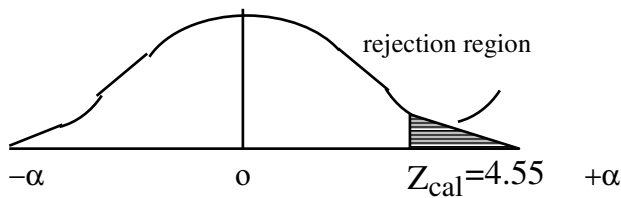
$$2. z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{36720.36 - 35000}{2390 / \sqrt{40}} = 4.55$$

Where, $\bar{x} = 3720.36$

$$\mu_0 = 35000$$

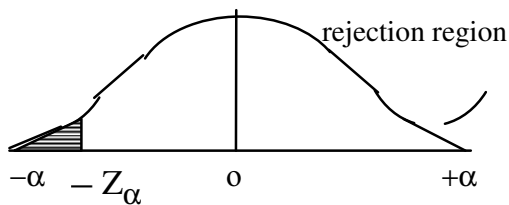
$$s = 2390$$

$$n = 40$$



3. The critical value

$$C = -z_{\alpha};$$



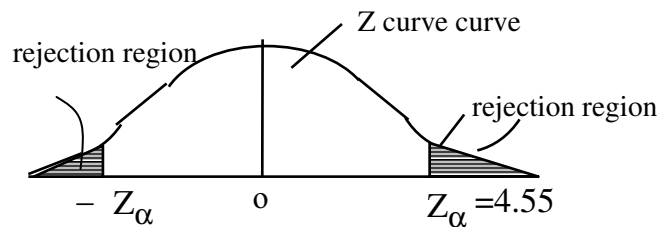
4. Decision: The critical region $C > z$, the dealer reject the null hypothesis i.e. the dealer purchase the new brand tire.

Two tailed test: When the hypothesis is not specified i.e. when $H : \mu = \mu_0$ against

$H_A : \mu \neq \mu_0$ - then

The test procedure is called two tailed test procedure:

1. The hypothesis:
null hypothesis $H_0 : \mu = \mu_0$ against
alternative hypothesis $H_A : \mu \neq \mu_0$
- ii. Assumption: Sample size is at least 30 ($n > 30$)
- iii. Significance level: $\alpha = .05, .01$ or $.10$
- iv. The critical region : $C = \pm z_{\alpha}$



- v. If $P[z < -c \text{ or } z > c] = \alpha$ then reject the null hypothesis.

Example: Let the actual mean yield of wheat denoted as μ , in bushels per acre, resulting from using the new fertilizer. The farmer wants to test the hypothesis.

$H_0 : \mu = 40$ bushels / acre

$H_A : \mu \neq 40$ bushels / acre

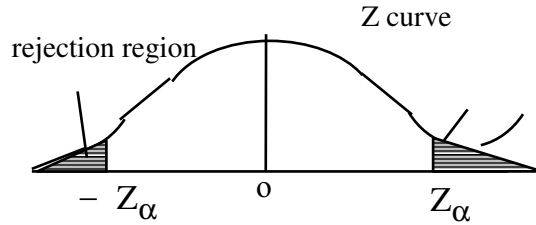
The sample mean $\bar{x} = 38.2$ and $\sigma = 5.3$ and $n = 100$ then what will be the decision about production.

Answer: The test procedure is given below:

1. The hypothesis—
 $H_0 : \mu = 40$
 $H_A : \mu \neq 40$
2. Assumption: Sample size is at least 30 ie. $n > 30$ here $n = 100$
3. The significance level $\alpha = .05$
4. The critical values—
 $-C = -z_{.05}$ and $c = z_{.05}$

$$\text{Now, } z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{38.2 - 40}{5.3 / \sqrt{100}} = -3.40$$

Since $z = -3.40$



We see that $z < -1.96$ and consequently we reject the null hypothesis i.e. the new fertilizer does not give the same mean.

Self-Assessment Questions:

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) "Reject the null hypothesis when it is accepted". this type of error is called:
 - (a) Type II error
 - (b) Type III error
 - (c) Type I error
 - (d) Type IV error
- (ii) "Accept the null hypothesis when it is rejected" this type of error is called:
 - (a) Type I error
 - (b) Type II error
 - (c) Type III error
 - (d) Type IV error
- (iii) "Prob {Reject H_0 | H_0 is true}" is called
 - (a) power of test
 - (b) level of significance
 - (c) null hypothesis
 - (d) alternative hypothesis
- (iv) "Prob {Accept H_0 | H_A is true}" is called
 - (a) α
 - (b) β
 - (c) Type I error
 - (d) Type II error
- (v) Reject H_0 when it is called
 - (a) Correct decision
 - (b) Types I error
 - (c) Types II error
 - (d) Type III error

Answer

Multiple Choice Questions

(i)- c (ii)- b (iii)- b (iv)- a (v)- a

Lesson 3: Testing Hypothesis about Population Mean and Difference between Two Population Mean

Lesson Objectives

After studying this lesson, you will be able to explain

- Testing hypothesis about population mean;
- Testing hypothesis about difference between two population means;
- Testing hypothesis with some examples.

Introduction

Test of means plays an important role in hypothesis testing. The test of hypothesis about the mean of a single population is discussed in previous lesson. Since we need at least one comparison or control group in the case of testing hypothesis about difference between two population means. If we test hypothesis about the mean of two population. One of which is a comparison or control group.

There are two types of test of hypothesis about mean:

- A. Large sample size test of population mean
- B. Small sample test of population mean.

A. Testing Hypothesis about Mean (Small sample). Assume that we wish to test a hypothesis about the mean of a normal population at the level of significant α . we then select a random sample of size $n < 30$ from the population. we have

$$x_i \sim N(\mu, \sigma^2); i = 1, 2, \dots, n$$

where, the random variables are independent.

When σ is unknown, we examine :

1. the computing hypothesis
2. assumption
3. the test statistic
4. the best test
5. an example.

1. The computing hypothesis

- a. $H_0 : \mu = k$ against
 $H_A : \mu \neq k$, i. e.
- b. $H_0 : \mu = k$ against
 $H_A : \mu < k$

We wish to test a hypothesis about the mean of a normal population at the level of significant α we then select a random sample of size $n < 30$ from the population. we have

$$x_i \sim N(\mu, \sigma^2); i = 1, 2, \dots, n$$

- c. $H_0 : \mu = k$ against
 $H_A : \mu > k$

where k is any real number.

2. **Assumptions:**

We have a random sample of size n from a normal population with:

a. mean $= \mu$; unknown

b. variance $= \sigma^2$; unknown i.e.

$x_i \sim N(\mu, \sigma^2)$; $i = 1, 2, \dots, n$

and $\sigma^2 > 0$

3. **Test statistic:** Under the null hypothesis the test statistics—

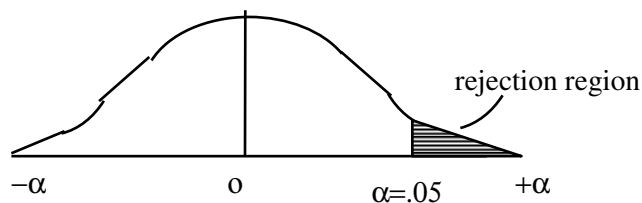
$$t = \frac{\bar{x} - k}{s/\sqrt{n}} \sim t_{n-1} \text{ distribution with } (n-1) \text{ degrees of freedom.}$$

where s is standard deviation (computed)

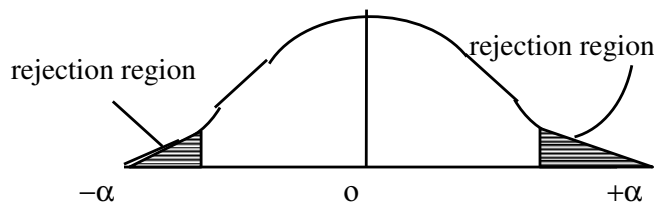
4. **Best test corresponding to types of computing hypothesis:**

- a. Reject null hypothesis, H_0 if $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} > t_{n-1, \alpha}$, t distribution with $(n-1)$ degrees of freedom at $\alpha\%$ level of significance otherwise accept the null hypothesis

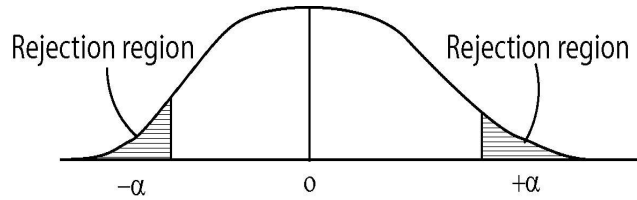
Reject null hypothesis, H_0 if $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} > t_{n-1, \alpha}$ otherwise accept the null hypothesis



- b. Reject the null hypothesis if $t = \frac{\bar{x} - k}{s/\sqrt{n}} > t_{n-1, \alpha}$, otherwise accept the null hypothesis.



- c. Reject the null hypothesis if $\frac{\bar{x}-k}{s/\sqrt{n}} > t_{n-1, \alpha/2}$, otherwise the null hypothesis is accepted.



5. **Example:**

A beer manufacturer puts out a 16-oz can of beer. He wants to make sure that the machine being used to fill the cans is working properly i.e. he wants to see whether the mean volume, μ , of beer put into the can is 16 fluid oz.

To check this he takes a sample of 20 cans of beer and determines the volume of each. 20 cans of beer, mean $\bar{x} = 16.02$ and standard deviation $s = 0.18$ and $\alpha = .10$

Test the following hypotheses:

1. $H_0 : \mu = 16$ against $H_A : \mu > 16$
2. $H_0 : \mu = 16$ against $H_A : \mu < 16$
3. $H_0 : \mu = 16$ against $H_A : \mu \neq 16$

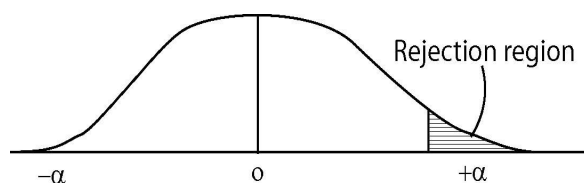
Answer: i) The significance level $\alpha = .10$

ii) The test statistic —

$$\begin{aligned} t &= \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t\text{-distribution with } n-1 \text{ d.f.} \\ &= \frac{16.02 - 16}{0.18/\sqrt{20}} \\ &= .50 \\ \therefore t &= .50 \end{aligned}$$

iii) **Decision:**

Reject the null hypothesis if $t_{cal} > t_{n-1, \alpha}$



iv. If the tabulated $t_{19;.10} = 1.75$ is less than of $t_{cal} = .05$ the null hypothesis is rejected. Since $t_{cal} < t_{tab}$ i.e. the null hypothesis is accepted i.e. the manufacturer will assume that the can machine is working properly.

2. i) The test statistics under the null hypothesis $t_{cal} = -\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ distribution

$$\therefore t_{cal} = -\left[\frac{16.02-16}{.18/\sqrt{20}} \right]$$

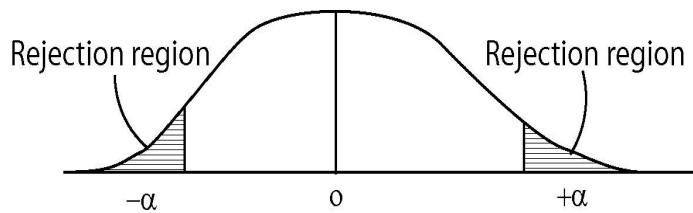
$$= -.50$$

$$\therefore t_{cal} = -.50$$

ii) **Decision** : Reject the null hypothesis if $-t_{cal} > t_{20-1, .10}$ otherwise accept the null hypothesis.

Now,

$$t_{cal} = -.50 \text{ and } -t_{19;.10} = -1.73$$



$-t_{19;.10} = -1.73$; reject the null hypothesis.

Now, $t_{cal} < t_{tab}$ i.e. the null hypothesis is rejected i.e. the manufacturer will assume that his cans filling machine is working properly.

3. Under the null hypothesis—

i) the test statistic –

$$t_{cal} = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1; \alpha/2} \text{ distribution,}$$

$$\text{Now, } t_{cal} = \left| \frac{16.02-16}{.18/\sqrt{20}} \right|$$

$$= .50$$

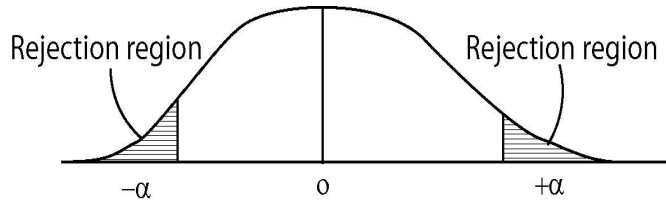
2. **Decision:**

Reject the null hypothesis if $t_{cal} < t_{n-1; \alpha/2}$ otherwise null hypothesis is accepted.

$$\text{Here, } t_{20-1, .10/2} = t_{19,.05} = 1.73$$

$$-t_{19,.05} = -1.73$$

$$\text{and } t_{cal} = .50$$



Since $t_{cal} < t_{tab}$; we reject the null hypothesis i.e. the manufacturer will assume that his can filling machine is working properly.

Testing Hypothesis about Mean (Large sample) when sample size $n > 30$ in a hypothesis test then the sample are called large sample test procedure:

i. The hypotheses are as follows:

a. $H_0 : \mu = \mu_0$ against

$H_A : \mu > \mu_0$

b. $H_0 : \mu = \mu_0$ against

$H_A : \mu < \mu_0$

c. $H_0 : \mu = \mu_0$ against

$H_A : \mu \neq \mu_0$

2. Assumptions:

i. Samples are taken from normal distribution

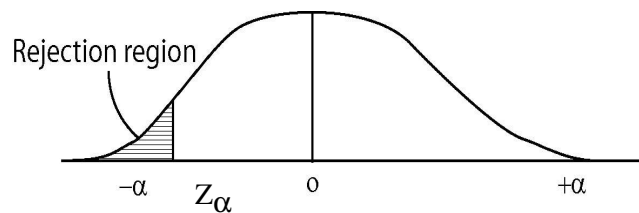
ii. $n > 30$

3. **Test statistics:** Under the null hypothesis the test statistics

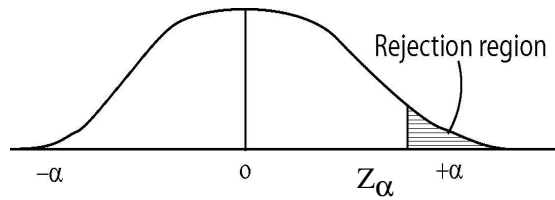
$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} ; \text{ Normal distribution with } n \text{ degrees of freedom.}$$

4. Decision:

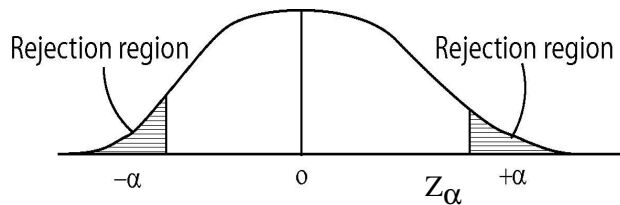
a. Reject null hypothesis if $Z_{cal} > Z_{\alpha}$ otherwise accept the null hypothesis.



- b. Reject null hypothesis if $Z_{cal} > -Z_{\alpha}$ otherwise accept the null hypothesis



- c. Reject null hypothesis if $Z_{cal} > Z_{tab}$ otherwise accept null hypothesis.



5. Example: A soft-drink manufacturer seals a “one liter” bottle of soda. The Food Development authority (FDA) is concerned that the manufacturer may be short changing the customer. The FDA is concerned that mean μ is less than one liter and therefore decides to perform the hypothesis test:

- Ho: $\mu = 1$ liter against
HA: $\mu > 1$ liter
- Ho: $\mu = 1$ liter
HA: $\mu < 1$ liter
- Ho: $\mu = 1$ liter
HA : $\mu \neq 1$ liter at 5% level of significance the mean of the sample
 $\bar{x} = .997$ liter
 $s = .021$ liters and
 $n = 100$

Answer:

- Since the sample size $n=100$ ($n>30$) so the sample is large sample.
- Samples are drawn randomly
- Under null hypothesis, the test statistic

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim \text{normal distribution with } n \text{ degrees of freedom.}$$

$$= \frac{.997 - 1.00}{.021/\sqrt{100}} = -1.43$$

$$\therefore Z = -1.43$$

4. a. Reject the null hypothesis if $Z_{cal} > Z_{n;\alpha}$ otherwise accepted.

Since $Z_{n;\alpha} = Z_{.05} = 1.64$ Now

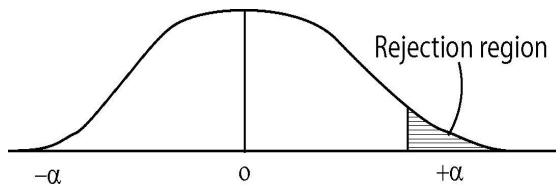


fig.

$Z_{cal} < Z_{tab}$ so the null hypothesis is accept. i.e. the FDA can not reject the null hypothesis at 5% level of significance that μ is one liter.

- b. Reject the null hypothesis if $-Z_{cal} > -Z_{n;\alpha}$ otherwise the null hypothesis is accepted

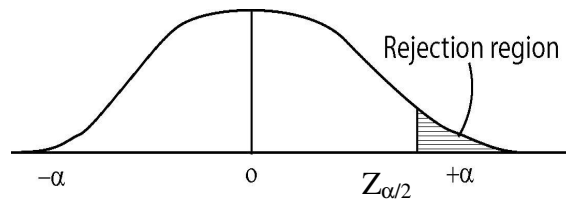
Here, $Z_{.05} = -1.64$ and $Z_{cal} = -1.43$ i.e.

$Z_{cal} < Z_{tab}$ i.e. the null hypothesis is accepted.

i.e. the FDA can not reject the null hypothesis.

- c. Reject the null hypothesis if $Z_{cal} > Z_{n;\alpha}$ otherwise accept the null hypothesis.

Hare $Z_{cal} = 1.43$ $Z_{.05} = 1.64$



Now, $Z_{cal} < Z_{tab}$ so the null hypothesis is accepted i. e. the FDA can not reject the null hypothesis at 5% level of significance that μ is one liter.

Testing Hypothesis about the Difference between Two Means:

Now we wish to test hypothesis about the difference between the means of two population (small and large sample size). Therefore, we select two independent random sample. One random sample of size n_1 is selected from the first population, and a second independent random sample of size n_2 is selected from the second population. Again we may distinguish between known and unknown variance for the population. For small sample size:

1. **The computing hypothesis:**

- a. $H_0 : \mu_1 - \mu_2 = k$ against
 $H_A : \mu_1 - \mu_2 < k$
- b. $H_0 : \mu_1 - \mu_2 = k$ against
 $H_A : \mu_1 - \mu_2 > k$
- c. $H_0 : \mu_1 - \mu_2 = k$ against
 $H_A : \mu_1 - \mu_2 \neq k$

2. **Assumption:**

We have two independent random samples of size n_1 and n_2 from two normal population with unknown means and unknown but equal variances,

$$\text{i.e. } x_{1i} \sim N(\mu_1; \sigma^2); i = 1, 2, \dots, n_1$$

$$x_{2j} \sim N(\mu_2; \sigma^2) j = 1, 2, \dots, n_2$$

Test statistics:

$$t_{\text{cal}} = \frac{(\bar{x}_1 - \bar{x}_2) - k}{S \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t\text{-distribution with } n_1 + n_2 - 2 \text{ degrees of freedom.}$$

$$\text{Where, } S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$S_1^2 = \sum_{i=1}^{n_1} \frac{(x_{1i} - \bar{x}_1)^2}{n_1 - 1}$$

$$S_2^2 = \sum_{i=1}^{n_2} \frac{(x_{2i} - \bar{x}_2)^2}{n_2 - 1}$$

4. **Best test:**

- a. Reject H_0 if $t_{\text{cal}} < -t_{(n_1+n_2-2)\alpha}$; otherwise accept the null hypothesis.
- b. Reject the H_0 if $t_{\text{cal}} > t_{(n_1+n_2-2)\alpha}$; otherwise accept the null hypothesis.
- c. Reject the null hypothesis H_0 if $t_{\text{cal}} > t_{(n_1+n_2-2)\alpha/2}$; otherwise accept the null hypothesis.

5. Example: A manufacturer of automobile products is interested in comparing a newly developed bulb with the bulb he is prudently producing. Specifically he wishes to perform the hypothesis.

- a. $H_0 : \mu_1 = \mu_2$ against
 $H_A : \mu_1 < \mu_2$

b. $H_0: \mu_1 = \mu_2$ against

$H_A: \mu_1 > \mu_2$

c. $H_0: \mu_1 = \mu_2$ against

$H_A: \mu_1 \neq \mu_2$

where μ_1 = the mean effectiveness time for the bulb presently.

μ_2 = the mean effectiveness time for the new developed bulb

The result of the experiment:

Time of effectiveness for the bulb presently	Time effectiveness for newly developed bulb
89	94
86	91
83	88
87	92
89	87

Answer:

1. Computing mean and variance:

$$x_1 = 86.80 \text{ and } x_2 = 90.40$$

$$s_1 = 2.49 \quad s_2 = 2.88$$

$$n_1 = 5 \quad n_2 = 5$$

$$\therefore S = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(5-1)(2.49)^2 + (5-1)(2.88)^2}{(5+5-2)}}$$

$$\therefore S = 2.69 \quad = 2.69$$

2. The competing hypothesis:

a. $H_0: \mu_1 = \mu_2$ against $H_A: \mu_1 > \mu_2$

b. $H_0: \mu_1 = \mu_2$ against $H_A: \mu_1 < \mu_2$

c. $H_0: \mu_1 = \mu_2$ against $H_A: \mu_1 \neq \mu_2$

3. Test statistics:

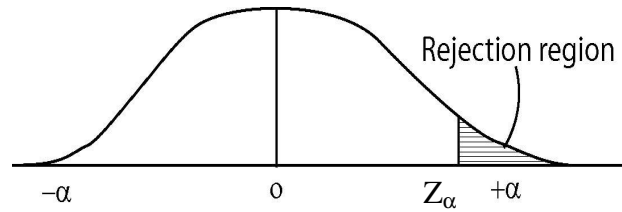
$$t_{\text{cal}} = \frac{\bar{x}_1 - \bar{x}_2}{S \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \text{distribution with } n_1 + n_2 - 2 \text{ degrees of freedom.}$$

$$= \frac{86.80 - 90.40}{2.69 \sqrt{\left(\frac{1}{5} + \frac{1}{5}\right)}} = -2.12$$

4. a. Reject the null hypothesis if $t_{cal} > t_{(n1+n2-2), \alpha}$

Now $t_{cal} = 2.12$ and

$$t_{tab} = t_{5+5-2, .05} = .86$$



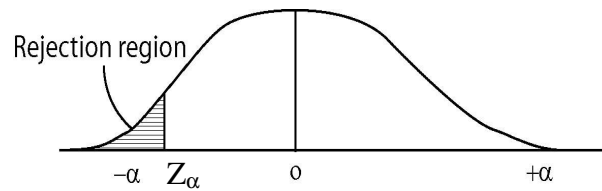
Since, $t_{cal} > t_{tab}$ so the null hypothesis is rejected.

i.e. the newly developed bulb has a longer mean effectiveness time than the present bulb.

- b. Rejected null hypothesis if $t_{cal} < -t_{n1+n2-2, \alpha}$

Here, $t_{cal} = -2.12$ and

$$t_{tab} = t_{5+5-2, .05} = t_{8, .05} = -1.86$$



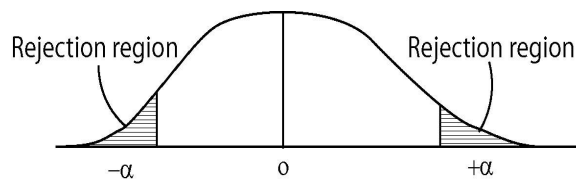
Since, $t_{cal} < -t_{tab}$ so the null hypothesis is rejected.

i.e. the newly developed bulb has a longer mean effective time than the present bulb.

- c. Reject the null hypothesis if $t_{cal} > t_{n1+n2-2; \alpha/2}$

Here, $t_{cal} = |-2.12| = 2.12$ and

$$t_{n1+n2-2, \alpha/2} = t_{8; .02} = 2.306$$



Since $t_{cal} < t_{tab}$ so the null hypothesis is accepted.

For Large Sample:

Let $\bar{x}_1$ be the sample mean of a random sample of size n_1 from a population with mean μ_1 and $\bar{x}_2$ be the sample mean of a random sample of size n_2 from another population with mean μ_2 . Also assume that the two samples are independent. If $(n_1, n_2) > 30$ i.e. for large sample, the test statistic will be-

$$Z_{\text{cal}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} \sim z\text{-distribution with } (n_1 + n_2 - 2) \text{ d-f}$$

where

$\bar{x}_1$ = mean of first sample

$\bar{x}_2$ = mean of second sample

s_1 = standard deviation of
the first sample

s_2 = standard deviation of the second
sample

n_1 = first sample size

n_2 = second sample size

1. Computing hypothesis:

- $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 > \mu_2$
- $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 < \mu_2$
- $H_0 : \mu_1 = \mu_2$ against $H_A : \mu_1 \neq \mu_2$

2. Test Statistics:

$$Z_{\text{cal}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)}} \sim Z\text{distribution}$$

3. Assumptions:

$H_0 : \mu_1 = \mu_2$ against

$H_A : \mu_1 > \mu_2$,

$H_0 : \mu_1 = \mu_2$ Against and

$H_A : \mu_1 < \mu_2$

$H_0 : \mu_1 = \mu_2$ against

$H_A : \mu_1 \neq \mu_2$

4. Best test:

- Reject null hypothesis if $z_{\text{cal}} > z_{\text{tab}}$; otherwise accept the null hypothesis.
- Reject the null hypothesis if $z_{\text{cal}} < -z_{\text{tab}}$; otherwise accept the null hypothesis.
- Reject the null hypothesis if $z_{\text{cal}} > z_{\text{tab}}$; otherwise accept the null hypothesis.

5. Example: A scientist is working for a company that produces bulbs. He has developed new types of filament that will prolong the life of the light bulbs. Then the statistician decides to take random sample of 100 light bulb using the presently used filament, and 100 light bulbs equipped with the new type of filament. The results are as follow:

$$\begin{array}{ll} \bar{x}_1 = 1214.52 & \bar{x}_2 = 1260.09 \\ s_1 = 119.86 & s_2 = 123.57 \\ n_1 = 100 & n_2 = 100 \end{array}$$

find there is any evidence of the experiment.

Answer:

1. Let us consider the null hypothesis
 - a. $H_0 : \mu_1 = \mu_2$ again
 $H_A : \mu_1 > \mu_2$
 - b. $H_0 : \mu_1 = \mu_2$ again
 $H_A : \mu_1 < \mu_2$
 - c. $H_0 : \mu_1 = \mu_2$ against
 $H_A : \mu_1 \neq \mu_2$
2. The level of significant $\alpha = .01$
3. **Assumption:** 1. randomly distributed from the population
2. sample are independent.

4. Test statistics:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} \sim t \text{ distribution with } n_1 + n_2 - 2 \text{ degrees of freedom}$$

$$= \frac{1214.52 - 1260.09}{\sqrt{\left[\frac{(119.86)^2}{100}\right] + \left[\frac{(123.57)^2}{100}\right]}}$$

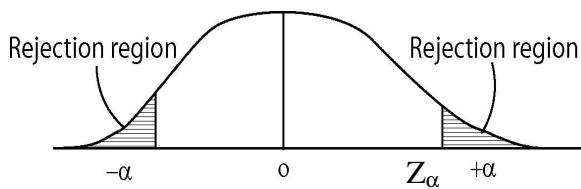
$$= -2.56$$

$$\therefore z = -2.56$$

5. a. Reject the null hypothesis if $z_{cal} > z_{.01}$ otherwise accept the null hypothesis.

$$\text{Here } z_{cal} = 2.56$$

$$z_{.01} = 2.33$$

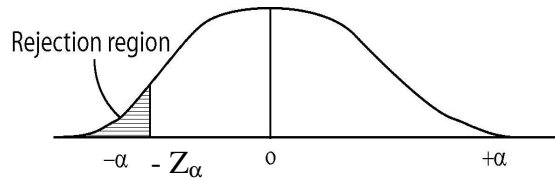


Since $z_{cal} > z_{tab}$; so the null hypothesis is accepted i.e. the new filament improved the life of the light bulb.

- b. Reject the null hypothesis if $z_{cal} < -z_{\alpha}$; otherwise the null hypothesis is accepted.

Here- $z_{cal} = -2.56$

$$z_{n_1+n_2-2; \alpha} = z_{198; .01} = -2.33$$



Since $z_{cal} < -z_{tab}$; So the null hypothesis is accepted i.e. the new filament improve the life of the light bulb.

- c. Reject the null hypothesis if $z_{cal} > z_{\alpha}$ otherwise accept the null hypothesis.

Here $z_{cal} = 2.56$

$$z_{tab} = z_{.01} = z_{.01} = 2.33$$

Since $z_{cal} > z_{tab}$; so the null hypothesis is accepted i.e. the new filament improve the life of light-bulb.

Self-Assessment Questions:

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) When the sample size "n" is greater than 30, i.e. $n > 30$ then the sample is called.
 - (a) small sample
 - (b) large sample
 - (c) random sample
 - (d) systematic sample
- (ii) When the sample size "n" is less than 30, then the sample is called.
 - (a) systematic sample
 - (b) random sample
 - (c) small sample
 - (d) large sample.
- (iii) For large sample, the test statistics followed
 - (a) t–statistic
 - (b) z–statistic
 - (c) χ^2 – statistic
 - (d) F–statistic
- (iv) For small sample, the test statistics followed
 - (a) F–statistic
 - (b) χ^2 –statistic
 - (c) t–statistic
 - (d) z–statistic

Answer:

Multiple Choice Questions

(i)- b (ii)- c (iii)- b (iv)- c

Lesson 4: Testing Hypothesis about Population Proportion and Difference between Two Population Proportion

Lesson Objectives:

After studying this lesson, you will be able to:

- Testing the hypothesis about population proportion;
- Comprehend testing the hypothesis between two difference population proportion;
- Explain with example of above cases.

Introduction

If the population represent by a binomial probability function with $n=1$ and parameter p equal to the proportion of individuals in the population. For bernoulli trail, we have sampling distribution of estimated p using central limit theorem for large sample being approximately normal.

Test of Hypothesis about Population Proportion:

The value of p , the proportion of individual in the sample of a population has to be considered when the normal curve approximation for the binomial. Therefore if np or nq is less than 5, use exact tables if np or nq is greater than or equal to 5, use normal approximation. If the sample size is smaller than 100(say), include the correction for continuity. Let us consider a test about one sample population test.

1. Types of computing hypothesis:

- a. $H_0 : P = P_0$ against
 $H_A: P > P_0$
- b. $H_0: P=P_0$ against
 $H_A: P < P_0$
- c. $H_0: P=P_0$ against
 $H_A: P \neq P_0$

2. **Assumption:** We have bernoulli trails from the population of interest.

3. **Test statistics:** Under the assumption and null hypothesis, we have

- a. For small sample: $t(p) = \frac{n!}{x!(n-x)!} p^x q^{n-x}; x=0,1$
Or otherwise.

- b. For large sample: If P_0 is the population proportion and P is the sample proportion, based on a sample size n , then the random variable:

For small sample:

$$t(p) = \frac{n!}{x!(n-x)!} p^x q^{n-x}; n=0,1$$

$$z = \frac{P - p_0}{\sqrt{p_0(1-p_0)/n}}$$

$\sim N(0,1)$ i.e. normal distribution with $(n-1)$ degrees of freedom.

$z = \frac{P - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$ i.e. normal distribution with $(n-1)$ degrees of freedom.

The approximation is good if both np_0 and $n(1-p_0)$ are at least 5.

4. Best test for large sample:

- Reject null hypothesis if $z_{cal} < -z_{\alpha}$; otherwise accept the null hypothesis.
- Reject null hypothesis if $z_{cal} > z_{\alpha}$; otherwise accept the null hypothesis.
- Reject the null hypothesis if $|z_{cal}| > z_{\alpha/2}$; otherwise accept the null hypothesis.

5. Example: A political incumbent received 58% of the vote during the last election. He feels that his 5 years in office have been good ones, and believes that his popularity has increased. To obtain relevant information concerning his belief, he decides to take a random sample of 300 voter to test.

- Ho: $P = .58$ against
H_A: $P > .58$
- Ho : $P = .58$ against
H_A: $P < .58$
- Ho: $P = .58$
H_A: $P \neq .58$, the result of the pollition 179 said they intended to vote for him.

Answer: Here, the hypotheses are given as:

- Ho: $P = p_0$ against
H_A: $P > p_0$
 - Ho: $P = p_0$ against and
H_A: $P < p_0$
 - Ho: $P = p_0$ against
H_A: $P \neq p_0$

2. Test statistics: $z = \frac{P - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0,1)$

Now, $P = \frac{179}{300} = .597$; $p_0 = .58$

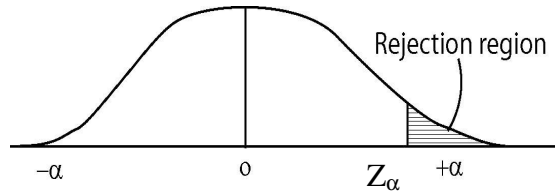
$$\therefore z = \frac{.597 - .58}{\sqrt{.58(1-.58)/300}} = .60$$

3. Best test:

- a. Reject the null hypothesis if $z_{cal} > z_{\alpha}$; otherwise accept the null hypothesis.

Here, $\alpha = .01$

$$z_{cal} = .60 < 1.28; \text{ tabulated value}$$



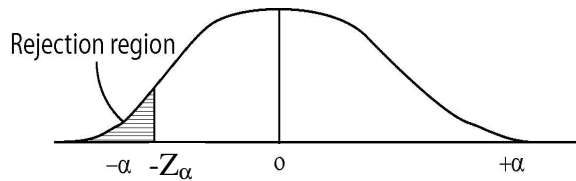
Z_{α}

Since $z_{cal} < z_{tab}$; So the null hypothesis is rejected i.e. his popularity is increased.

- B. Reject the null hypothesis if $z_{cal} < -z_{\alpha}$; otherwise accept the null hypothesis.

Here $z_{cal} = .60$

$$-z_{\alpha} = z_{.01} = -1.28; \text{ tabulated value}$$

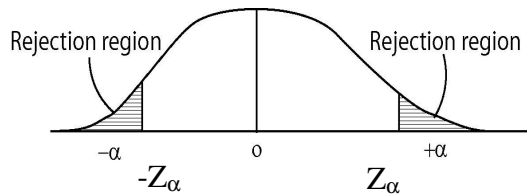


Since $z_{cal} < -z_{tab}$ so the null hypothesis popularity is increased.

- c. Reject the null hypothesis if $z_{cal} > z_{(n-1); \alpha / 2}$; otherwise accept the null hypothesis.

Here $z_{cal} = .60$

$$z_{\alpha/2} = z_{.05} \text{ Tabulated value}$$



Since $z_{cal} < z_{tab}$; so the null hypothesis is rejected i.e. the political popularity is increased.

Testing Hypothesis between Two Different Population Proportion:

Let us consider, we want to test hypothesis about two different proportion P_1 and P_2 , each proportion corresponding to the parameter P of a binomial population. Therefore we select two independent random samples; one sample size n_1 , from the first population and a second sample size n_2 from the second population.

1. Types of computing hypothesis:

- a. $H_0: P_1 - P_2 = 0$ against
 $H_A: P_1 - P_2 > 0$
- b. $H_0: P_1 - P_2 = 0$ against and
 $H_A: P_1 - P_2 < 0$
- c. $H_0: P_1 - P_2 = 0$ against
 $H_A: P_1 - P_2 \neq 0$

2. Assumption:

1. Bernoulli trial, with success P_1 and P_2
2. trials are independent.

3. Test Statistics:

Under null hypothesis, we have—

$$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0, 1)$$

4. Best Test:

- a. Reject the null hypothesis H_0 if $z_{cal} > z_{\alpha}$ otherwise accept the null hypothesis H_0 .
- b. Reject the null hypothesis H_0 if $z_{cal} < -z_{\alpha}$ otherwise accept the null hypothesis H_0 .
- c. Reject the null hypothesis H_0 if $z_{cal} > z_{\alpha/2}$ otherwise accept the null hypothesis.

5. Example: School of Business of BOU wishes to determine whether or not there is any relationship between male and female participation in MBA Programme. They select a random sample of 100 male and female from different R.R.C (Scored 1 if they are male and 0 otherwise). Is there any relationship? given $\alpha = .05$, $P_1 = 0.7$, $P_2 = .08$; $n_1 = 40$ and $n_2 = 60$.

Answer:

1. Computing null hypothesis:

- a. $H_0: P_1 - P_2 = 0$ against
 $H_A: P_1 - P_2 > 0$

b. $H_0: P_1 - P_2 = 0$ against

$H_A: P_1 - P_2 < 0$ and

c. $H_0: P_1 - P_2 = 0$ against

$H_A: P_1 - P_2 \neq 0$

2. Test statistics: Under null hypothesis—

$$Z = \frac{(P_1 - P_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim \text{Normal distribution}$$

$$= \frac{.70 - .80}{\sqrt{\frac{(.70 \times .30)}{4} + \frac{(.80 \times .20)}{6}}} ; q = 1 - p$$

$$= 1.10$$

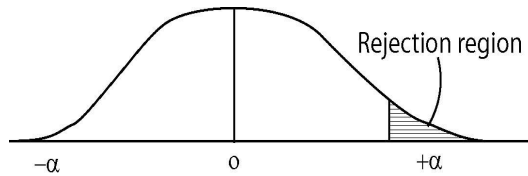
$$\therefore z_{cal} = 1.10$$

3. Test:

a. Reject the null hypothesis if $z_{cal} > z_{\alpha}$ otherwise accept the null hypothesis.

Here $z_{cal} = 1.10$

$$z_{tab} = z_{\alpha} = z_{.05} = 1.96$$

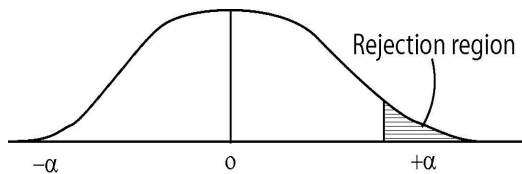


Since $z_{cal} < z_{tab}$; so the null hypothesis is accepted.

b. Reject the null hypothesis if $z_{cal} > z_{\alpha}$; otherwise accept the null hypothesis.

Here $z_{cal} = -1.10$ (say)

$$z_{tab} = z_{\alpha} = z_{.05} = 1.96$$

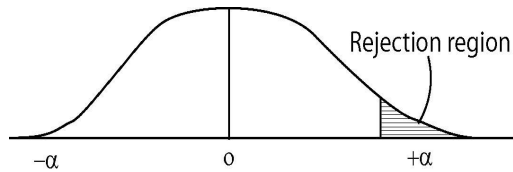


Since $z_{cal} > z_{tab}$ so the null hypothesis is accepted.

- c. Reject the null hypothesis if $t_{cal} > z_{\alpha/2}$ otherwise accept the null hypothesis

Here $z_{cal} = 1.10$

$$z_{tab} = z_{\alpha/2} = z_{0.025} = 1.63$$



Since, $z_{cal} < z_{tab}$ so the null hypothesis is accepted.

Self-Assessment Question:

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Intesting hypothesis about population, if np or nq is less than 5, then use—
(a) exact table (b) table
(c) χ^2 -table (d) t-table
- (ii) Intesting hypothesis about population proportion; if np or nq is greater than 5, then use
(a) χ^2 -table (b) t-table
(c) z-table (d) F-table
- (iii) For large sample p is estimated from sampling distribution by using.
(a) sampling theorem (b) central limit theorem
(c) starting approximation (d) possion therorm
- (iv) The sample proportion is always less than.
(a) ten (b) one
(c) -one (d) -ten.

Answer:

Multiple Choice Questions:

- (i)- c (ii)- c (iii)- b (iv)- b

Lesson 5: Non-Parametric Test

Lesson Objectives:

After completing this lesson, you will be able to:

- Define non-parametric test;
- Know test procedures;
- Solve the problem regarding non-parametric test.

Introduction

A purpose of this lesson is to extend this unit to allow for samples and for populations that may not be normal. In contrast, nonparametric statistical tests do not test hypothesis about specific population parameters. This unit rounds out the overentation of non-parametric techniques by treating non-parametric statistics that the test has difference in the shape or location of the pupulation identifying two or more groups.

Non-Parametric Test:

A statistical test procedures is non-parametric if it satisfies at least one of the following criteria:

- i. The method may be used on data with a nominal scale of measurement.
- ii. The method may be used on data with an ordinal scale of measurement.
- iii. An interval or ratio scale of measurement where the distribution function of the random variable producing the data is either unspecified or specified except for an infinite number of unknown parameter.

There are many test procedures in the non-parametric approaches. We discuss only the following:

- (a) Sign test for paired data
- (b) Rank sum tests:
- (c) Kruskal Wallis test

(a) Sign Test for Paired Date

One of the easiest non-parametric test is known as sign test. The sign test is useful for testing whether one random variable in a pair trends to be longer than the other random variable in a pair. It may be used to test for trend in a series of ordinal measurement or as a test for correlation.

DATA: The data consist of observations on a bivariate random sample $(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)$; where there are N pair of observation.

Within each pair (X_i, Y_i) a comparison in made, and the pair is classified as “+” if $X_i < Y_i$ or “-” if $X_i > Y_i$ and “O” if $X_i = Y_i$. Thus the measurement Scale is ordinal.

Assumption

- (i) The pair of variable (X_i, Y_i) are mutually in dependent.
- (ii) Ordinal measurements scale.
- (iii) The pair (X_i, Y_i) are internally constant.

Hypothesis $H_0: E(X_i) = E(Y_i)$; for all;

$H_A: (X_i) \neq E(Y_i)$; for all;

The Statistics:

T = table number of “+”s.

Decision rule: disregard all ties pair and let n equal the number of pair that are not ties.

N = total number of “+”s and “-”s.

then,

$$t = \frac{1}{2} \left[n + W_{\alpha/2} \sqrt{n} \right]; W_{\alpha/2} \text{ is obtain from table}$$

Reject H_0 if $T \leq t$ or if, $T \geq n - t$ at the level of significance of 2α , otherwise accept H_0 .

Example: Use the sign test to see whether there is a difference between the number of days required to collect an account receivable before and after a new collection policy use the 0.05 significance level.

Before	33	36	41	32	39	47	34	29	32	34	40	42	33	36	29
After	35	29	38	34	37	47	36	32	30	34	41	38	37	35	28

Solution:

Let,

$$H_0: P(+) = P(-)$$

$$H_A: P(+) \neq P(-)$$

no. of “+”s = 7, no. of “-”s = 6 and no. of tie’s = 2

n = no. of “+”s + no. of “-”s

$$= 7+6 = 13$$

T = no. of “+”s = 7

Now, $X_1 = 2266$ (From table)

$$n - t = 9 - 6.20 = 2.78$$

Since $T = 7$, H_0 is accepted because $T > n - t$

Activity:

The following data shows employees' satisfaction level before and after their company was brought by a larger firm. did the buyout increase employee satisfaction? Use the 0.05 significance level.

Before	98.4	96.6	82.4	96.3	75.4	82.6	81.6	91.4	90.4	92.4
After	82.4	95.4	94.2	97.3	77.5	82.5	81.6	84.5	89.4	90.6

Rank Sum Test (Mann-Whitney U Test):

The purpose of the mann-whitney U-test is to help decided whether or not the distributions of scores in two independent group were drawn from two identical population distribution.

Hypothesis: H_0 : The distribution of the scores in two population from which the group were drawn are identical.

$$H_A: \mu_1 \neq \mu_2$$

Test statistic: For $n < 20$

$$U_1 = n_L n_S + \frac{n_L(n_L + 1)}{2} - T_L; U_2 = n_L n_S - U_1$$

Where: n_2 = number of subjects in the group with the larger sum of ranks.

n_3 = number of subjects in the other group

Maun-whitney U = smaller of U_1 and U_2

Decision rules: H_0 is rejected if the probability of observing a value of less than or equal to U. otherwise accept the H_0 .

Test statistics: $n > 20$

$$Z_0 = \frac{U - n_1 n_2}{S_U} = \frac{U - n_1 n_2}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{2}}} \approx Z - \text{distribution}$$

Decision Rule: H_0 is rejected if $Z_U \leq Z_{\alpha/2}$, otherwise accept it.

Example: Using first Graders' data from the study of teacher exceptency are given below:

	C	C	C	E	E	C	E	C	E	E
Raw core	90	99	102	107	111	114	117	121	122	125
Rank	1	2	3	4	5	6	7	8	9	10

Solution:

Here Sum of rank,

Control Group, $C = 1+2+3+6+8 = 20$

Experimental group, $E = 4+5+7+9+10 = 35$

$n < 20$ i.e.,

H_0 : The distribution scores in the two population from which the group were draws are identical.

H_A : $\mu_1 \neq \mu_2$

Test Statistics:

$$U = n_L n_S + \frac{n_L (n_L + 1)}{2} - T_L$$

$$U_2 = n_L n_S - U_1$$

n_L = number of subjects in the group.

n_S = number of subjects in the other groups.

$$U_1 = 5 \times 5 + \frac{5(5+1)}{2} - 35$$

$$= 5$$

$$U_2 = n_L n_S - U_1 = 5 \times 5 - 5 = 20$$

The probability of the observing value is 0.075 (From table)

So, the H_0 is accepted.

Activity:

Test the hypothesis of no difference between the ages of male and female employees of a certain company using the Mann-whitney U test for the sample data. Use the 0.10 level of significance.

Male	31	25	38	33	42	40	44	26	43	35
Female	44	30	34	47	35	32	35	47	48	34

KRUSKAL-WALLIS TEST: The Kruskal-wallis one way analysis of variance by rank is the non-parametric analog to the one way analysis of variance (ANOVA).

Hypothesis:

H_0 : the distribution of scores in the population under each group are identical.

H_A : the population differ in their average.

Test statistics:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n} - 3(N+1); K-1 \text{ degrees of freedom}$$

and where, R_i = Sum of the ranks in group i

K = number of groups

n = number of subjects in a group

N = total sample size.

Decision Rule: Reject H_0 ; if the value of $H_{\text{observed}} \geq X_{\alpha(k-1)}^2$ tabulated value, reject the H_0 . Otherwise accept the null hypothesis.

Example: Students belief scores and ranks in three classes of introductory psychology.

Subjects	Instructor	Rank	Instructor II	Rank	Instructor III	Rank
1	3	4	2	1.5	8	18.5
2	4	6.5	5	9	7	15.5
3	6	12	7	15.5	3	4
4	5	9	6	12	5	9
5	8	18.5	9	20.5	7	15.5
6	9	20.5	3	4	2	1.5
7	6	12	7	15.5	4	6.5
Sum of Rank		82.5		78		70.5

Solution:

Let, H_0 : each three groups are identical

H_A : They are not identical

Test statistic:

$$\begin{aligned}
 H &= \frac{12}{N(N+1)} \sum_{i=1}^3 \frac{R_i^2}{n} - 3(N+1) \\
 &= \frac{12}{12(12+1)} \left[\frac{82.5^2}{7} + \frac{78^2}{7} + \frac{70.5^2}{5} \right] - 3(4+1) \\
 &= 0.27
 \end{aligned}$$

Here table value at $K-1=2$ is 5.99

which is grater H_{observed} . So the null hypothesis is accept.

Tryout: Students belief scores is three classes of introductory psychologies.

Sl.no.	Subject	Instructor		
		I	II	III
1.	1	3	2	8
	2	4	5	7
	3	6	7	3
	4	5	6	5
	5	8	9	2
	6	9	3	2
	7	6	4	7

Test it through Kruskal-Wallis procedure and Comment on it.

Self-Assessment Questions

Write a short note on the following:

1. Sign test
2. Rank Sum test
3. Kruskal-Wallis test

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which test is non-parametric?
- (a) F-test (b) Sign test
- (c) Z-test (d) t-test

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) Mann-Whitney test is parametric test
- (ii) Non-parametric test are always distribution free test
- (iii) Non-parametric may be used on the data with a nominal scales

Answer:

Multiple Choice Questions

(i)- b

True/False

(i) F (ii) T (iii) T

Exercise

1. Explain the general procedure for testing of a hypothesis.
2. What do you mean by null hypothesis and level of significance? Point out the assumption in hypothesis testing in large samples.
3. Distinguish between type I and type II error test of a hypothesis.
4. Differentiate the following:
 - a. Critical region and Acceptance region
 - b. Null hypothesis and Alternative hypothesis
 - c. One tailed test and two tailed test.
5. Discuss the testing hypothesis about population mean for small sample.
6. Discuss the testing hypothesis about population portion.
7. (a) What is meant by test of hypothesis? Define null hypothesis, level of significance, critical region with example.
 (b) The following are the per acre production of rice of some farmers: Production of rice (md/acre)

60.6, 70.2, 50.8, 45.7, 80.2, 60.0, 72.5, 48.7, 60.3, 70.0

Does the production data follow a distribution with mean 65.0?

8. (a) Define type I and type II error.
- (b) Two independent random samples are drawn from two populations. The sample observations are:

x:	10.2	10.6	11.7	15.2	9.6	16.4	8.7	12.3	14.4		
y:	11.6	16.7	17.2	15.0	18.7	14.2	12.7	11.4	9.6	11.0	13.4

Do you think that the populations are same?

9. (a) What is meant by test of hypothesis? What is 't'-test? Write down the uses of t-test.
- (b) From a population a random sample of 12 observations are drawn. The sample observations are:
 10, 12, 11, 9, 15, 17, 10, 10, 14, 16, 18, 12

10. (a) Define test statistic, critical region, level of significance, null hypothesis with examples.
- (b) The following information are given from two samples:
- | Sample 1 | Sample 2 |
|-------------------------------------|---|
| $n_1=32, \Sigma=48, \Sigma x^2=126$ | $n_2=40, \Sigma y = 100, \Sigma y^2=3000$ |
- Do you think that the samples are same?
11. (a) Define parameter, statistics, critical region, level of significance with examples.
- (b) What is the need of test of proportion? In a locality 80% people are involved in farming and in another locality the percentage of people involved in farming is 70%. The number of people in former and latter localities is 500 and 1000 respectively. Are the percentages of people involved in farming in two localities similar?
12. Define non-parametric test. Distinguish between parametric and non-parametric test. State Kruskal Wallis Nonparametric procedure with an example.

CHI-SQUARE (χ^2) TEST



The chi-square test is a very general test that can be used whenever we wish to evaluate whether or not frequencies which have been empirically obtained differ significantly from those which would be expected under a certain set of theoretical assumptions. The test has many applications, the most common of which in the social and business field of studies are “contingency” problems in which two nominal-scale variables have been cross classified.

Some measures of the differences between observed and expected frequencies must be obtained. There are, of course, a large number of possible measures, but we need one for which the sampling distribution is known and tabulated. For this reason, we make use of a measure based on a statistical test, named chi-square.

Sometimes chi-square test is used in analysing the enumeration data. Enumeration data is that data which classifies the experimental units into some distinct groups based on some qualitative characters. For example, a population is sampled and the number of smokers and non-smokers are observed or the numbers of owner, tenant and owner-cum-tenant farmers of the sampled unions of a given district are observed and so on.

Lesson 1: Chi-Square Distribution

Lesson Objectives:

After completing this lesson, you are able to know

- What the chi-square distribution is;
- The expected value and variance of the chi-square distribution;
- The character of the shape of chi-square distribution;
- The use through example.

Introduction

The chi-square statistics provides a measure of how much the observed and expected frequencies differ from one another. But how much differences should be tolerated before concluding that the observed frequencies were not sampled from a distribution represented by the expected frequencies. In other words, how large should chi-square (observed) be in order to reject the null hypothesis that the observed frequencies were sampled from a distribution represent by the expected frequencies?

Chi-Square Distribution

Defination: The chi-square distribution is a theoretical distribution-actually, a family of theoritial distribution. Karl Pearson (1900) first used the word "Chi-quare:

It can be defined as—

$$\chi^2 = c \chi^{(v-2)/2} e^{-\chi^2/2} \chi > 0$$

Where χ is the Greek letter chi and χ^2 read chi-square. C is a constant that norms the distribution so as to produce an area of unity, with C depending on v, the degrees of freedom.

The χ^2 -distribution corresponding to various values of v are shown in Fig: 11.1.

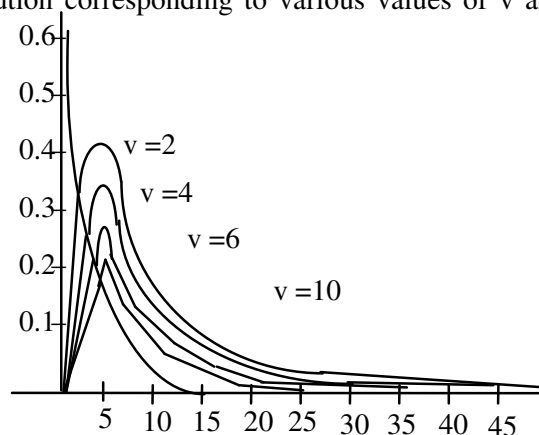


Fig. 11.1: Chi-square distribution for various values of v.

The expected value of χ^2 -distribution is v and its variance is $2v$. It turns out that a chi-square distribution can also be interpreted as a sum of a number of squared and independently distributed standardized normal variables.

The fig: 11.1 represent a family of curves that always vary from zero to infinity and are skewed to the right, with the degrees of skewness diminishing as v increases. The expected value of χ^2 -distribution is v and its variance is $2v$. It turns out that a chi-square distribution can also be interpreted as a sum of a number of squared and independently distributed standardized normal variables.

Properties:

- i. chi-square distribution possess a convenient additive property. If there are two independently distributed chi-square distribution, with v_1 and v_2 degrees of freedom respectively, their sum also be distributed as chi-square distribution with v_1+v_2 degrees of freedom.
- ii. χ^2 - distribution is unimodal
- iii. χ^2 - distribution is positive by skewed.
- iv. χ^2 - distribution approaches normal distribution when degrees of freedom grows infinitely large.
- v. The mean and variance of the χ^2 - distribution is $E(\chi^2) = v$ and $V(\chi^2) = 2v$ respectively where v is the number of degrees of freedom.

Chi-Square Test

Chi-square test defined as a measure of the discrepancy exist between the observed and expected frequencies is supplied by the statistic χ^2 given by

$$\chi^2_{\text{observed}} = \sum \frac{(\text{observed frequencies} - \text{expected frequencies})^2}{\text{expected frequencies}}$$

$$\text{or } \chi^2_{\text{observed}} = \sum \frac{\text{observed frequencies}}{\text{expected frequencies}} - N; N = \text{number of frequencies.}$$

where “ Σ ” defined as summation, use Greek letter sigma.

Suppose that in a particular sample a set of possible events $E_1, E_2, \dots, E_k$ (see table B) are observed to occur with frequencies $O_1, O_2, \dots, O_k$, called observed frequencies, and that according to probability rules they are expected to occur with frequencies $e_1, e_2, \dots, e_k$, called expected or theoretical frequencies.

Table-B

Events	$E_1 \ E_2 \ \dots \ E_k$
Observed frequencies	$O_1 \ O_2 \ \dots \ O_k$
Expected frequencies	$e_1 \ e_2 \ \dots \ e_k$

Now we are interested to know whether the observed frequencies differ significantly from the expected frequencies.

Then the χ^2 -test statistics are given by $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$

$$\text{or } \chi^2 = \sum_{i=1}^k \frac{O_i^2}{E_i} - N; \quad i=1, 2, \dots, K \text{ and } N = \sum_{i=1}^k O_i$$

of $\chi^2 = 0$, the observed and theoretical frequencies agree exactly

$\chi^2 > 0$, they do not agree exactly,

Calculate the Degrees of Freedom (d-f):

The degrees of freedom for chi-square statistic depends on the number of categories (k) and not the number of subjects in the sample (N), i.e. the degrees of freedom $df = \text{number of categories} - 1 = k - 1$.

In order to determine whether to reject the null hypothesis, we compared with χ^2_{observed} with χ^2_{critical} with $k-1$ degrees of freedom.

Where $\chi^2_{\text{observed}} = \text{Chi-square statistic value from observed data}$

$\chi^2_{\text{critical}} = \text{Chi-square statistic value from } \chi^2 \text{ table at different level of significance (see table)}$

Hypothesis: Chi-square is typically used for non directional (two tailed) test. For such test, the null hypothesis and alternative hypothesis are as follows:

H_0 : The observed distributions of frequencies equal the expected distribution of frequencies in each categories.

H_A : The observed distribution of frequencies does not equal the expected distribution of frequencies.

Assumptions and Requirements:

For testing whether the observed frequencies in two or more categories come from the hypothesised frequencies in these categories, we make the following assumptions:

- i. Each observation must fall in one and only one category.
- ii. The observations in the sample are independent of one another.
- iii. The observations are measured as frequencies.
- iv. The expected frequencies for each category is not less than 5 for $df > 2$ and not less than 10 for $df = 2$.
- v. The observed values of χ^2_{observed} with degrees of freedom must be corrected for continuity in order to use the table of values of χ^2_{critical} .
- vi. The sample should contain at least 50 observations.

The degrees of freedom for chi-square statistic depends on the number of categories (k) and not the number of subjects in the sample (N), i.e. the degrees of freedom $df = \text{number of categories} - 1 = k - 1$.

The expected frequencies for each category is not less than 5 for $df > 2$ and not less than 10 for $df = 2$.

Decision: If the observed χ^2 value exceeds tabulated χ^2 value; then the null hypothesis should be rejected i.e. if $\chi^2_{\text{observed}} > \chi^2_{\text{critical}}$; then null hypothesis should not be rejected

Reject H_0 if $\chi^2_{\text{observed}} > \chi^2_{\text{critical}} (\alpha, df)$

Do not reject H_0 if $\chi^2_{\text{observed}} < \chi^2_{\text{critical}} (\alpha, df)$

Example: A weight control clinic was interested in the types of foods that dieters found most difficult to avoid. Sixty clients were randomly selected and individually asked the question: Which of the following foods were difficult to avoid?

- Bread and rice
- Cookies and cakes
- Ice cream and frozen desserts
- Pastries and pies.

The data were obtained as in table-C

Table-C

Number of clients respondent

Bread and rice	Cookies and cakes	Ice cream and frozen desserts	Pastries and pies
17	13	21	9

Conduct a χ^2 -test at $\alpha=.01$ to determine whether these observed frequencies differ from what would be expected if these foods were all equally difficult to avoid.

Solution : Computation of χ^2 observed for data on dieters provided in table- C*

Expected proportion: the sum of expected proportion should be equal to 1.

Table-C*

Categories	observed frequencies, O_i	Expected proportion= $1/\text{categories}$	Expected frequencies $E_i = N \cdot \text{expected proportion}$	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
a	17	.25	15	2	4	.25
b	13	.25	15	-2	4	.26
c	21	.25	15	6	36	2.4
d	9	.25	15	-6	36	2.4

$$\begin{aligned} \text{Now, The observed, } \chi^2_{\text{observed}} &= \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i} \\ &= 5.32 \end{aligned}$$

Hypothesis: The null and alternative hypothesis are:

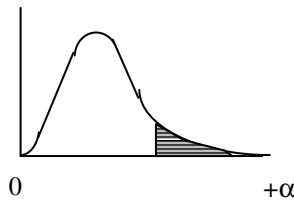
H_0 : There is no difference in the population between observed and expected frequencies.

H_A : There is a significant difference in the population between observed and expected frequencies.

Degrees of freedom (df) : $df = k-1 = 4-1 = 3$

χ^2_{observed} χ^2_{critical} from χ^2 -table with $k-1=4-1=3$

df at $\alpha = .01$ level of significance is $\chi^2_{\text{critical}} (\alpha = .01, df = 3) = 6.25139$



$$\chi^2_{3, .01} = 6.251$$

Decision: Since χ^2_{observed} does not exceed χ^2_{critical} , the null hypothesis should not be rejected. There is insufficient evidence that dieters found most difficult to avoid dieting.

Problem: The table-D shows the observed and expected frequencies in producing a bulb industry was defected bulb per hour in 120 time.

Table-D

Defected bulb per hour	1	2	3	4	5	6
Observed frequencies O_i	25	17	15	23	24	16
Expected frequencies E_i	20	20	20	20	20	20

conduct a χ^2 test at $\alpha = .01$ to determine whether these observed frequencies differ from what would be expected if those bulb were equally difficult to product.

Hints: Include the following items in your answer.

- Null and alternative hypothesis,
- Verification that the assumption and requirement have been met.
- Computation of χ^2_{observed}
- Computation of degrees of freedom (df)
- χ^2_{critical} from the χ^2 -table with required df at $\alpha = .05$ level of significance.
- Conclusions.

Summary: The Chi-square tests are used to determine whether the observed frequencies differ from the expected frequencies. The mean of the χ^2 -distribution is v and the variance of the χ^2 -distribution is $2v$. χ^2 -distribution possessed v additive properties. χ^2 -distribution is unimodal and positive skewed. It also approaches normal distribution when the df grows infinite large. When $\chi^2_{\text{observed}} > \chi^2_{\text{critical}}$ then reject the null hypothesis, on the other hand, if $\chi^2_{\text{observed}} < \chi^2_{\text{critical}}$, the alternative hypothesis is rejected.

Self-Assessment Questions:

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The shape of the chi-square distribution is
 - (a) negative skewed
 - (b) positive skewed
 - (c) both positive and negative skewed
 - (d) U shaped
- (ii) The mean of χ^2 -distribution is
 - (a) equal to degrees of freedom
 - (b) equal to number of categories
 - (c) twice of degrees of freedom
 - (d) twice of number of categories.
- (iii) The sum of expected proportion should be equal to
 - (a) one
 - (b) ten
 - (c) hundred
 - (d) number of cell frequencies.
- (iv) The degrees of freedom for χ^2 -distribution depends on
 - (a) number of categories
 - (b) member of categories-1
 - (c) number of frequencies
 - (d) member of frequencies-1
- (v) The expected frequencies for each categories is not less than
 - (a) 5 for $df > 2$
 - (b) 10 for $df > 2$
 - (c) 15 for $df > 2$
 - (d) 5 for < 2
- (vi) The sum expected frequencies should be equal to
 - (a) number of categories
 - (b) degrees of freedom
 - (c) number of observation
 - (d) one
- (vii) χ^2 distribution have always.
 - (a) negative value
 - (b) positive value
 - (c) Imaginary value
 - (d) infinity value.
- (viii) For large degrees of freedom, χ^2 approaches
 - (a) Binomial distribution
 - (b) Normal distribution
 - (c) Hyper Geometric distribution
 - (d) Uniform distribution.
- (ix) The standard deviation of $\sqrt{2\chi^2}$ distribution is equal to
 - (a) 0
 - (b) -1
 - (c) +1
 - (d) α

Answer

Multiple Choice Questions

- (i) b (ii) a (iii) a (iv) b (v) a (vi) c
- (vii) b (viii) b (ix) c.

Lesson 2: Condition for the Application of χ^2 Test and Uses of χ^2 Table

Lesson Objectives:

After completing this lesson, you will be able to know

- About corrected continuity;
- How to calculate χ^2_{observed} value;
- Yates's correction and
- The uses through example.

Introduction

You have already been known about chi-square tests that are frequently used in social and business studies when the number of subjects falling into particular categories.

For testing whether the observed frequencies in two or more categories come from hypothesised frequencies in their categories, we make five assumptions. If there are three or more, and if one or more of the categories has an expected frequency of less than 5, they may, in some cases, be combined in order to increase the expected frequency in a collapsed category. This procedure is only recommended when combination of categories can be made meaningfully. If χ^2_{observed} values with one degrees of freedom it must be corrected for continuity.

Correction Continuity

The chi-square distribution is a smooth, continuous curve. Observed frequencies, however, change directly. Thus, discrepancies may arise between the smooth curve of the value of theoretical chi-square distribution and sampling distribution of observed values. This inconsistency between the theoretical and actual sampling distribution of chi-square is only serious enough to affect the outcome of a test with one degrees of freedom. In this case, a correction is applied to the observed frequencies known as “Correction for continuity” in order to smooth them.

Yates (1934) apply a strategy for “Corrected continuity” Known as Yates correction for continuity as

- i. Subtract 0.5 from the observed frequencies if the observed frequency is greater than the expected frequency i.e. if $O_i > E_i$; $i=1, \dots, k$, subtract 0.5 from O_i .
- ii. Add 0.5 to the observed frequency if the observed frequency is less than the expected frequency i.e. if $O_i < E_i$; $i=1, \dots, k$, add 0.5 to O_i .

Then the corrected chi-square statistic are given by,

The chi-square distribution is a smooth, continuous curve. Observed frequencies, however, changes directly. Thus, discrepancies may arise between the smooth curve of the value of theoretical chi-square distribution and sampling distribution of observed values

Yates, F (1934) Contegency tables involving small numbers and the f^2 test.

$$\chi^2_{\text{observed (corrected)}} = \sum_{i=1}^k \frac{[O_i - E_i]^2}{E_i}; i=1, \dots, k \text{ and } O_i = \text{observed}$$

frecuencios

i.e. $O_i = O_i \pm 0.5$ [if $O_i > E_i$; then $O_i - 0.5$ and

if $O_i < E_i$; then $O_i + 0.5$

Example: A Statistics class is required by the students of School of Business of Bangladesh Open University in which two-thirds of the students are men and the rest are women. A tutor observed that in a class of 50 students, 30 are men and 20 are women. The tutor wonders whether there are more men in the class than would be expected from the distribution of men and women in the School of Business. In order to find out the reality, the tutor conducts a χ^2 -test at $\alpha = .05$.

Solution:

Since there are two groups, this test is 1 df χ^2 test, and a correction for continuity must be made. The computational table A is given below:

Table-A

Computational table for the statistics class example

category	observed frequency O_i	Expected proportion	Expected frequency E_i	$O_i^* = O_i$ corrected for continuty	$O_i - E_i$	$\frac{(Q_i - E_i)^2}{E_i}$
men	30	.67	33.33	30.5	-2.83	.24
women	20	.33	16.67	10.5	2.83	.48
Total	N=50	1.00	50	50	O	.72

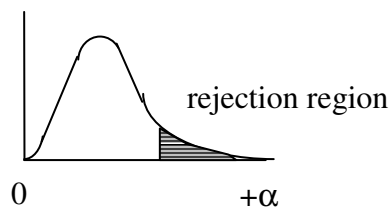
i. observed corrected χ^2 ,

$$\chi^2_{\text{observed (corrected)}} = \sum_{i=1}^k \frac{[O_i^* - E_i]^2}{E_i} = .72$$

ii. Hypothesis: H_0 : no difference in the population between observed and expected frequencies

H_A : a difference between them.

iii. Degrees of Freedom: df = the number of catagories-1
= 2-1
= 1.



$$\chi^2_{1,.05} = 3.841 \text{ rejection region}$$

- iv. Critical χ^2 value at $\alpha=.05$ with 1 df is $\chi^2_{\text{critical}}(1,.05) = 3.841$
- v. Conclusion: Since $\chi^2_{\text{observed}}(\text{corrected}) < \chi^2_{\text{critical}}(1,.05)$

χ^2 observed value does not exceed χ^2 critical, the null hypothesis cannot be rejected. There is insufficient evidence to lead to the conclusion that more men than women enrolled in this statistics course of School of Business of Bangladesh Open University than would be expected from the number of men and women in the school.

Example 1

A large industrial company recently instituted a new type of management training programme. Records of previous programmes showed that above 75% of the participants in these programmes were rated as “Successful” managers by their superiors. Three months after completing the new management training programme participants were rated as “Successful” or unsuccessful managers by their superiors. The number of participants in each category are as follows:

Successful: 37

Unsuccessful: 3

Conduct a χ^2 -test at $\alpha=.01$ to determine whether the number of participants rated as successful was greater than would be expected from previous experience.

- Hints: i. Computation of χ^2 observed
- ii. χ^2_{critical} from χ^2 table
 - iii. Conclusion.

Example 2

In 360 tosses of a pair of dice, 74 sevens and 24 eights are observed. Using the .05 significance level, test the hypothesis that the dice are fair.

Hints: [A pair of dice can fall in 36 way. A seven can occur in 6 ways and an eight can occur in 2 ways. Thus $\Pr[\text{seven}] = \frac{6}{36} = \frac{1}{6}$ and $\Pr[\text{eight}] = \frac{2}{36}$ Thus in 360 tosses we would have expected frequencies, $360 \times \frac{1}{6} = 60$ for seven and $360 \times \frac{2}{36} = 20$ for eight]

Uses of Chi-Square Test

- i. Chi-square test is used in behavioral research.
- ii. Chi-square test is used to test whether observed frequencies differ significantly from the expected frequencies.
- iii. The chi-square test can be used to determine how well theoretical distributions fit empirical distribution.
- iv. Some important applications of χ^2 test:
 - i. Sampling Distribution of the sampling variance:

When the parent population is normal with variance σ^2 and if random sample of size n with sample variance s^2 the χ^2 statistic can be shown as-

$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2$ distribution with v degrees of freedom.

When the parent population is normal with variance σ^2 and if random sample of size n with sample variance s^2 the χ^2 statistic can be shown as-

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2 \text{ distribution with } v \text{ degrees of freedom.}$$

ii. Confidence Interval for variance:

The sampling distribution of χ^2 is $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

A $100(1-\alpha)$ percent Confidence Interval (C.I) for σ^2 is

$$\chi^2_{\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{(1-\alpha/2)}$$

$$\text{or, } \frac{(n-1)S^2}{\chi^2_{(1-\alpha/2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\alpha/2}}$$

Two value of χ^2 is to the left of the smaller value and $\alpha/2$ is to the right of the larger value.

iii. Test of hypothesis concerning variance; If the null hypothesis $H_0: \sigma^2 = \sigma_0^2$; σ_0^2 is the specified value of the population variance.

Let $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2$ distribution for large sample of size n then $\chi^2 < \chi^2_{1-\alpha/2}$ and $\chi^2 > \chi^2_{\alpha/2}$ i.e. when the computed χ^2 lies in rejection region, we reject the null hypothesis.

Summary: The theoretical sampling distribution of χ^2 and the actual sampling distribution is called Yates's correction. if $O_i > E_i$, $i=1,2, \dots, K$, subtract 0.5 from O_i and if $O_i < E_i$, $i=1,2, \dots, k$; add .05 to O_i .

Self-Assessment Questions

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) When Yates's correction is necessary ?
 - a. $df = 5$
 - b. $df = .05$
 - c. $df = 1$
 - d. $df = .005$
- (ii) When add 0.5 to the observed frequency ?
 - a. if $O_i > E_i$; $i=1,2, \dots, k$
 - b. if $O_i E_i = 0$; $i=1,2, \dots, k$
 - c. if $O_i < E_i$; $i=1,2, \dots, k$
 - c. $\frac{O_i}{E_i} = 0$; $i=1,2, \dots, k$
- (iii) When subtract 0.5 from the observed frequency ?
 - a. if $O_i \infty E_i = 0$; $i=1,2, \dots, k$
 - b. If $> E_i$; $i=1,2, \dots, k$
 - c. If $O_i/E_i = 0$; $i=1,2, \dots, k$
 - d. If $O_i < E_i$; $i=1,2, \dots, k$
- (iv) The sum of difference between observed and expected frequency is equal to
 - a. 10
 - b. 5
 - c. 1
 - d. 0

Answer:

Multiple Choice Questions: (i) c (ii) a (iii) d (iv) 4.

Lesson 3: Test of Indipendence

Lesson Objectives:

After completing this lesson, you will be able to know:

- Define contingency table;
- Explain test of independence;
- Construct the test procedure.

Introduction

We are frequently concerned with enumeration data, which consists of frequencies or counts rather than qualitative measurements. Enumeration data are often classified according to several variables. For example, a person may be classified as a smoker or a non-smoker and as one with or without coronary disease. There are two variables of classification. Where the individuals constituting a sample are classified according to two attributes and arranged in a rectangular array, known as contingency table. Now the test of independence in a contingency table indicates whether the two classifications are independent or not if they are not independent there is said to be interaction and for this the test is often called a test of interaction.

Contingency Table

When χ^2 -test consists of at least two independent variables, each with two or more levels, and a dependent variable in the form of a frequency count, then we can arrive at 2 way classification table, or contingency table, in which the observed frequencies occupy rows and columns. Such table is often called contingency table.

Corresponding to each observed frequency in a contingency table (see, table A), there is an expected or theoretical frequency that is computed subject to the hypothesis according to the rule of probability. These frequencies, which occupy the cell of a contingency table, are called cell frequency. The total frequency in each row or each column is called marginal frequency.

When χ^2 -test consists of at least two independent variables, each with two or more levels, and a dependent variable in the form of a frequency count, then we can arrive at 2 way classification table, or contingency table

Table A
Contingency table

	Attribute B				Total
	B ₁	B ₂	B ₃	B _c	
A ₁	O ₁₁	O ₁₂	O ₁₃	O _{1c}	R ₁
A ₂	O ₂₁	O ₂₂	O ₂₃	O _{2c}	R ₂
A ₃	O ₃₁	O ₃₂	O ₃₃	O _{3c}	R ₃
.	.	.	.	.	R _r
.	.	.	.	.	
.	.	.	.	.	
.	.	.	.	.	
A _r	O _{r1}	O _{r2}	O _{r3}	O _{rc}	
Total	C ₁	C ₂	C ₃	C _c	N

Test Criterion: The test criterion is given by the equation

$$\chi^2_{\text{observed}} = \sum_{i=1}^r \sum_{j=1}^c \frac{[O_{ij} - E_{ij}]^2}{E_{ij}}; \quad i=1,2,\dots,r, j=1,2,\dots,c$$
$$= \sum_{i=1}^r \sum_{j=1}^c \frac{O_{ij}^2}{E_{ij}} - N; \quad \text{which is } \chi^2\text{-distribution with}$$

$(r-1)(c-1)$ degrees of freedom.

Where $O_{ij}; i=1,2,\dots,r, j=1,2,\dots,c$ = observed frequency in the i th row j th column of the contingency table.

$E_{ij}; i=1,2,\dots,r, j=1,2,\dots,c$ = is the corresponding expected frequency

and $E_{ij} = \frac{R_i C_j}{N}$; where

R_i = i th row total

C_j = j th column total and

N = grand total of the frequencies.

Assumption:

- i. Each observation must fall in one and only one cell.
- ii. Each observation is independent of every other observations
- iii. The observations are measured as frequencies.
- iv. The expected frequency for any cell is not less than 5 for $df > 2$ and not less than 10 for $df = 1$.
- v. The observed values of χ^2 with one df , contingency table must be corrected for continuity in order to use the table of value χ^2 critical.

The Null Hypothesis

H_0 : the two attributes are independent

H_A : the two attributes are not independent.

Computed Expected Frequency:

Compute expected values using the formula: Expected number in cell

$$= \frac{\text{Row total} \times \text{Column total}}{\text{Total number in Sample}}$$

and place each expected value in the lower right-hand corner of the appropriate cell in the table of observed frequencies.

Decision Rules for Rejecting H_0 :

In order to decide whether to reject H_0 , we must compare χ^2_{observed} with $\chi^2_{\text{critical}} (\alpha, df)$. For a χ^2 contingency table, $df = (r-1)(c-1)$. with α df known.

χ^2 table can be used to find χ^2_{critical} .

If $\chi^2_{\text{observed}} > \chi^2_{\text{critical}} (\alpha, df)$; we reject the null hypothesis. i.e. two attributes are associated.

$\chi^2_{\text{observed}} < \chi^2_{\text{critical}} (\alpha, df)$: we accept the null hypothesis.

Example

It was found that the sex of the students and the students' choice of academic major were related; women tended to choose majors which gave higher grades. The result explained the apparent difference in achievement between men and women; it was not necessary to invoke certain masculine and feminine characteristics as hypothesised in the past. Table-B presents the data on choice of academic major for 897 male and female undergraduates at UCLA.

Table-B
Academic major

Sex	Physic	Engineering	English	Design	Total
Male	108	345	94	17	564
Female	8	12	253	60	333
Total	116	357	347	77	897

Ans: The data presented as 2x4 contingency table.

$$\text{Computation } \chi^2_{\text{observed}}: \chi^2_{\text{observed}} = \sum_{i=1}^r \sum_{j=1}^c \frac{[O_{ij} - E_{ij}]^2}{E_{ij}}; i=1,2 -$$

-- r, j=1,2 ----c

Expected frequency table can be given as—

Sex	Physics	Engineering	English	Design	Total
Male	72.94	224.47	218.18	48.42	564.01
Female	43.04	132.53	128.82	28.59	332.98
Total	115.98	357	347	77.01	897

$$\begin{aligned}
 \text{Now, } \chi^2_{\text{observed}} &= \frac{[108-72.94]^2}{72.94} + \frac{[345-224.40]^2}{224.47} + \frac{[94-218.18]^2}{218.18} \\
 &\quad + \frac{[17-48.42]^2}{48.42} + \frac{[8-43.04]^2}{43.04} + \\
 &\quad + \frac{[12-132.53]^2}{132.53} + \frac{[253-128.82]^2}{128.82} + \frac{[60-28.59]^2}{28.59} \\
 &= 16.85 + 64.72 + 70.68 + 20.39 + 28.55 + 109.62 + 119.74 + 34.51 \\
 &= 465.06 \\
 \therefore \chi^2_{\text{observed}} &= 465.06
 \end{aligned}$$

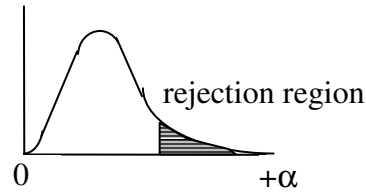


Fig: rejection region; $\chi^2_{\text{tab}} > \chi^2_{3,0.01} = 11.3$

degrees of freedom, $df = (r-1)(c-1) = (2-1)(4-1) = 1 \times 3 = 3$

If the test is conducted at $\alpha = .01$

$$\chi^2_{\text{critical}} (.01, 3) = 11.345$$

Null Hypothesis

H_0 : Sex and choice of academic major are independent in the population.

H_A : Sex and choice of academic major are related in the population.

Decision

Since the observed χ^2 , $\chi^2_{\text{observed}} > \chi^2_{\text{critical}}(3, .01)$ i.e, χ^2_{observed} value exceed $\chi^2_{\text{critical}}(3, .01)$ value, so the null hypothesis should be rejected clearly, students's sex is related to choice of aacademic major.

Example: In a trial to determine the effect of a certain vaccination upon the development of a certain deasee, one got the following result.

	Disease	Not-disease	Total
Not vaccinated	265 (142.5)	35 (157.5)	300
Vaecinated	20 (142.5)	280 (157.5)	300
Total	285	315	600

The figures within the bracket are the number of estimated person and those without the braket are the observed number of person.

Can you draw χ^2 -test on the effectiveness of this new vaccine to control the disease at $\alpha=.05$?

Hint: i. Null and alternative hypothesis

ii. Verification that the assumptions and requirement have been met

iii. Computation of χ^2_{observed}

iv. χ^2_{critical}

v. Conclusion

The 2x2 χ^2 -test is a statistical test with one degrees of freedom. Whenever the theoritical sampling distribution of chi-square is used with 1 df, Yates's correction for continuity should be used.

Correction for Continuity for the 2 x 2 χ^2 -test :

The 2x2 χ^2 -test is a statistical test with one degrees of freedom. Whenever the theoritical sampling distribution of chi-square is used with 1 df, Yates's correction for continuity should be used.

Since 2x2 chi-square is only 2x2 contingency table with one df, a simple computational formula which incorporates Yeats' correction is available. Consider the following 2x2 contingency table:

a	b	a+b
c	d	c+d
a+c	b+d	N=a+b+c+d

the letters a, b, c, d represent the cell frequencies. From this table, the value of χ^2 can be found by using the following formula.

$$\chi^2 = \frac{N[|ad-bc| - N/2]^2}{(a+b)(b+c)(c+a)(d+a)} \text{ with df} = 1$$

Example: Consider some data are taken from a hypothetical surveys are given below of Bangladeshi and Indian businessmen about whether they preferred soft drink with lunch.

Table:

For hypothetical survey of soft drink preference of Bangladesh and Indian businessmen.

Nationality	Do you Prefer soft drink with lunch		Total
	Yes	No	
Bangladeshi	a = 54	b=6	a+b=60
Indian	c = 16	d=24	c+d = 40
Total	a+c=70	b+d = 30	a+b+c+d = 100

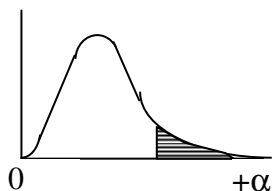
Solution : Before proceeding with statistical test, we should check the smallest expected frequencies to see whether it is at test 10 for this df=1 test now the expected frequency table are as following table for expected frequencies.

Nationality	Yes	No	Total
Bangladeshi	a=42	b=18	60
Indian	c=28	d=12	40
Total	70	30	100

So, smallest expected frequency is greater than 10. Then the

$$\text{i. } \chi^2_{\text{observed}} = 100 \frac{[|54 \times 24 - 6 \times 16| - 100/2]^2}{60 \times 40 \times 70 \times 30} = \frac{100[|129696| 50]^2}{5040000}$$

$$\chi^2_{\text{observed}} = \frac{132250000}{5040000} = 26.24$$



- ii. χ^2 critical ($\alpha=.05,1$) = 3.841 ; From the χ^2 -table at .05 level of χ^2 -curve significance with $df=1$.
- iii. **null hypothesis:** H_0 : There is no difference between Bangladeshi and Indian businessmen regarding soft drink preference
 H_A : There is a significant different between Bangladeshi and Indian businessmen regarding soft drink preference.
- iv. **Conclusion:** Since $\chi^2_{\text{observed}} > \chi^2_{\text{critical}} (.05,1)$ i.e. the null hypothesis should be rejected. We conclude that there is a relation between nationality and soft drink Preference.

Activity:

Example: From the following information test χ^2 criterion Yeat's correction

	effected	not affected	Total
vaccinated	3	188	191
not vaccinated	30	112	142
Total	33	300	333

Summary: When we can arrive at two-way classification tables or $r \times c$ tables, in which the observed frequencies occupy 'r' row and 'c' columns. Such table is often called contingency table. χ^2 -statistics provided the expected frequency are too small that it should be greater than 10. The expected frequency is computed subjects to some hypothesis according to the rule of probabilities. The total frequencies in each row or each column are called marginal frequencies.

Self-Assessment Questions

Multiple-Choice Question:

1. **Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:**
 - (i) $r \times c$ table; in which the observed frequencies occupy 'r' rows and 'c' column is called:
 - (a) frequency table
 - (b) χ^2 -table
 - (c) contingency table
 - (d) F-table
 - (ii) The total frequencies in each row or each column are called.
 - (a) partial frequency
 - (b) expected frequency
 - (c) proportional frequency
 - (d) marginal frequency.
 - (iii) The frequencies, which occupy the cells of a contingency table, are called:
 - (a) cell frequencies
 - (b) marginal frequencies
 - (c) expected frequencies.
 - (d) proportional frequencies.
 - (iv) When the contingency table is arranged, 2×2 called:
 - (a) 2×2 contingency table
 - (b) 2×2 design
 - (c) 2×2 factorial design
 - (d) rectangular table.

Answer:

Multiple Choice Questions

- (i) c (ii) d (iii) a (iv) a

Lesson 4: Test of Goodness of Fit and Test of Homogeneity

Lesson Objectives:

After studying this lesson, you will be able to explain:

- What goodness of fit test is?
- Test of homogeneity with example.

Introduction

In the chi-square test for one way classification, there must be one independent variable with two or more levels. A subject may be counted in one and only one cell of the classification. And the dependent variable is a count in the form of frequencies, proportions, probabilities or percentages.

Goodness of Fit: The purpose of the chi-square test is to determine whether the observed frequencies differ systematically from the theoretically expected frequencies or, whether the difference may be due to chance. Often this chi-square test is called a goodness of fit because it examines how closely the observed frequencies fit the theoretical expected frequencies.

The chi-square test is called a goodness of fit because it examines how closely the observed frequencies fit the theoretical expected frequencies.

In some cases, we wish to compare an observed frequency distribution with some theoretical frequency distribution like the normal distribution. In developing χ^2 -statistic to measure how closely the observed frequencies, (denoted by O) match the theoretical expected frequencies (denoted by E), we might subtract E from O in each categories and then add the difference over all of k categories. But, the result will always be zero.

$$\sum_{i=1}^k (O_i - E_i) = \sum_{i=1}^k O_i - \sum_{i=1}^k E_i = N - N = 0$$

In order to avoid this problem we can square the difference between O and E : $(O - E)^2$

Thus one possible statistic for comparing observed and expected frequencies is:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}; i = 1, 2, \dots, k$$

Prcedure for a χ^2 Goodness of Fit Test:

The population are grouped according to the probability distribution:

Group	Probability
1	P_1
2	P_2
.	.
k	P_k

Hypothesis: The null hypothesis, H_0 : the population are grouped according to the probabality distribution.

The alternative hypothesis, H_A : the population is not grouped according to the probability distribution in the null hypothesis.

Assumption: 1. No accepted frequency is less than 1

2. At most 20% of the expected frequency are less than 5.

Level of Significance : The level of significan, $\alpha = .05, .01$

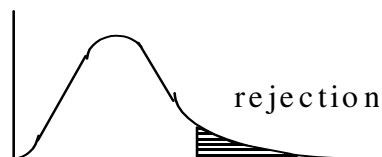
Test Statistics: The test statistics are given as $\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$;

$i=1,2,----k$

Make a work table of observed and expected frequencies in each group and calculate as given below:

Group	O_i	E_i	$O_i - E_i$	$\frac{(O_i - E_i)^2}{E_i}$
1	O_1	E_1	$O_1 - E_1$	$\frac{O_1 - E_1}{E_1}$
2	O_2	E_2	$O_2 - E_2$	$\frac{O_2 - E_2}{E_2}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
K	O_k	E_k	$O_k - E_k$	$\frac{O_k - E_k}{E_k}$
Total				$\sum_{i=1}^k \frac{O_i - E_i}{E_i}$

Decision: If $\chi^2 > \chi^2_{tab}$, reject the null hypothesis where $\chi^2_{tab} = \chi^2_{(n-1)}$, a from table at desired level of significance.



rejection region.

Example: A new tract has the same age distribution as the order tract 3198 did in 1999. The age distributions for the new tract are given below—

Age group	Probability
0-20	0.392
21-65	0.584
65 over	0.024

Is the distribution rely on the fact? Where the sample size is 300?

Solution : We proceed through the following steps:

1. Hypothesis; Ho: There is age difference between the old and the new tracts.

H_A : There is no age difference between the old and the new tracts.

2. Expected frequency table:

Age group	Probability	Expected frequency E = np
0.20	0.392	117.6
21-66	0.884	175.2
65>over	0.024	7.2

Now we check the assumption:

- a. no expected value i.e. E is less then 1 Here no value is less then one.
- b. at most 20% of the E value are less than 5.

So we can perform the test statistic as χ^2 -good ness of fit.

Now: Consider $\alpha = .05$

A. The test statistic—
$$\chi^2_{cal} = \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i}$$

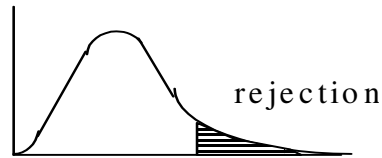
Now,

Age group	O	E	O-E	(O-E) ² /E	
0-20	126	117.6	8.4	.60	$\chi^2 = \frac{(O-E)^2}{E}$ $= 1.64$
21-64	165	175.2	-10.2	.59	
65>over	9	7.2	1.8	.45	
Total					

$\therefore \chi^2_{cal} = 1.64$

- B. Best test: Reject the null hypothesis if $\chi^2_{cal} > \chi^2_{(k-1)}$; a otherwise accept the null hypothesis,

$$\text{Now, } \chi^2_{tab} = \chi^2_{3-1, .05} = \chi^2_{2, .05} = 5.99$$



Rejection region

- C. Decision: Here $\chi_{cal} < \chi^2_{tab} = 80$ the null hypothesis is accepted i.e. there is insufficient evidence i.e. there is a difference in age distribution for new and old tracts.

Activity:

School of Business of Bangladesh Open University has agreed upon the grading policy shown below for its introductory course.

Grades	Observed frequency	Probability
A	10	.10
B	10	.20
C	10	.40
D	14	.20
E	6	.10

The Dean of School of Business is not sure that the policy was followed last semester. Since the introductory course had 2000 student in 50 Section, he does not wish to tally the grades on all there grade sheets. perform the hypothesis test at the .05 significance level.

Test of Homogeneity:

In the independence test of χ^2 ; we tried to determine whether two characteristics of individuals in the same population were independent. But in test of homogeneity test of χ^2 , we look at characteristics of individuals from different population and lead to very similar data tables with the use of same computational methods as independent test.

Where two populations do have identical percentages for each category in a grouping, they are called homogeneous with respect to that grouping.

Where two populations do have identical percentages for each category in a grouping, they are called homogeneous with respect to that grouping.

Differences between the "Test of Homogeneity "and the" Test of Independence"

Test of Homogeneity	Test of Independence
1. We are concernd wheather the different samples come from the same population.	1. We are concerned with the prolem whether the two attributes are independent or not.
2. The test involves two or more samples on the from each population.	2. The best involves a gingle sample from each population.

For test of homogeneity we follow the following steps:

1. **Hypothesis:** H_0 : the population are homogeneous with respect to the racial categories

H_A : the population are not homogenous with respect to the given categories.

2. **Test statistic:**
$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}; i=1,2, \dots, k$$

Where O = observed frequency

E = expected frequency

3. **Best test:** Reject the null hypothesis if $\chi^2_{cal} > \chi^2_{(k-1)}$; otherwise accept the null hypothesis.

Cautions while Applying χ^2 -Test

χ^2 - test is a very important and popular practice in the business field. The test must be used with great care. Some sources of error in the application are given below:

- i. Small theoretical frequencies.
- ii. Uses of non-frequency data
- iii. Neglect of frequencies of non-occurrence.
- iv. Incorrect categorising
- v. Indeterminate theoretical frequencies.

Example: There are three categories which are given in the following table:

	White	Black	Other	Total
A	83	5	12	100
B	87	6	7	100
Total	170	11	19	200

Test whether each one cells for a test of homogeneity.

Solution: Now we compute an expected value for the cells:

Table for observed value					Table for Expected frequency				
	White	Black	other	Total	White	Black	other	Total	
A	83	5	12	100	A	85	5.5	9.5	100
B	87	6	7	100	B	85	5.5	9.5	100
Total	170	11	19	200	Total	170	11	19	200

1. Hypothesis: H_0 : The population are homogenous with respect to the racial categories white, black and other.

H_A : The population are non homogenous with respect to the given categoris.

$$\begin{aligned}
 2. \quad \text{Test Statistics: } \chi^2 &= \sum_{i=1}^3 \frac{(O_i - E_i)^2}{E_i} = \frac{(83-85)^2}{85} + \frac{(5-5.5)^2}{5.5} + \\
 &\quad \frac{(12-9.5)^2}{9.5} + \frac{(87-85)^2}{85} + \frac{(6-5.5)^2}{5.5} + \frac{(7-9.5)^2}{9.5} \\
 &= .0005 + 0.005 + 0.66 + 0.05 + .05 + .66 \\
 &= 1.52 \\
 \therefore \chi^2_{\text{cal}} &= 1.52
 \end{aligned}$$

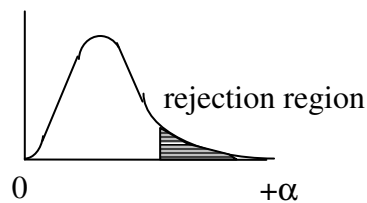


Fig : Rejection region

3. Best test: reject the null hypothesis if $\chi^2_{\text{cal}} > \chi^2_{(n-1)}$; otherwise accept the null hypothesis.
4. Decision: Here $\chi^2_{\text{cal}} < \chi^2_{\text{tab}}$ so the null hypothesis is accepted i.e. population are homogenous with respect to the racial catagories.

Activity:

Example

On February 1998, the Television shows different programme of BOU. On the first night SSC programme was watched by 41 percent of the viewing audience while the remaining of the audience was split between B.Ed programme and MBA programme. A sample of size 100 in two city, Noapara and Fultala under Khulna division shows the following break-down.

	SSC prog.	B.Ed Prog.	MBA prog.
Noapara	42	28	29
Fultala	39	33	28

Using $\alpha = .01$, perform a χ^2 -test of homogeneity to decide whether the viewing audiences in the two cities were distributed among the three programmes shows in the same way.

Self-Assessment Question:

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) When the observed frequencies is very close to the theoretical expected frequencies then χ^2 -test is called:
 - (a) Test of homogeneity (b) Goodness of fit.
 - (c) Independent test (d) non-parametric test.
- (ii) In the chi-square test for one way classification there must be one:
 - (a) dependent variable (b) independent variable
 - (c) inter-dependent variable (d) random variable.
- (iii) When χ^2 -test finds out the characteristics of individual from different population, it is called
 - (a) Goodness of fit (b) Non-parametric test
 - (c) Test of homogeneity (d) Bank test.
- (iv) In which test, the null hypothesis usually states that the sample is drawn from theoretical population distribution.
 - (a) Test of homogeneity (b) Non-parametric test
 - (c) Test of goodness of fit. (d) Sign test.
- (v) In which test χ^2 -test contain two more samples, one from each population.
 - (a) Sign test (b) Test of goodness of fit
 - (c) Test of homogeneity (d) Test of Independence

Answer:

Multiple Choice Questions

- (i) b (ii) b (iii) c (iv) c (v) c.

Exercise

1. What is χ^2 -test? Point out its role in Business decision making.
2. What is χ^2 -test of independence? Explain correction continuity for small frequencies in contingency table.
3. What is χ^2 -test of goodness of fit? What cautions are necessary while applying this test?
4. Discuss the chi-square test of homogeneity. State the conditions for the validity of chi-square test.
5. For 2×2 contingency table

	A	not A
B	a	b
not B	c	d

Prove that χ^2 test for independence gives—

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

6. Write short notes on the following:
 - (i) Goodness of fit
 - (ii) Yates's corrections for continuity
 - (iii) Degrees of freedom
7. The following data related to the sales in a time of trade depression of goods in wide demand

District wise sales are:

	District	
	Not hit by depression	Hit by depression
Satisfactory	120	40
Not satisfactory	20	40

Do the data suggest that the sales are significantly affected by depression?

8. The following table shows the classification of 2000 workers in a factory

Promotional experience \ Disciplinary action	Promoted	Not promoted
Not-offenders	146	462
Offenders	54	1338

Test whether the disciplinary action taken and promoted experience are independent.

SAMPLING AND SAMPLING DISTRIBUTION

12

The importance of sampling in statistics has been widely recognized not only by the statisticians but by the users of statistical data as well. For this reason the basic idea about sampling i.e. methods of selection of samples and obtaining of estimates from them must be known at this level. The major points of sampling and non-sampling error, sampling distribution, acceptance statistical quantity control, sampling and sampling plan have been explained with the hope and you should be led to realize the importance of samples. The uses and importance of complete count and its relation with the sampling have been given in the unit.

Lesson 1: Sampling – Purposes & Types

Lesson Objectives:

After completing this lesson you will be able to explain

- Define Sampling;
- Explain the purpose of sampling;
- Mention the principles of sampling;
- Describe the methods of sampling.

Introduction

The proper decisions in business as well as in the different field of our daily lives depend on the adequate and reliable data which can be popularly known as obtained by either complete enumeration survey, census method or by using sampling technique. Due to time constraint and the cost involved in census method to obtain information, it becomes obvious to draw inferences for the population on the basis of sampling information. The sample is a part or small section selected from a population. This sampling theory is a study of relationship that exists between the population and the samples drawn from the population. The complete enumeration, popularly known as census, may not be feasible either due to non-availability of time or because of high cost involved. Therefore, it becomes essential to draw inferences for the population on the basis of sample information. Thus sampling helps us to get as much information as possible of the whole universe and based on which we can make the inference about the whole universe.

Sampling

The sample is a part or small section selected from the population, and the process of such selection is known as sampling. Thus, the sampling theory is a study of relationship that exists between the population and the samples drawn from the population. Sampling becomes essential to draw inferences for the population on the basis of sample information. Thus sampling helps us to get as much information as possible of the whole population.

The sample is a part or small section selected from the population, and the process of such selection is known as sampling

Purpose of Sampling

The purpose of sampling is to make inference about a population. When one has to make an inference about a lot and it is not practically able to examine each individual member of the lot; one invisibly takes resources to sampling that is to say, one examines only a few members of the lot and, on the basis of this sample information, one makes decision about the whole lot. For example, a person wanting to purchase a basket of mangoes, may examine a few mangos from the lot as sample and on the basis of this sample information can make his decision about the mangoes of the basket.

Sample results reflect the characteristics of the population.

In order that sample results reflect the characteristics of the population, it is necessary that the sample selected should be

- i. Truly representative; i.e. the selected sample truly represents the universe so that the results can be generalized.
- ii. Adequate; i.e. the size of the sample or the sample size should be adequate enough to represent the various characteristics of the universe.
- iii. Independent; i.e. the elementary units selected should be independent of one another and all units of the population should have the same chance of being selected in the sample, and
- iv. Homogenous; i.e. there should not be any basic difference between the characteristics of the unit in the sample and that of the population. This means that if two or more samples are drawn from the same population the result should be identical.

Principles of Sampling

There are three basic principles for the sampling. These principles are:

- i. Validity and ii. Optimization.
- i. **Validity:** By validity of sampling we mean that the sample should be selected in such a way that the results be interpreted objectively in terms of probability. In other words, valid tests or estimates about the population characteristics must be available.
- ii. **Optimization:** The optimization takes into account the factors, which can be stated as:
 - (a) **Efficiency:** Efficiency is measured by the inverse of the sampling variance of the estimator.
 - (b) **Cost:** Cost is measured by expenditure incurred in terms of money or man-hours.

The term optimization ensures, that a given level of efficiency will be reached with minimum cost or that the maximum possible efficiency will be attained with a given level of cost.

Thus, the term optimization ensures, that a given level of efficiency will be reached with minimum cost or that the maximum possible efficiency will be attained with a given level of cost.

Merits and Demerits of Sampling

Merits of Sampling:

The benefits on merits of sampling can be stated in the following way:

1. **Time Saving:** Sampling involves less amount of time than a complete count both in the execution and the analysis of data.
- ii. **Economy:** The expenditure or labour increases as the amount of work increases. The size of the sample may for example be only 5% of the size of the population or even less. Naturally, the overall

expenses and labour required to study a sample will be considerably low.

- iii. **Quality and Accuracy:** It is often experienced that the results from a sample, are as good as those obtained by complete count.
- iv. **Feasibility:** Some time the data are obtained by tests that are destructive. To know the average life of a certain types of electric bulb (phillips), we will take a sample of theses bulbs and keep them on still they burn out. We cannot think of testing the whole lot. We must have recourse to a sample and a rather small sample at that.
- v. **Sampling Error Reduce:** A properly designed sample will itself give an idea of the magnitude of the sampling error involved in the estimator.

Demerits of Sampling: In the basic facts about each and every unit in the population are desired, census becomes indispensable a requirements. Some example in which the use of a sample is not allowed are:

- i. When the population is very high i.e. finite and it would be impossible to conduct census surveys.
- ii. Sometimes it is difficult to draw inference about the population when a small sample is taken from it. A sample if it is small, may not contain units from some of theses groups and thus will not reveal any characteristic concerning them.
- iii. An inventory of all goods is necessary to know the total amount of goods of a firm. The incomplete list will provide an underestimate/overestimate of the total value of goods. This underestimate/overestimate will depend upon the extent to which the inventory is incomplete selection of a sample will not help to find this value.
- iv. The lists of income tax payers are also prepared very carefully so that nobody is spared from paying the income tax. Incomplete information breaks the sampling techniques.

Sampling Methods

The meaningfulness of estimates obtained from a sample depends on the methods of selection of a sample. The classification of sampling can be shown in the following figure:

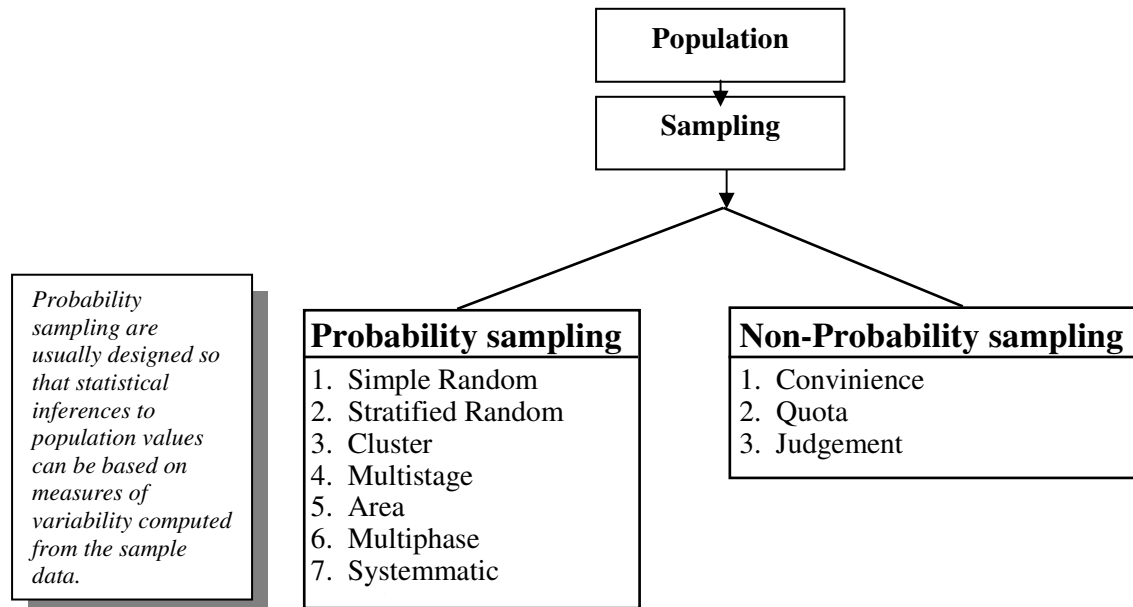


Figure12.1: Showing the Methods of Sampling

Probability Sampling: In probability sampling; each and every element in the population has an equal chance of being included in the sample. This probability is attained through some mechanical operation of randomization. An ideal probability sampling, the inferences to the population can be made entirely by the statistical methods. Probability sampling is usually designed so that statistical inferences to population values can be based on measures of variability computed from the sample data. This method is also known as the method of chance selection because of the selection of items in the sample depends completely on the chance.

i. Simple Random Sampling: In simple random sampling, drawing of elements from the population is done randomly and the choice of an element is made in such a way that each and every element has the same probability of being chosen. Simple random sampling can be arranged in two types a. Simple random sampling with replacement b. Simple random sampling without replacement.

ii. Stratified Sampling: In stratified sampling, the population is subdivided into strata before the sample is drawn. Strata mean the homogeneous groups or classes of population under study. To design a more efficient sample, the researcher has to divide the whole population into different strata and then he/she can proceed to select the sample from each group by simple random method and the outcome is known as the stratified sample. A stratified sampling can be either proportional or disproportionate. In proportional stratified sampling, the total population is divided into different strata and the number of sample items

A compromise between cluster sampling and the direct sampling of unit can be achieved by selecting a sampling of cluster and studying only a sample of units in each sample cluster instead of completely studying all the units in the sample of clusters.

is drawn from each stratum proportionately based on the size of the strata. We can make it more clear by putting an appropriate example. Suppose, a sample of 1,000 is to be drawn from the population divided into four strata. The respective size of the samples from each stratum is 10, 20, 25 and 45 percent of the total samples. So, the task of completing the sampling can be done in the following manner.

From Stratum – 1 :	$1,000 (0.10) = 100$
From Stratum – 2 :	$1,000 (0.20) = 200$
From Stratum – 3:	$1,000 (0.30) = 250$
From Stratum – 4:	$1,000 (0.45) = 450$
Sample Size = 1,000	

This method of stratified sampling represents the universe because of the proportional selection of the samples from each stratum based on the size of the population in each stratum. On the other hand, when an equal number of samples/items is chosen from each stratum irrespective of the size of the population is known as the disproportionate stratified sampling.

iii. Cluster Sampling: In stratified sampling we divide our population into homogenous groups that we call strata, and chose sample from every stratum. Sometimes the population is divided into a large number of group called clusters, and to sample among the cluster is called cluster sampling and is frequently used in social survey in order to cut down on the cost of gathering data. For example, we can divide city into several hundred ceusus tracts and then may select 40 tracts randomly for our sample. It is important to note here that though the clusters are chosen random, but the each and every item of the selected clusters are considered for the information. The aim in cluster sampling is to select clusters that are heterogeneous as possible but that are small enough to cut down expenses.

iv. Multistage Sampling

We have seen that through cluster sampling is economical under certain circumstances, it is generally less efficient than the sampling of individual units directly. A compromised between cluster sampling and the direct sampling of unit can be achieved by selecting a sampling of cluster and studying only a sample of units in each sample cluster instead of completely studying all the units in the sample of clusters. This procedure is known as two-stage sampling, since the units are selected into two stages. This procedure can be easily generalized to give rise to multistage sampling, where the sampling units at each stage are cluster of units of the next stage and the ultimate observational units are selected in stages, sampling at each stage being done from each of the sampling units or cluster selected in the previous stage. This procedure being a compromise between direct sampling of units and cluster sampling, is (i) more efficient than direct sampling and less efficient than cluster sampling as for as the operational convenience and cost are concerned, and (ii) less efficient than direct sampling and more efficient than cluster

sampling from the view point of sampling variability. The multistage sampling systems is as follow:

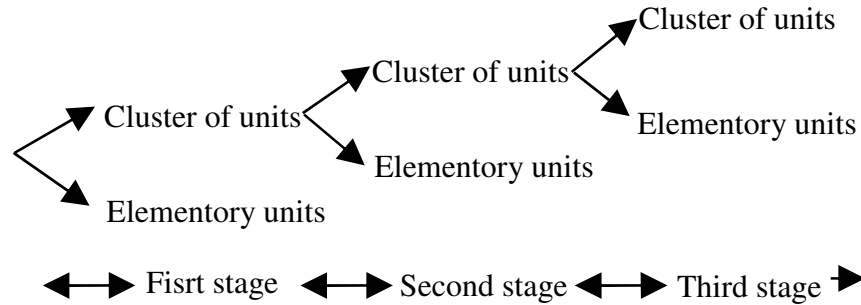


Fig. 12.2: Multistage sampling units

The sampling process at each stage may either be random stratified. We have simple cluster sampling of one or more stages where the sampling units chosen at each stage are selected by the method of simple random sampling. We have stratified cluster sampling of one or more stage if stratification is employed for selection of sample units.

A sample of suitable size is obtained by taking every unit, one of the first units in this ordered arrangement is chosen at random and the sample is completed by selecting every unit from the rest of the list. This mode of selection is called systematic sampling.

v. Area Sampling: A form of cluster sampling that is widely and is the one that associate the elementary unit of population with geographical area. In most applications of this method, the unit under study are human being. Each of this unit must be associated with a single definable area. By random methods, a sample of area is selected. If needed, sub samples of the chosen sample areas may also be selected by random methods. This types of sampling methods are called area sampling.

vi. Multiphase Sampling: The term multiphase sampling is used when sampling unit of the same type are the objects of different phases observation. In one of these phases all the units in a sample are studied with respect to characteristics, while at a later phase, some of the units, a sub sample of the full sample is studied with respect to some additional characteristics. Thus, we should have two phases or double sampling. This type of sampling can be generalized for multiphase and is called multiphase sampling.

vii. Systematic Sampling: When the numbers of population to be sampled are arrange in order, the order corresponding to consecutive numbers. A sample of suitable size is obtained by taking every unit, one of the first units in this ordered arrangement is chosen at random and the sample is completed by selecting every unit from the rest of the list. This mode of selection is called systematic sampling.

Suppose N units of the population are numbered from 1 to N and sample size n is to be selected such that $\frac{n}{N} = \frac{1}{K}$; K being an inteer. Consider a population of eight companies, say, a, b, c, d, e, f, g and h. If a sample of size 2 is to be chosen, then $K = \frac{8}{2} = 4$ in this case, the possible sample in systematic sampling will be shown in figure.

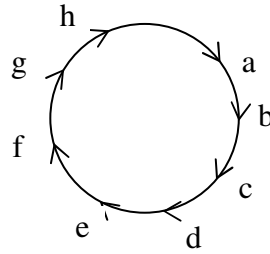


Figure 12.3: Systematic sampling in a cyclinal diagram

Non-Probability Sampling

When the selection of sample does not depend on the chance, rather it depends on the judgements or exercises of the investigators or set criteria, then it is called non-random or non-probability sampling. Sometimes the non-random selection of sample is also justified. The following sections describe the important types of non-probability sampling.

i. Convenience Sampling: In this method a sample is obtained by selecting “convenient” population elements for example, a sample selected from the readily available sources or list such as telephone directory or a register of the small-scale industry units, etc. will give us a convenient sample. In these, case even if a random approach is used for identifying the units, the scheme will not be consider as simple random sampling For exampe, if one studies the wage stucture in a close by as textile industry by interviewing a few selected workers, then the method adopted here is convinient sampling. The result obtained by conveniencing sampling method can hardly be said to be representative of the population parameter therefore the results oblained are generally unsatisfactory. However, convenient sampling is used for making pilot studies.

ii. Quota Sampling: In this methods of sampling the basic parameters which describe the population are identified first, then the sample is selected which conform to these parameters. Thus, in a quota sample, quotas are fixed according to the parameters, and each field investigator is assigned with quotas of the number of units to be interviewed. Within the preassigned quotas the selection of the samples elements depends on the personal judgment. Quota sampling method is generally used in public opinion studies.

iii. Judgment Sampling: Judgment sampling method can also be called as sampling by opinion. In this method, some one who is well acquainted with the population decides which member in his/her judgment would constitute a proper cross-section representing the parameters of relevance to the study. This method of sampling is generally used in studies involving performance of personnel.

Summary: The sampling theory is a study of relationship that exists between the population and the sample drawn from the population. There are two basic concept about sampling i. probability sampling and ii. Non-probability sampling. The probability sampling includes the following methods: i. simple randon sampling, stratified sampling, cluster sampling multistage sampling, area sampling, multiphase sampling, systemic sampling ii. The non-probability sampling contain, convenied sampling, quota sampling and judgment sampling.

Self-Assessment Question

1. Do you agree with the following statements? Answer Yes or No.

- (i) The sample is the small part of the population.
- (ii) Sampling helps us to get as much more information as possible of the whole population.
- (iii) Sampling is not essential to draw inferences for the population on the basis of sample information.
- (iv) The principles of sampling are validity, efficiency and minimum cost.
- (v) Optimization ensures that a given level of efficiency will be reach with minimum cost.
- (vi) The maximum possible efficiency will not be allowed with a given level of cost.
- (vii) The size of the sample may be 5% of the size of the population.
- (viii) Probability sampling has a known non-zero probability of being selected in the population.
- (ix) An ideal probability sampling, the inference to the population cannot is made entirely by the statistical methods.
- (x) Simple random sampling without replacement is not simple random sampling.
- (xi) Stratified sampling, the population is divided into strata before the sample is drawn.
- (xii) Cluster sampling is frequently used in social surveys in order to cut down on the cost of gathering data.
- (xiii) When the sample is taken that associated with geographical area are called area sampling.
- (xiv) In a quota sample, quotas are not fixed according to the parameters.

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The total error is a function
- (ii) In stratified sampling we divided our population in homogenous groups.

3. Fill Up

- (i) _____ is measured by the inverse of the sampling variance of the estimator.
- (ii) There are two basic principles for the sampling (a) _____
(b) _____

Answer:

True/False

- (i) T (ii) T

Fill up

- (i) Efficiency (ii)-a validity (b) optimization

Lesson 2: Sampling and Non - Sampling Error

Lesson Objectives:

After studying this lesson, you will be able to explain

- Define sampling error;
- Define non-sampling error;
- Determine the sample size;
- Explain the relationship between sampling error and non-sampling error.

Introduction

In choosing the sampling methods, one must be able to identify the degree of efficiency the sampling method should possess in order to generalize of the results on the population. A study of various sampling methods would reveal which method provides the desired degree of efficiency at minimum cost. In any sampling method, it is not to be expected that the value of the sample statistic would be exactly equal to the population parameter that is being estimated. The observed difference between the value of the population parameter and the corresponding value of a sample statistics can be attributed to the combined influence of sampling and non-sampling error.

Sampling Error: The term "Error" refers to the difference between the value of a "Statistic" and that of corresponding "parameter". Sampling error can be defined as the difference that can be assigned to the fact that the results have been obtained on the basis of the sample rather than the complete universe. The magnitude of the sampling error differs from one sampling method to another, even for the same sample size. For example, the expected sampling error associated with a simple random sampling will be greater than the expected sampling error associated with the stratified random sample of the same size, given that the stratification is meaningful.

Sampling error can be defined as the difference that can be assigned to the fact that the results have been obtained on the basis of the sample rather than the complete universe.

Causes for Sampling Error: There are two types of sampling error: a) Biased error b) Unbiased error

- a. Biased Error: These errors arise from any bias in selection, estimation etc.
- b. Unbiased Error: These errors arise due to chance differences between the members of population included in the sample and those that are not included.

Non-Sampling Error: Non-sampling error can be defined as the difference between the true value of the population parameter and the actual value of the sample statistic that can be specifically assigned to carelessness in the design of the sampling scheme, carelessness in data collection, carelessness in data tabulation and analysis etc.

Non-sampling error can be defined as the difference between the true value of the population parameter and the actual value of the sample statistic.

Causes for Non-Sampling Error: Non-sampling errors may arise from the following factors:

- i. Data specification being inadequate and inconsistent with respect to the objectives.
- ii. Lack of trained and experienced investigator.
- iii. Error due to non-responses.
- v. Lack of adequate inspection.
- vi. Inadequate scrutiny of the database.
- vii. Errors in data processing operation.
- viii. Omission of units due to imprecise definition.

The above are the main factors but may not be complete list of arising non-sampling errors.

The total error is thus a function of two independent sources of error and can not be substantially reduced unless both types are simultaneously controlled.

Relation Between Sampling and Non-Sampling Error: In designing a sampling study, the objective must be to constrain the degree of error in the estimate to the levels specified by the situation in which estimate is to be used. This results to inspecting a sampling scheme and sample size such that the combined total of expected sampling error and non-sampling error is within a specified limit. The total error is thus a function of two independent sources of error and cannot be substantially reduced unless both types are simultaneously controlled. The relation between sampling error and non-sampling error can be shown in the following figure.

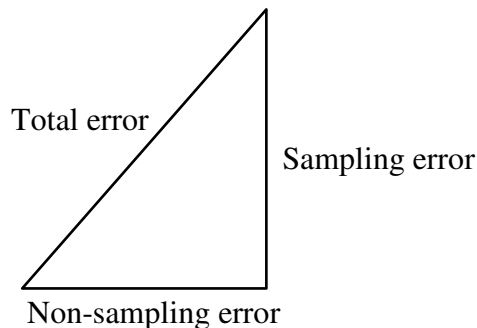


Fig. 12.4: Relationship between total error, sampling error and non-sampling error.

If non-sampling such as response or interviewing error are large, there is no point in taking a huge sample in order to reduce the standard error of the estimate since total error will be primarily determined by the length of the base of the triangle. Like wise, if one is willing to go to great pains to reduce non-sampling errors to a minimum, it will be foolish to make use of a small sample, thereby having a large sampling error. A proper balance between sampling and non-sampling error should be maintained.

Size of the Sample: The size of the sample is defined as the area or number of unity in the sample which are taken from the population. Size of the sample is determined for reducing cost of data collection without cost of some useful information about the population. For a high level of precision, we need to take a large sample. How large should be the sample and what should be the level of precision? In specifying a sample size, attention should be given such that neither so few are selected so as to render the risk of sampling error intotally large, nor too many units are included, which should raised the cost of the study to make it inefficient. It is therefore necessary to make a trade-off between increasing sample size, which should be reduced the sampling error but increase the cost, and decreasing the sample size, which might increase the sampling error while decreasing the cost. Therefore, one has to make a compromise between/obtaining data with greater precision and with that of lower cost of data collection. For determining the sample size, we can use the following relationship-

Size of the sample are determined for reducing cost of data collection without cost of some useful information about the population.

$$\text{Standard error (estimated)} = \sigma_x = \frac{\sigma}{\sqrt{n}}$$

σ = standard deviation of the population

n = number of observations in the sample

If we know the uper and lower limit = Confidence Limit (CL) then;

$C.L = Z_{.05} \frac{\sigma}{\sqrt{n}}$; Where z is the value of the normal variaty at 5% level of significance.

$$\text{or, } \sqrt{n} = z_{.05} \sigma / C.L.$$

$$\text{or, } n = [z_{.05} \sigma / CL.]^2$$

i. e. the sixe of the sample,

$$n = \left[\frac{z \text{ value at 5\% level of significance} \times \text{standard deviation of the upper and lower confidence limit}}{\text{}} \right]^2$$

Example: A departmental store is performing a survey to determine the annual salary earned by managers numbering 5000 in the defferent sectors. How large a sample size it should take in order to estimate mean annual earning withen ± 1000 ? When the standard deviation is $\sigma = 3000$.

Answer: Since the desired upper and lower confidence limit

$$CL = \pm 1,000$$

$$\sigma = 3,000$$

$$Z_{.05} = 1.96 \quad (\text{value from } z \text{ able at 5\% level of signifiance then } n=?)$$

$$\sqrt{n} = [Z_{0.05} \sigma / C.L.]$$

$$\therefore n = [Z_{0.05} \sigma / C.L.]^2$$

$$= \left[\frac{5880}{1000} \right]^2$$

$$= [5.88]^2$$

$$= 34.5744$$

$\therefore$ The desired sample size is $n = 35$

Sample Size in Proportion: If the proportion of the success denoted by P, then the standard error can be defined as:

$$\sigma_p = \sqrt{\frac{PQ}{n}}; \quad P = \text{probability of success}$$

Q = probability of failure

n = number of the sample.

Then -

confidence limit,

$$C.L. = Z_{0.05} \cdot \sigma_p$$

$$\text{or, } C.L. = Z_{0.05} \times \sqrt{\frac{PQ}{n}}$$

$$\text{or, } \sqrt{n} \times C.L. = Z_{0.05} \times \sqrt{PQ}$$

$$\text{or, } \sqrt{n} = \frac{Z_{0.05} \times \sqrt{PQ}}{C.L.}$$

$$\text{or, } n =$$

Example: Suppose, P is estimated as 60% and $\alpha = .05$ level of significance. If the allowable error in estimating the population proportion is not to be greater than 2%, calculate the sample size.

Answer: Since $Z_{\alpha} = .05 = 1.96$; Tabulated value at 5% level of significance.

$$P = \frac{60}{100} = .60 \quad \therefore Q = 1 - P = 1 - .60 = .40$$

$$\text{and } C.L. = \frac{2}{100} = .02$$

$$\text{then } n = ?$$

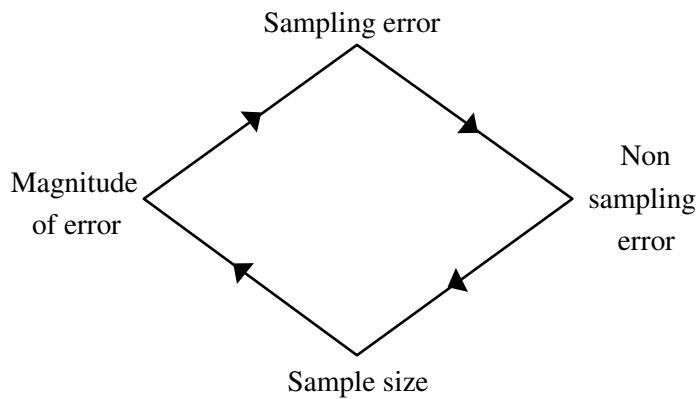
Now—

$$\begin{aligned}
 n &= \left[\frac{Z_{.05} \times \sqrt{PQ}}{C.L} \right]^2 \\
 &= \left[\frac{1.96 \times \sqrt{.60 \times .40}}{.02} \right]^2 \\
 &= \left[\frac{.96019}{.02} \right]^2 \\
 &= [48.009998]^2 \\
 &= 2304.9
 \end{aligned}$$

∴ The sample size $n = 2305$.

Relationship Among Sampling Error, Non-Sampling Error and Sample Size

Relationship among sampling error, non-sampling error and sample size is seen in practical simulations that non-sampling error increases with the increase in sample size whereas sampling error decreases with increase in sample size, keeping in view these relations, a suitable size which gives the minimum value of both types of errors should be taken and can be seen as under fig.-12.5



Non- sampling error increases with the increase in sample size whereas sampling error decreases with increase in sample size.

Fig.12.5: Showing the relationship among sampling error, non-sampling error and sample size.

Self-Assessment Questions

1. Do you agree with the following statements? Answer Yes or No

- (i) The magnitude of the sampling error deffer from one sample to another, even for the sample size.
- (ii) Non-sampling error occurs only carelessness in data collection.
- (iii) The sample size, $n = [z_{0.05}\sigma/cl]^2$
- (iv) If the proportion of the success denoted by P, then, the size of sample is given by $n = \left[\frac{z_{0.05}\sqrt{PQ}}{CL} \right]^2$; Q = Probability of failure.
- (v) Sampling error is decrease with increase in sample size.
- (vi) Minimum error = Sampling error + Non-sampling error.
- (vii) The difference between the value of a statistic and the population value is called sampling error.
- (viii) The sampling error decreases as the size of the sample increases and is zero if sample size is equal to the population size.

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) Non sampling error can be defined as the difference between the true value of the population parameter and the actual value of the sample statistics.
- (ii) The total error is a functin of two independent sources of error.

3. Fill up

- (i) Non sampling error _____ with the increase in _____.
- (ii) _____ refers to the difference between the value of a statistic.

Answer

True/False

- (i) T (ii) T

Fill up

- (i) Increase, sample size (ii) Error

Lesson 3: Sampling Distribution

Lesson Objectives:

After completing this lesson, you will be able to explain

- Sampling distribution;
- Sampling distribution of mean;
- Sampling distribution of the difference between two means;
- Sampling distribution of proportion;
- Sampling distribution of the difference between two proportion.

Introduction

In sampling theory, all possible sample size N that can be drawn from a given population, the statistic that will vary from sample to sample and obtain a distribution called sampling distribution. If the particular statistic used in the sample is the mean, then the distribution called the sampling distribution of mean. In the same way we will have sampling distribution of variance, standard deviation, proportion etc.

Sampling Distribution

When all possible sample of size N that can be drawn either with or without replacement from a given population, the statistic [such mean, median, variance, standard deviation (s.d.) and proportion] that vary from sample to sample and obtained a distribution of the statistic called its sampling distribution.

For each sampling distribution, we compute mean, standard deviation (s.d.). We have the following distribution:

- i. Distribution of mean
- ii. Distribution of difference between two mean.
- iii. Distribution of proportion.
- iv. Distribution of difference between two proportion.
- v. Distribution of standard deviation.

Distribution of Mean

Let us consider that all possible sample of size N (i) with replacement (ii) without replacement from a finite population of size $N_p > N$. If the mean and the standard deviation (s.d.) of the sampling distribution are respectively denoted as $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$, and the population mean and standard deviation are μ and σ then,

a. $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} \sqrt{\frac{N_p - N}{N_p - 1}}$; in the case of without replacement; and

b. $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$; in the case of with replacement.

When all possible sample of size N that can be drawn either with or without replacement from a given population, the statistic that vary from sample to sample and obtained a distribution of the statistic called its sampling distribution.

If a population distribution is normal, the sampling distribution of the mean is also normal for samples of all sizes.

For large sample i.e. $N > 30$, the sampling distribution of mean is approximately a normal distribution with mean μ_x and standard deviation (s.d) σ_x , irrespective of the population.

- ii. Sampling distribution of the difference between means: Consider that there are two populations with sample size N_1 and N_2 drawn from these populations whose mean and standard deviation are respectively - mean = μ_{x1} and standard deviation = s_{x1} from the 1st sample, and mean = μ_{x2} and standard deviation s_{x2} from the 2nd sample, then all possible combinations of these samples from the two populations can be obtained by the differences between means, called the sampling distribution of difference between means. Then the difference of the sample means $\bar{x}_1 - \bar{x}_2$ and this can also be presented as: $\mu_{\bar{x}_1} - \mu_{\bar{x}_2}$ which also equals to the difference between the population means. We can express the same as: $\mu_{x_1} - \mu_{x_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$

Again, the standard deviation of the sampling distribution can be

presented as
$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This result also holds for finite population if the sampling is with replacement and without replacement.

- iii. **Sampling Distribution of Proportion:** Consider that a population is infinite and the probability of success is P and the probability of failure is $q = 1 - p$. The all possible samples of size N drawn from the population, and for each sample determined the proportion P of success, then a sampling distribution of proportion whose mean μ_p and standard deviation σ_p are given by

$$\mu_p = \mu \text{ and } \sigma_p = \sqrt{\frac{Pq}{N}} = \sqrt{\frac{P(1-P)}{N}}$$

For large sample, the sampling distribution is very close to normal distribution.

- i. If sample is taken with replacement then $\mu_p = \mu$ and $\sigma_p = \sqrt{\frac{P(1-P)}{N}}$
- ii. If sample is taken without replacement then $\mu_p = P$ and $\sigma_p = \frac{\sqrt{Pq}}{\sqrt{N}} \times \sqrt{\frac{Np-N}{Np-1}}$

iv. **Sampling Distribution of Differences Between Proportion:**

Consider that two finite populations with the respective probability of success are P_1, P_2 and the probability of failure are $q_1=(1-P_1), \frac{P}{2} = (1-q_2)$ for both. The all possible case the sample size N_1 and N_2 of the two populations and for each cases the sample determines the difference of two proportion whose mean, μ_{p_1}, μ_{p_2} and standard deviation $\sigma_{x_1}, \sigma_{p_2}$ receptively. then sampling distribution of deference between two proportion can be given—

$$\mu_{p_1-p_2} = \mu_{p_1} - \mu_{p_2} = P_1 - P_2 \text{ and}$$

$$\sigma_{p_1-p_2} = \sqrt{\sigma_{p_1}^2 + \sigma_{p_2}^2} = \sqrt{\frac{P_1 q_1}{N_1} + \frac{P_2 q_2}{N_2}}$$

If [N_1 and $N_2 > 30$], then the sampling distribution of difference of proportion is very close to normal distribution.

Distribution of Standard Deviation

When the standard deviation of the population is unknow then standard deviation σ must be estimated by the sample standard deviation.

For large n , the standard deviation of random sample is aproximately distributed as normal with standard deviation $\sigma/\sqrt{2n}$ which is known as the standard error of the standard deviation and is denoted as

$$S = \sigma/\sqrt{2n}; \text{ where } S = \text{Standard error of the standard deviation.}$$

Example: If a coin tossed is eight times in unbiasedly and get a head. Consider the probability of success is 0.50. in any trial what is the probability that the number of success is less than or equal to 5.0.

Salution: Given that - probability of success is .05.

then, the mean, $\mu = np = 8 \times 0.50 = 4$; wher $n=8$ and

$$\text{Standard error, } \sigma = \sqrt{npq} = \sqrt{8 \times 0.50 \times 0.50} = 1.4142 \text{ then}$$

$$z = \frac{x - \mu}{\sigma} \Rightarrow z = \frac{5 - 4}{\sqrt{8 \times 0.50 \times 0.50}} = \frac{5 - 4}{1.4142} = .707$$

From the table, the corresponding value of z is .758, ie. $\Pr[z_{0.5}] = .758$

Some mean and standerd deviation of some sampling distributions are given in the following table:

When the standard deviation of the population is unknow then standard deviation, r must be estimated by the sample standard deviation.

Table-12.3.1; Mean and standard deviation some sampling distributions.

Sampling distribution	mean	Standard deviation
mean	$\mu_x = \mu_1$, the population mean, in all cases	$\sigma_x = \frac{\sigma}{\sqrt{N}}$; this is true for large and small samples
proportion	$\mu_p = \mu$; in all cases	$\sigma_p = \sqrt{\frac{P(1-P)}{N}} = \sqrt{\frac{PN}{N}}$
Standard deviation	$\mu_2 = a_2$ and $\mu_4 = 364$	$\sigma_3 = \frac{\sigma}{\sqrt{2N}}$; for large sample $\sigma_3 = \sqrt{\frac{\mu_4 - \mu_2^2}{4N\mu_2}}$; for approximately normal
median	$\mu_{med} = \mu$; for large sample	$\sigma_{med} = \sigma \sqrt{\frac{\pi}{2N}} = \frac{1.2533\sigma}{\sqrt{N}}$
First and 3rd.	μ_{q1}, μ_{q3} , quartile	$sq_1 = sq_3 = \frac{1.36260}{\sqrt{N}}$
variances	$\mu_{s2} = \sigma^2(N-1)/N$; for large N	$\sigma_{s2} = \sigma^2 \sqrt{\frac{2}{N}}$ and population is normal $\sigma_{s2} = \sqrt{\frac{\mu_4 - \mu_2^2}{N}}$; population are not normal
Co-efficient of Variation	$v = \frac{\sigma}{\mu}$	$\sigma_v = \frac{v}{\sqrt{2N}} \sqrt{1+2v^2}$

Activity :

The mean length of life of a electric bulb is 21.5 hours with a standard deviation of 1.5 hours. What is the probability that a simple random sample of size 50 drawn from this population will have a mean of between 30.5 hours and 45.5 hours?

Self-Assessment Question

1. Do you agree with the following statements? Answer yes or no.

- (i) If the mean of the sample distribution is denoted as μ then-

$$\mu = \frac{\text{sum of observations}}{\text{total number of observations}}$$

- (ii) The standard deviation of the sample distribution $\sigma_x = \sigma/\sqrt{n}$
 $\sqrt{\frac{Np-N}{Np-1}}$; in the case of with replacement.

- (iii) The standard deviation of the sample distribution, then $\sigma_x =$
 $\frac{\sigma}{\sqrt{n}}$; in the case of with replacement.

- (iv) For large sample, the sampling distribution is very close to the normal distribution if samples are taken with replacement then

$$\sigma_p = \sqrt{\frac{P(1-p)}{N}} \text{ and } \mu_p = \mu$$

- (v) For large sample, the sampling distribution of difference of are not very close to the normal distribution.

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) For large sample, the sample distribution is very close to normal distribution.

- (ii) If the mean of the sample distribution is denoted as μ then
 $\mu = \sqrt{\frac{\text{Sum of observation}}{\text{number of observation}}}$

Fill up

- (i) Coefficient of variation, $\sigma_v =$ _____ of sample distribution.
 (ii) Distribution of standard deviation $S =$ _____

Answer

True/False

- (i) T (ii) F

Fill up

- (i) $\sqrt{\frac{\sigma}{x}}$ (ii) $\frac{\sigma}{\sqrt{2n}}$

Lesson 4: Statistical Quality Control

Lesson Objectives:

After studying this lesson, you will be able to explain

- What is meant by quality control;
- How to construct quality control chart;
- Different types of control chart;
- The different types of problems regarding quality control.

Introduction

In this era of ever-growing competition, it has become absolutely necessary for businessman to keep a continuous watch over the quality of goods produced. As we know that quality means conformance to requirements. So, it is considered a very important function to management in the overall value delivery system of the organization.

Quality means an attribute of the product that determines its fitness for use.

Quality: Quality means the concept of consistency, reliability and lack of errors and defectiveness. In other words, quality is an attribute of the product that determined its fitness for use. In business sector quality of goods, production of goods etc. are frequently use. Quality ensure the degree of goodness of the goods.

Quality Control: Quality Control is a powerful productivity device for effectively identified the nature of a problem or a lack of quality in any of the materials, processes and end products. Quality Control ensures the whole processes to get the qualitative goods.

Statistical Quality Control (SQC): Statistical quality control means planned collection and effective use of data for studying causes of variations in quality either as between process, procedures, materials, machines etc. over periods of time.

Control Chart: A control chart is a statistical device principally used for the study and control of respective process. Dr. W.A. Shewart (1920) was the originator of the control chart. He said “Control Chart is essentially a graphic device for presenting data”.

A control Chart consists of three horizontal line. The lines are as followed.

Contruccion of Control Chart: A control chart consists of three horizontal line. The lines are as follows.

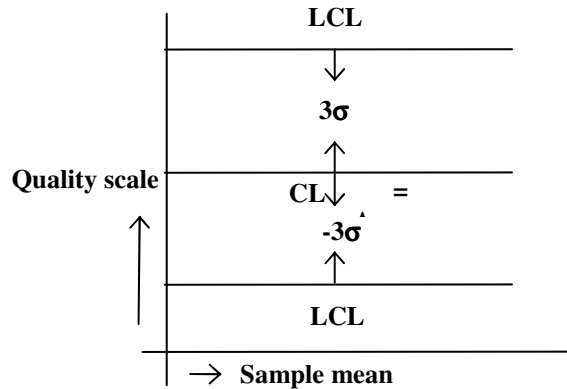
Central Line (CL): A control line is to indicate the desired standard or level of the process.

Upper Control Limit (UCL): An upper control limit indicates the maximun or upper level of the desired process.

Lower Control Limit (LCL): A lower control limit indicates the lower level of the desired process.

The three control lines are as follows:

The different levels are shown in the following graph:



There are two types of control chart,

- i. Control Chart for variables ($\bar{X}$ =mean chart, σ =S.D. Chart and R=Range chart): Control chart for variables are those quality characteristics of a product which are measurable and can be expressed in a specific unit of measurements the different types of variable control chart are as follows:
 - (a) $\bar{X}$ Chart : Control Chart for mean
 - (b) S.D Chart : Control Chart for S.D
 - (c) R-Chart : Control Chart for Range
- ii. Control chart for attributes (p-chart and c-chart).
 Control chart for attributes are those product characteristics which are not amenable to measurement. Such characteristics can only be identified by their presence or absence from the product. The different types of attribute chart are as follows:
 - (a) p-chart: control chart for fraction defective
 - (b) Control chart for the number of defects per unit

$\bar{X}$ - Chart (mean chart) :

X-Chart is used when production process is related to machine setting:

Let $\bar{X}_{ij}$; j = 1,2, ...n be the measurements on the ith sample (i=1,2, ... k), the mean $\bar{X}_i$ and Range R_i can be estimated through the following steps:

Step-1: Find out the sample mean of the data.

Step-2: Find out the mean of the sample mean i.e.

$$\bar{\bar{X}} = \frac{\sum \bar{X}_i}{\text{No. of the samples}}$$

Step-3: Find out the range of the sample, R_i

Step-4: Find out the mean of the Ranges $\bar{R}$

Step-5: Calculate the following:

$$CL = \bar{\bar{X}}$$

$$UCL = \bar{\bar{X}} + A_2 \bar{R} = \bar{\bar{X}} + \frac{3\bar{R}}{d_2 \sqrt{n}}; A_2 \text{ and } d_2 \text{ are tabulated value}$$

$$LCL = \bar{\bar{X}} - A_2 \bar{R} = \bar{\bar{X}} - \frac{3\bar{R}}{d_2 \sqrt{n}}; A_2 \text{ and } d_2 \text{ are tabulated value}$$

Example: A Bicycle manufacturing company construct a control chart for mean and the range for the following data on the basis of fuses, samples of 5 being taken every hours. The data are given as:

18	42	15	42	69	19	64	36	61	42	60	51
20	65	30	45	109	24	90	54	78	51	60	74
27	75	39	68	113	80	93	69	94	57	72	75
42	78	62	72	118	81	109	77	109	59	95	78
60	87	84	90	153	81	112	84	136	78	138	132

Show-that the production process is under control or not?

Solution: Step: 1 Arranging the data as given in the following table:

Table- Arranging the given data

Sample No.	Sample data	Total	Sample mean $\bar{X}_i$	Sample range R_i
1	18 20 27 42 60	167	33.4	42
2	42 65 75 78 87	347	69.4	45
3	15 30 39 62 84	230	46.0	69
4	42 45 68 72 90	317	63.4	48
5	69 109 113 118 153	562	112.4	84
6	19 24 80 81 81	285	57.0	62
7	64 90 93 109 112	468	93.6	48
8	36 54 69 77 84	320	64.0	48
9	61 78 94 109 136	478	95.6	75
10	42 51 57 59 78	287	57.4	36
11	60 60 72 95 138	425	85.0	78
12	51 74 75 78 132	410	82.0	81
			$\sum \bar{X}_i = 859.2$	$\bar{R} = 59.67$

From the above, we have

$$\bar{\bar{X}} = \frac{1}{12} \sum_{i=1}^{12} \bar{X}_i = \frac{859.2}{12} = 71.60$$

$$\bar{R} = \frac{1}{12} \sum_{i=1}^{12} R_i = \frac{716}{12} = 59.67$$

$$\therefore \bar{\bar{X}} - \text{Chart: } CL = \bar{\bar{X}} = 71.60$$

$$UCL = \bar{\bar{X}} + A_2 \bar{R} = 71.6 + 0.58 \times 59.67 = 106.21$$

$$LCL = \bar{\bar{X}} - 0.58 \times 59.67 = 36.99$$

From the table for n=5

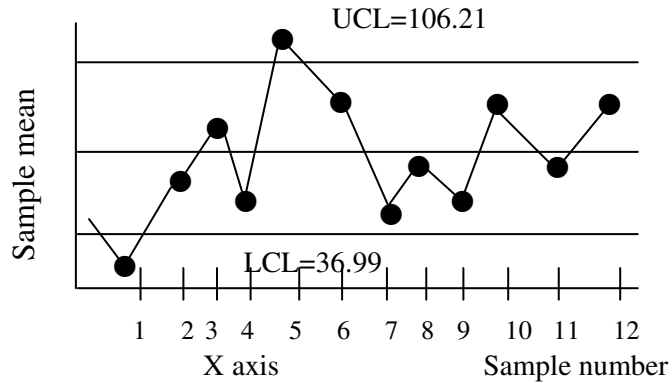
$$\therefore A_2 = 0.58$$

$$D_3 = 0$$

$$D_4 = 2.11$$

Now, Putting the Values on graph paper:

Fig.



From the $\bar{X}$ – Chart, two point 1st and 5th Sample is out of control but rest of the sample is under control. So we can say the process is not exactly under control.

Activity:

Bangladesh Metal Tools factory uses an extraction process to produce various kind of aluminium brackets.

The data are given as follows:

Hour	Brackets Diameters (mm)			
1	5.03	5.06	4.86	4.90
2	4.97	4.94	5.09	4.78
3	5.02	4.98	4.94	4.95
4	4.92	4.93	4.90	4.92
5	5.01	4.99	4.93	5.06

Comment on the process whether it is under control or not?

R-Chart:

If X_{ij} ; $j = 1, 2, \dots, n$ be the measurements on the i th sample ($i = 1, 2, \dots, k$). Then

$$\bar{X}_i = \frac{\sum X_{ij}}{n}; i = 1, 2, \dots, k; j = 1, 2, \dots, n$$

$$\bar{\bar{X}} = \frac{\sum \bar{X}_i}{K}; i = 1, 2, \dots, k$$

$$R = \frac{\sum R_i}{K}$$

Then $\bar{R}$ chart can be obtained, as follows:

$$CL = \bar{R}$$

$$UCL = \bar{R} + \frac{3d_3 \bar{R}}{d_2} = \bar{R} \left[1 + \frac{3d_3}{d_2} \right] [\because d_3, d_2 \text{ are the tabulated value at } n \text{ sample size.}]$$

$$LCL = \bar{R} - \frac{3d_3\bar{R}}{d_2} = \bar{R} \left[1 - \frac{3d_3}{d_2} \right]$$

Which, can be simply shown as:

$$CL = \bar{R}$$

$$UCL = \bar{R}D_4$$

$$LCL = \bar{R}D_3 \text{ [} D_4, D_3 \text{ are also tabulated value at } n \text{ sample size.]}$$

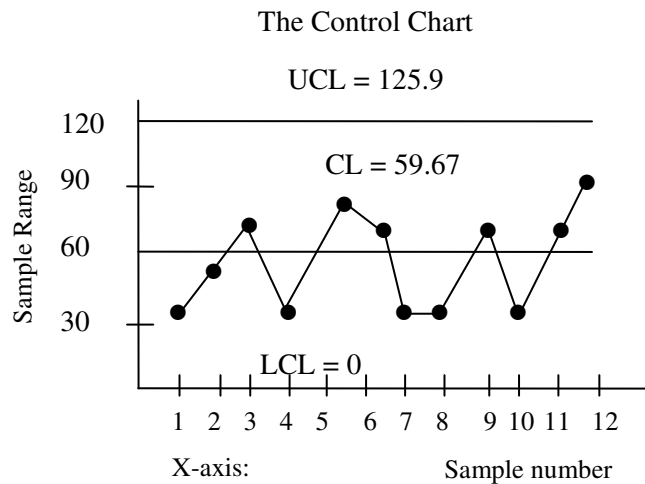
Example: From example $\bar{X}$ chart

Solution: From the table A, for $n = 5$

$$CL = 59.67,$$

$$UCL = 2.11 \times 59.67 = 125.904$$

$$LCL = 0 \times 59.57 = 0 \text{ [} D_4 = 2.11, D_3 = 0 \text{]}$$



The above chart shows that the process is with in control.

Activity:

For each of the following cases, find out the CL, UCL and LCL for an R-Chart based on given information.

(a) $n=3, \bar{X} = 18.4, \bar{R} = 3.1$

(b) $n=24, \bar{X} = 8.6, \bar{R} = 1.4$

(c) $\bar{R} = 6.0, LCL = 3.0$ Find the UCL when $n=3$.

Self-Assessment Questions:

Multiple-Choice Question:

1. Find out the right answer:

- (i) Who was the originator of control chart
(a) R. A. Fisher (b) D'Morgan
(c) W. A. Shewart (d) G. M. Shaha

2. Write "T" if the statement is true and "F" if the statement is false:

- (i) A control chart is a statistical device for industrial quality control.
(ii) UCL indicate the lower level of the desired process.
(iii) $\bar{X}$ - Chart is known as mean chart.
(iv) P-Chart usually used in case of fraction defective.

Fill up

- (i) Quality ensure the _____ of _____ of the goods.
(ii) Quality Control ensure the whole _____ to get the qualitative goods.
(iii) For $\bar{X}$ - Chart, $CL = \bar{\bar{X}} + \text{-----}$
(iv) Define the following
(i) SQC (ii) UCL (iii) LCL (iv) CC.

Answer:

Multiple Choice Questions:

- (i)- c

True/False

- (i) T (ii) F (iii) T (iv) T

Fill up

- (i) degree, goodness
(ii) whole process, get the qualitative
(iii) $\frac{3R}{d_2\sqrt{n}}$

Lesson 5: Acceptance Sampling by Attribute and Sampling Plan

Lesson Objectives:

After completing this lesson, you will be able to explain:

- What is acceptance sampling by attribute;
- What is sampling plan;
- Uses of the acceptance plan.

Introduction

In many manufacturing process, the producer in order to ensure that the manufactured goods are according to specification of the customer, gets his lot checked at initial strategic, an ideal way of doing this seems to inspect each and every item presented for acceptance. But it is impossible because the process is destructive. From practical and economic considerations sampling procedure are adopted on the basis of the samples drawn out randomly from the lot for making decision accepted or rejected after the inspection. It has been found that if a scientifically designed sampling inspection plan is used, it provides adequate protection to producer as well as consumer very economically.

It has been found that if a scientifically designed sampling inspection plan is used, it provides adequate protection to producer as well as consumer very economically.

Acceptance Sampling by Attribute: From economic consideration, it is not practicable to inspect fully in lot control; one has to take resource to sampling inspecting to accept or reject the lot. This means that the items are judged on some character of the goods named attribute i.e. good or defective by inspection and the quality of the lot is judged from the sample fraction defective. The process is taken on the basis of the samples drawn at random of the goods known as acceptance sampling by attributes for example, If testing is destructive, as for instance is the case of crackers, shells, bulb etc., it is absolutely non-sensical to talk of 100% inspection. Even in those cases where 100% inspection is possible it may not be desirable; it is costly and due to fatigue, impossibility of proper check and variations in efficiencies of inspection in time, person and place.

The process is taken on the basis of the samples drawn at random of the goods known as acceptance sampling by attributes

Acceptance Quality Level (AQL): A lot with relatively small fraction defective i.e., sufficiently good quality, say P_1 that we do not want to reject more often than a small proportion of time is sometimes referred to as a good lot. Usually,

Prob. [Rejecting a lot of quality P_1] = .05

Which implies, Prob. [Accepting a lot of quality P_1] = .95

Where P_1 is known as an acceptance quality level (AQL) and a lot of this quality is considered as satisfactory by the consumer.

Lot Tolerance Proportion or Percentage Defective (LTPD): Lot proportion defective usually denoted by P_d is the lot quality which is considered to be bad by the consumer. The consumer is not willing to accept lots having proportion defective P_d or, greater. $100 P_d$ is called lot

tolerance percentage defective (LTPD). On the other hand, LTPD is the quality level which the consumer regards as rejectable and is usually called rejecting quality level (RQL). A lot of quality P_d known to be accepted some arbitrary and small fraction of time, usually 10 percent.

By consumer we shall mean the person, firm or department that receives the goods from the producer.

Consumer's Risk (C.R) By consumer we mean the person or firm or department that receives the goods from the producer. The consumer has to face the risk of accepting a lot of unsatisfactory quality on the basis of sampling inspection. Let P_d be the lot tolerance percentage defective ie. the maximum fraction defective in the lot that the consumer with tolerate. Then the probability of accepting in a lot with fraction defective P_d under the inspection plan is called consumer risk and which is denoted by P_c and defined as $P_c = \text{Prob} [\text{accepting a lot of quality } P_d] = B$; $B = .10$ [Dodge and roming]

Producer Risk (P.R): By producer be shall mean any person, farm or deparment that produce goods to be supplied to another person or farm or department not any sampling inspection plan for acceptance or, rejection of a lot possess the disadvantage of occasionally rejecting a lot of satisfactory quality. Consideration the producer claims that he has standardized the quality at the level of fraction defective $\bar{P}$, called the producer process average. The probability of rejecting a lot under the sampling inspection plan when the fraction defective is actually $\bar{P}$ is called producer risk and is denoted by P_p . clearly, P_p can be kept small by making $\bar{P}$ sufficiently small. But the producer may find it more economical to allow a fairly sign risk than to try to reduce $\bar{P}$.

P_p can defin d as—

$$P_p = \text{Prob} [\text{rejecting a lot of quality } \bar{P}] = \alpha$$

Process Average Fraction Defective (PAFD): When the quality of any production process tends to settle down to some level which may be expected to be more or less the same for a particular machine. But in practical, the quality of product may suddenly deteriorate the process avarage of any manufactured product is obtained by finding the percentage of defectives in the product over a fairly long time called process avarage fraction defective which is denoted as $\bar{P}$.

The inspection of the rejected lot and replacing the defective pieces found in the rejected lots by the good ones eliminates the number of defective in the lots to a great extent, thus improving the lots quality.

Sampling Plan

In a sampling plan in which a specified quality objective is attained through corrective inspection of rejected lots. The inspection of the rejected lot and replacing the defective pieces found in the rejected lots by the good ones eliminates the number of defective in the lots to a great extent, thus improving the lots quality. This plan is called rectifying inspection plan (RIP)

Most of the rectifying inspection plans for a lot by lot sampling call for 100% inspection of the rejected lot and replacing the defective pieces found by good one. According to Harold. F. Dodge and Harry G. Rowing the two inspection points related to rectifying inspection plan that are as-

- i. **Average Out Going Quality (AOQ):** The average quality of the product after sampling and 100% inspection of rejected lot called average out going quality. The expected fraction defective remaining in the lot after the application of the sampling plan is also called AOQ.
- ii. **Average Total Inspection (ATI):** the average amount of inspection required for rectifying inspection plan, called average total inspection.

Average outgoing quality limit (AOQL); The maximum value of the average out going quality, the maximum being taken with respect to P_L , known as average out going quality limit (AOQL). The average outgoing quality P^* , is the function of the incoming quality P , then—

$$P^* = AOQ = \frac{P(N-n)P_p}{N}$$

where, N = size of the lot
 n = size of the sample
 P_p = probability of acceptance of the lot
 P = incoming quality.

If n is small compared with N , then a good approximation of the outgoing quality is given by—

$$AOQ = PP_p$$

In general if the incoming quality and a rectifying inspection plan is used then the average outgoing quality of the lot will be—

$$\begin{aligned} AOQ &= P \cdot \text{prob} \{ \text{accepting of the lot of quality } p \} + 0 \cdot [1 - \text{prob} \{ \text{accepting of the lot of quality } P \}] \\ &= P \cdot \text{prob} \{ \text{accepting of the lot of quality } P \} \end{aligned}$$

because,

- i. the probability of accepting of the lot of quality and the outgoing of the lot will be approximately same as incoming lot quality P
- ii. the probability of rejection of the lot and the lot is rejected after sampling inspection, the average outgoing quality is zero.

For a given sampling plan the value of average outgoing quality can be plotted for differed value of P to obtained the AOQ curve as in figure-

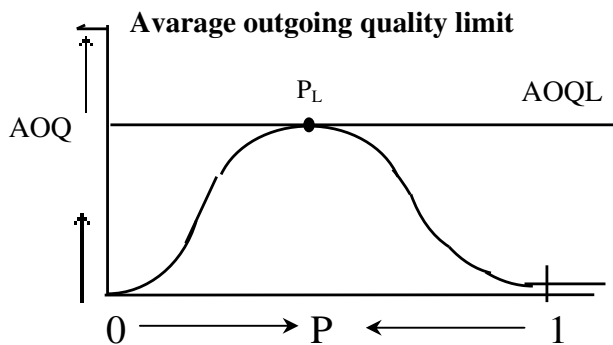


Fig: 12.6: Average outgoing quality limit

Average total inspection (ATI); the average amount of inspection required for the rectifying inspection plan, called average total inspection.

The probability of rejection of the lot and the lot is rejected after sampling inspection, the average outgoing quality is zero.

The incoming quality always lies, $0 < P < 1$, the AOQ will be positive and having a maximum value of the incoming quality.

If, the average outgoing quality limit, denoted as P_L then,

$$AOQL = P_L = P_m \cdot \text{Prob} \left\{ \text{acceptance of the lot of quality of } P \right\} \frac{(N-n)}{N}$$

Where P_m = the maximum value of p , if it is computed from $P = P_m$, then

$$AQQL = P_m \frac{N-n}{N} \text{prob} \left\{ \text{acceptance of the lot of quality of } P \right\}$$

Operating characteristic curve is a graphic representation of the relationship between the probability of acceptance and for the variation in the lot quality.

Operating Characteristic (O.C) Curve: Operating characteristic curve is a graphic representation of the relationship between the probability of acceptance and for the variation in the lot quality.

The OC Curve of an acceptance-sampling plan shows the ability of the plan to distinguish between good and bad lots. For any given fraction defective P in a submitted lot, the OC curve shown in figure 10.5.1 indicate that the probability $P(A/Q)$. Such that a lot will be accepted by given sampling plan

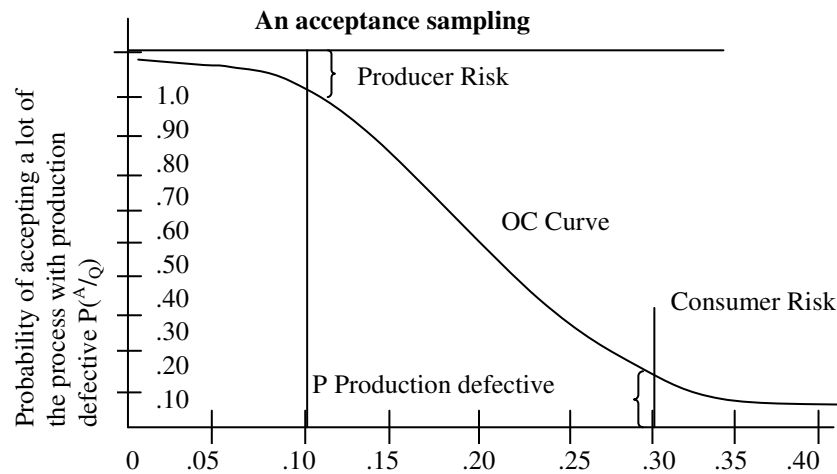


Figure 12.7: OC Curve

Average Sample Number (ASN): The average sample numbers can be defined as the expected value of the sample size required to arrive at to a decision about the acceptance or rejection of the lot in an acceptance-rejection sampling plan. It is a function of the incoming lot quality. On the other hand, the expected number of items inspected per lot to arrive at a decision in an acceptance rejection sampling inspection plan called average amount of total inspection (ATI).

Now, $ATI = ASN + [\text{average size of inspection of the remainder in the rejected lot}]$

If the lot is accepted on the basis of the sampling inspection plan, then

$$ASN = ATI \quad \text{otherwise}$$

$$ASN < ATI$$

In other words, ASN gives the average number of the unit inspected per accepted lot.

For example, in a simple sampling plan is used the number of the item inspected for each lot will be corresponding sample size n lies $ASN = n$.

Example: Suppose same bulbs are packaged 25 to a box and that the following acceptance sampling plan is used for accepting or rejecting boxes of these bulbs:

- A random sample of two bulbs is drawn from the box and the bulbs are tested.
- The box is accepted of both bulb in the sample are good; otherwise the box of bulb is rejected.

Answer: Let, the number of bulb, $N = 25$

the sample size, $n = 2$

the number of no defective bulb, $c = 0$

and then $P(A/\theta) = \text{Prob}[A]$ and $\text{prob}[x]$ are function of θ

Where $P(A/\theta) = \text{Probability of acceptance of a lot of fraction/defective}$.

$\theta = \text{Fraction defective}$

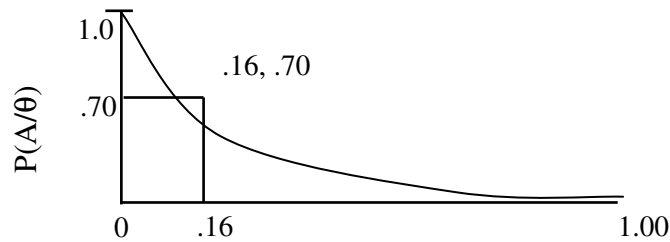
$$\text{i.e. } P(A/\theta) = \frac{\binom{N\theta}{0} \binom{N-N\theta}{2}}{\binom{N}{\theta}}$$

$$= \frac{\binom{250\theta}{0} \binom{25-25\theta}{2}}{\binom{25}{\theta}}$$

Where the possible values of Q are 0.04, .08 ----- then

θ	$P(A/\theta)$	θ	$P(A/\theta)$
0	1.00	.52	.22
.04	.92	.56	.18
.08	.84	.60	.15
.12	.77	.64	.12
.16	.64	.68	.10
.20	.57	.72	.07
.24	.51	.76	.05
.28	.45	.80	.02
.32	.40	.84	.01
.36	.35	.88	.00
.40	.30	.92	.00
.44	.26	.96	.00
.48		1.00	

Putting the value in the graph, then the oe are as follows:



Here, we find that if the box has four defective bulb, the probability of accepting the box of the bulb on the basis of the sampling plan in .70

Activity :

Suppose the following single sampling plan is used for accepting or rejecting large lot of mass-produced samples.

1. Draw a sample size 50 from the lot and inspect the 50 items.
2. Accept the lot if the sample contains not more than one defective other wise reject the lot.

Hints: $n=50$, $c=1$ if the frection defective is other $\mu = 50$.

$$\text{then } P(A/\theta) = \frac{(500)^x e^{-500}}{x!}$$

Self-Assessment Question:

1. Do you agree with the following statements? Answer yes or no.

- (i) The process is taken on the basis of the samples drown at random of the goods known as acceptance sampling.
- (ii) The probability of accepting in a lot with fraction defective is called consumer risk.
- (iii) PR _ prob [rejecting a lot of quatity of lot proportion defective]
- (iv) The avarage outgoing quantity, AOQ = incoming quality X probability of acceptance of the lot.
- (v) Oparating characteristic curve is a graphic representation of the relationship between the propability of acceptance and for the variation in the lot quality.
- (vi) Avarage sample number gives the avarage numbers of the unit inspected per accepted lot.

Lesson 6: Central-Limit Theorem

Lesson objectives:

After studying this lesson, you will be able to explain

- Central limit theorem;
- Uses of the central limit theorem with example.

Introduction:

Most of the statistical theorem implies that the populations are normal i.e.. the sampling distribution of the sample mean is normal if and only if the population is normal. But when the populations are not normal then the process of measurement produced a continuous random variable by applying a theorem called central limit theorem.

Central Limit Theorem

If x_i ; $i = 1, 2, \dots, n$ are independent random variables, with mean μ_i ; $i = 1, 2, \dots, n$ and variance σ_i^2 ; $i = 1, 2, \dots, n$, and $x = a_1x_1 + a_2x_2 + \dots + a_nx_n$, then as n increases, x becomes approximately normal i.e.

$$X \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right); n \rightarrow \infty$$

In other words, if $\bar{x}$ is the mean of the random variable and σ^2/n is the variance then,

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0, 1)$$

The question as to how large n has to be in order to achieve a good approximation. It depends on the shape of the density or probability function of x_i , i.e. the shape of the population. In general if the sample size is 30 or more it will give suitable approximation.

The sampling distribution of x will be approximately normal no matter what the shape of the x_i as long as n is large enough and it is the sampling distribution of x that is important in statistical inference with means and not the population of the x_i .

The application of the central limit theorem:

The central limit theorem can be applied on the following areas:

1. Non-normal population
2. Discrete population and continuous population
3. Estimation with the normal density function.
4. Chi-square sampling distribution
5. T-sampling distribution
6. F-family distribution.

The sampling distribution of x will be approximately normal no matter what the shape of the x_i as long as n is large enough and it is the sampling distribution of x that is important in statistical inference with means and not the population of the x_i .

1. Non-Normal Population

When the variables of the population are not normal then for large $n > 30$, central limit theorem are applied to make the population as approximately normal if $x_1, x_2, \dots, x_n$ are random variable of a population then-

$$z = \frac{X - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Where, z = approximately normal variable

μ = mean of the population

$n = 30$ or more.

2. Discrete Population and Continuous Population

The central limit theorem applies to discrete as well as continuous population. We represent the population of concern by a binomial population (see, fig 12.8)

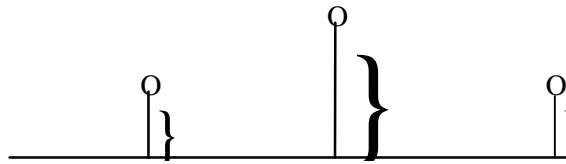


Fig 12.8: Population represent by binomial probability function.

Consider n Bernoulli trials are presented by $x_1, \dots, x_n$ where each x_i has mean p and variance pq . If X presents equal number of success then,

$$X = x_1 + x_2 + \dots + x_n$$

Where x has binomial sampling distribution with mean equal to np and variance npq . However in accordance with the central limit theorem, we can say, for large sample, $X \sim N(np, npq)$. In similar way we can define, $P \sim N(P, pq/n)$

Example : Consider $n = 10$ and $P = 0.5$ we select possible values of x from binomial table. we find-

$$P(x=10) = P(p=1) = \frac{10!}{10!0!} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^0 = \frac{1}{1024} = .0010$$

$$P(x=9) + P(p=\frac{9}{10}) = \frac{10!}{9!1!} \left(\frac{1}{2}\right)^9 \left(\frac{1}{2}\right)^1 = \frac{10}{1024} = .0098$$

$$P(x > 8) = \frac{1}{1024} + \frac{10}{1024} = .0108$$

According to central limit theorem-

$$\bar{X} \sim N(np, npq) \Rightarrow \bar{X} \sim N(10 \times 0.5, 10 \times 0.5 \times 0.5)$$

$$\Rightarrow \bar{X} \sim N(5, 2.5)$$

$$\text{and } P \sim N(0.5, .025)$$

If we let $z \sim 2N \left[z > \frac{x-\mu}{\sqrt{d.n}} \right]$

$$\Rightarrow Z \sim N \left[z > \frac{9-5}{\sqrt{2.5}} \right]$$

$$\Rightarrow z \sim N [z > 2.53]$$

$$\sim .00057$$

Which is better approximation than other method?

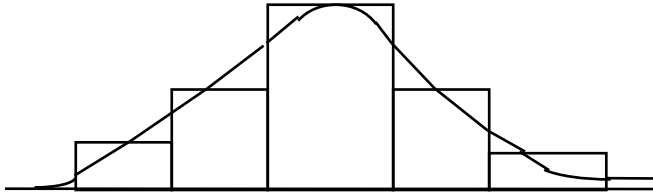


Fig 12.9: Probability function approximated by density function

3. Estimation with Normal Function

In case of estimating the parameter μ and σ^2 , we would measure μ and σ^2 as a measure of variability of normal density family then, central limit theorem is to estimate consistent, sufficient and most efficient unbiased estimator.

4. The Chi-Square Sampling Distribution

Many statistics are useful in the inference process in sampling distribution C of the χ^2 -family. The χ^2 -family is one parameter family with n degrees of freedom representing the parameter. The mean and variance of a χ^2 -density function with n degrees of freedom are n and $2n$ respectively. For large n , χ^2 is approximately and by normal density function.

5. The t-Sampling Distribution

In statistical inference concerning means, both in estimation and hypothesis testing we will frequently assume a t -density function as a sampling distribution. The t -family of density function can be thought of as arising from independent normal and χ^2 -density function. So central limit theorem can easily be applied for t -density as normal.

6. The F-Sampling Distribution

The F -family consists of members that are two parameter density functions which are generated from two χ^2 -density functions so central limit theorem can easily be applied for estimating F -density as a normal.

In case of estimating the parameter μ and σ^2 , we would with to measure μ and σ^2 as a measure of variability of normal density family then, central limit theorem to estimate consistent, sufficient and most efficient unbiased estimator.

Self-Assessment Question

1. Do you agree with the following question statements? Answer yes or no.

- (i) Most of the statistical theorem implies that the population is normal.
- (ii) The sample mean is normal if and only if the population mean is normal.
- (iii) A good approximation depends on the probability function of x_i .
- (iv) The central limit theorem cannot be applied on Non-normal population for large sample.
- (v) Large sample size, $N < 10$.

Exercise

- 1. Describe the different methods of sampling and the requirement of a good sample.
- 2. What is sampling? Explain the importance of sampling in solving business problem.
- 3. Explain the concepts of sampling distribution and standard error? Explain the difference between the sampling error and Non-sampling error.
- 5. Find the mean and variance of the sampling distribution of the sample mean. Distinguish between standard error and standard deviation.
- 6. State the probabilistic and non-probabilistic sampling techniques. Explain stratified random sampling techniques.
- 7. Define acceptance sampling plan. Explain rectifying inspection plan and show that $AOQL = P_m \frac{N-n}{N} \times \text{Prob} [\text{acceptance of the lot of the quality of } P]$
- 8. Define central limit theorem. Describe the uses of central limit theorem with examples.

BUSINESS FORECASTING & TIME SERIES ANALYSIS



When estimates of future conditions are made on a systematic basis, the process is referred as forecasting. Forecasting refers to the analysis of past and present conditions with the object of drawing inferences about probable future business conditions. In fact when a man assumes the responsibility of running business he takes the responsibility for attempting to forecast the future. Forcasts depend on some information regarding time sequences data. Such data are generally referred as time series.

In this unit we discuss about forecasting and time series data and analysis of data.

Lesson 1: Time Series

Lesson Objectives:

After completing this lesson, you will be able to:

- Understand the meaning of time-series and forecasting;
- Define components of time series;
- Describe models of time series;
- Narrate the selection criterion of models of time series.

Introduction

Data collected for the purpose of analysis at a single point of time are called **cross-sectional** data. For example, data for household expenditure on food can be collected in different locations at a single point of time. Alternatively data collected over time are called **time series** data. Generally we use the term time series to refer to any group of statistical information accumulated at regular intervals of time.

Time Series

Forecasting is a systematic procedure to determine the future value of a variable of interest. Forecasting or predicting, is an essential tool in any decision making process. The quality of the forecasting management can make is strongly related to the information that can be extracted and used from past data. Time series analysis is used to detect patterns of change in statistical information over regular intervals of time. We project these patterns to arrive at an estimate for the future. Thus, time-series analysis helps us cope with uncertainty about the future. The distinction between prediction and forecasting is that forecasting generally refers to the scientific methodology that often uses past data along with some well-defined assumptions or 'model' to come up with a 'forecast' of future. In that sense, forecasting is an objective estimate. A prediction is a subjective estimate made by an individual by using his intuitive 'hunch' which may in fact come out true.

Time-series analysis helps us cope with uncertainty about the future.

The future is inherently uncertain and any forecast at best is an educated guess with no guarantee of coming true. In certain purely deterministic systems an unequivocal relationship between cause and effect has been clearly established and it is possible to predict very accurately the course or events in the future, once the future patterns of causes are inferred from past behaviour. Economic systems, however, are more complex because (i) there is a large number of governing factors in a complex structural framework which may not be possible to identify and (ii) the individual factors themselves have a high degree of variability and uncertainty. In spite of these complexities, a forecast has to be made so that the people involved in business can plan for the future.

The future is inherently uncertain and any forecast at best is an educated guess with no guarantee of coming true.

Forecasting for Planning Decisions

The primary purpose of forecasting is to provide valuable information for planning the design and operation of the enterprise. Planning decisions may be classified as long term, medium term and short term.

Time series analysis is one of the methods used for both medium and short term forecasting.

Long-term decisions include decisions like plant expansion or new product introduction, which may require new technologies or a complete transformation in social or moral fabric of society. Such decisions are generally characterized by lack of quantitative information and absence of historical data on which to base the forecast of future events. Intuition and the collected opinion of experts in the field generally play a significant role in developing forecasts for such decisions.

Medium term decisions involve such decisions as planning the production level in a manufacturing plant over the next year. Short-term decisions include daily production planning and scheduling decisions. Time series analysis is one of the methods used for both medium and short term forecasting.

Components of Time Series

When the data are arranged on the basis of time of their occurrence, often they show fluctuations from time to time—from day to day, from week to week, from month to month, and from year to year. A constantly working composite force causes these fluctuations. This composite force has four components, commonly known as the components of a time-series, which are as follows:

- a) Secular trend or long term movements denoted by (T)
- b) Periodic movements or short term fluctuations, which comprise of
 - i) Seasonal fluctuations (S) ii) Cyclical fluctuations (C) and
- c) Irregular or random fluctuations (I)

Secular Trend

Trend is that irreversible movement in a time series, which continues in general, in the same direction over a long period of time. Simply stated, secular trend refers to the general tendency of the time series data to increase or decrease or to be steady over a long period of time. For example, an upward tendency is usually observed in time series relating to population, production and sales of certain products, prices, incomes, etc. On the other hand, a downward tendency can be observed in the time series relating to the rate of infantile mortality, deaths due to epidemics, etc., due to advancement in medical science, improved medical facilities and better sanitation etc. It should be noted that trend refers to only smooth, regular, long-term movement of the data and has nothing to do with sudden and erratic movements either in upward or downward direction.

Uses of Trend

- i) The study of trend enables us to have a general idea about the pattern of behavior of the phenomenon under consideration. This helps in business forecasting and planning future operations.
- ii) By isolating trend values from the given time-series, we can study the short-term and irregular movements.
- iii) The study of trend enables us to compare two or more time series over different periods of time and to draw important conclusions about them.

Secular trend refers to the general tendency of the time series data to increase or decrease over a long period of time.

Trend refers to only smooth, regular, long-term movement of the data and has nothing to do with sudden and erratic movements either in upward or downward direction.

Short-Term Fluctuations or Periodic Movements:

Periodic movements can be classified into two categories, viz., seasonal fluctuations and cyclical fluctuations.

i) **Seasonal Fluctuations:** Seasonal fluctuations are those periodic movements that occur regularly every year and have their origin in the nature of the year itself. Seasonal variations recur in regular and periodic manner over a span of less than a year, i.e., during a period of twelve months and have the same or almost the same pattern year after year. Thus in a time-series data where only annual figures are given, there are no seasonal variations. Seasonal component is denoted by S. It shows the patterns of change within a year that tends to be repeated from year to year. Here the variable moves in regular cycles within a year around the trend line. Note that the term season may be any unit of time, for example days within a week, weeks within a month, months or quarters or seasons within a year. Most of the time series relating to business and economics are influenced by seasonal forces, e.g., time-series relating to agricultural production and sales of agricultural products, bank clearings, bank deposits, sales and profits in a departmental store, etc.

Seasonal variations recur in regular and periodic manner over a span of less than a year

Seasonal variations have two main causes: a) Climate in its widest sense and b) Customs. Climate influences the timing of farmers' incomes. Sales of woollen clothing, umbrellas, etc., have a seasonal movement due to climate. Seasonal fluctuations also occur due to customs, habits, fashions, and conventions of the people in a society. Thus demand for clothes suddenly increases before festivals like Eid, Christmas, Durga Pooja, etc.

Seasonal variations have two main causes: a) Climate in its widest sense and b) Customs.

The main objective of studying seasonal variations is the establishment of the seasonal pattern which helps in planning future operations and formulating policy decision regarding purchase, production, inventory control, personnel requirements, selling and advertisement programs etc., during various parts of the year.

ii) **Cyclical fluctuations:** The wavelike movements in time-series with period of oscillation more than one year are called cyclical variations. Cyclical fluctuations generally exhibit semi-regular periodicity, as they are neither as regular as the seasonal variations nor as accidental as the erratic fluctuations. Most business and economic series exhibit wavelike changes of prosperity and depression. In times of prosperity, production, sales, employment and other economic activities are high; in times of depression, the opposite is true.

Cyclical fluctuations generally exhibit semi-regular periodicity, as they are neither as regular as the seasonal variations nor as accidental as the erratic fluctuations.

A **cyclical component** denoted by C, is given by rather less regular cycle over the medium to longer term around the trend line. The cycle is at least one year and it can be as many as 15 to 20 years. Generally such cycles are related to new technological breakthroughs. Thus a study of the cyclical fluctuations helps business executives in the formulation of policies aimed at stabilizing the level of business activity. It also helps the Government to take such measures as are feasible to prevent the deterioration of mild recession or early crisis into deep depression and keep the swings of prosperity within reasonable limits without developing into storming speculations.

In addition to the influence of long term and short term forces, every time series is subjected to occasional influences, which may occur just once, or several times, but without any pattern or regularity.

Random or Erratic Fluctuations :

In addition to the influence of long term and short term forces, every time series is subjected to occasional influences, which may occur just once, or several times, but without any pattern or regularity. The variations they produce are, therefore, called irregular or random fluctuations. Wars, earthquakes, floods, fires, strikes or lock outs and such other unforeseen or unforeseeable events are typical cause of erratic fluctuations. A random variation may last but a day or may last many months. An **irregular component** denoted by I, is entirely unpredictable. The stock market of 1996 in Bangladesh was dramatic and unexpected departure from trend in terms of share price. In this component the value of the variable changes in a random manner. The impact of nine-eleven in American Air Travel business is another example of random fluctuations.

Different component of time series are shown in figure 13.1

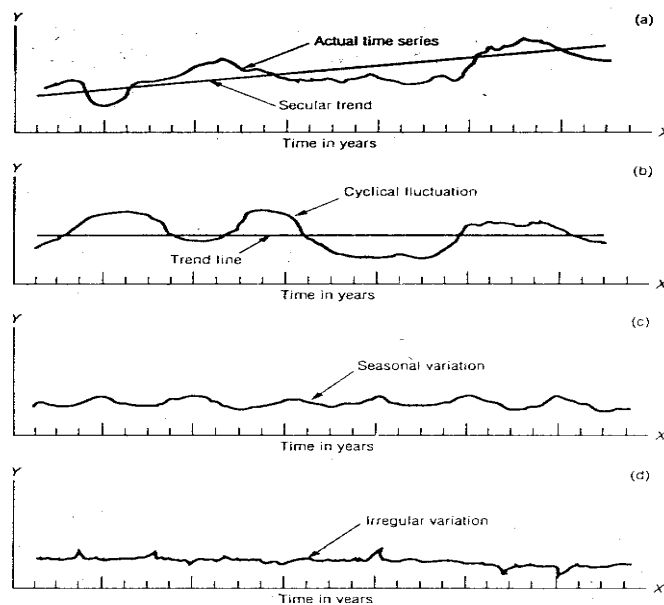


Figure 13.1

By analysis of time series we mean a process by which the effects of the various components of a time-series can be isolated and measured.

Analysis of Time series:

The value of a time series over a period of time changes due to the combined impact of its components. Therefore, it is necessary to isolate and measure the separate effects of these forces in a given time series. The process of analysis of time series aims at achieving this objective. Thus by analysis of time series we mean a process by which the effects of the various components of a time-series can be isolated and measured. The analysis of time series enables us to obtain answers to the following questions:

- i) What would be the size of the variable at different points of time if only long-term forces had affected it?
- ii) How much does the variable change in different parts of the year due to seasonal factors?

- iii) To what extent and in which direction has the cyclical variations pulled the variable?
- iv) What has been the effect of erratic forces?

Importance of Time-Series Analysis

Time-series analysis is of great importance and extremely useful in decision making due to the following reasons:

- i) It helps in understanding the past behaviour of a variable and thereby in predicting the future behaviour.
- ii) It helps in planning future operations. By studying the various components of time-series, a business executive can make intelligent choices regarding capital investment, production, sales and inventory etc.
- iii) It helps in evaluating current accomplishments. The performance can be compared with the expected performance and the reasons for variations analyzed.
- iv) It facilitates comparison between different time-series.

Models for Time Series

Mathematical Models for Time-Series Analysis:

There are two mathematical models that are commonly used for the decomposition of a time series into its components. These are:

- i) Additive model and ii) Multiplicative model

i) Additive Model

According to this model, a time series is the sum of its four components. Symbolically, $Y = T + S + C + I$

Where Y denotes the result of the four components,

T = Trends, S= Seasonal component, C = Cyclical component, and

I = Irregular component.

This model assumes that all the components of a time series are independent of one another. The model assumes that trend has no effect on the seasonal and cyclical components, nor seasonal swings have any influence on cyclical variations and vice versa. However, in most of the business and economic time series this assumption is not true. For example, the seasonal or cyclical fluctuations may virtually be wiped off by very sharp rising or falling trend. Similarly strong and powerful seasonal swings may intensify or even precipitate a change in cyclical fluctuations. The additive model also assumes that the different components are absolute quantities expressed in original units and can take positive and negative values.

ii) Multiplicative model

According to this model, a time-series is the product of its four components. Symbolically: $Y = T \times S \times C \times I$

The model assumes that trend has no effect on the seasonal and cyclical components, nor do seasonal swings have any influence on cyclical variations and vice versa.

Multiplicative model assumes that the four components of a time-series are due to different causes, but they are not necessarily independent and they can affect one another.

In this model only trend is expressed in terms of original values, while the seasonal and cyclical components are expressed as relatives or percentages. This model assumes that the four components of a time-series are due to different causes, but they are not necessarily independent and they can affect one another.

Algebraic Analysis of the models:

On the basis of these models, different components can be analysed using algebraic procedures. For example we can have

$C+I = Y-T-S$ in additive model and

$$C \times I = \frac{Y}{T \times S} \text{ in multiplicative model.}$$

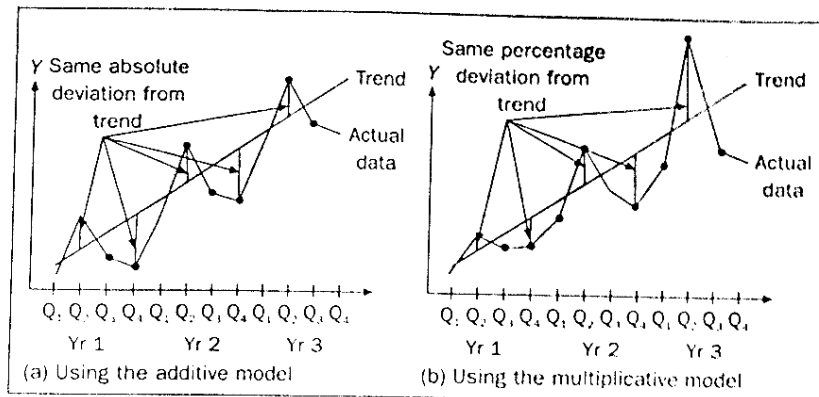
It is to be mentioned that most of the time-series relating to economic and business phenomenon conform to the multiplicative model, the additive model is used rarely.

Model Selection Criterion:

We use Additive model

$$Y = T + S + C + I$$

When the deviations from the trend line are of a similar absolute magnitude from one peak (or trough) to another. But when the deviations from the trend line are of a similar percentage from one peak (or trough) to another then multiplicative model $Y = T \times S \times C \times I$ is most appropriate. Following figures 13.2 illustrate the model selection criterions.



Self-Assessment Question:

Short questions

1. Define & explain different components of time series.
2. Describe the importance of time series analysis.
3. Discuss the types of models used in time series analysis.
4. Mention the model selection criterions.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) A time series is:
 - (a) A set of measurement on a variable collected at the some time on approximately the same period of time.
 - (b) A set of measurement on a variable taken over some time period in sequential order.
 - (c) A model that attempts to analyze the relationship between a dependent variable and one on more independent variable.
 - (d) A model that attempts to forecast the future value of a variable.
- (ii) A time series of annual data can contain which of the following components?
 - a) Secular trend.
 - b) Cyclical fluctuation.
 - c) Seasonal variation.
 - d) All of these.
- (iii) Which of the four time series components is more likely to exhibit the relative steady growth of the population of Bangladesh from 1964 to 2004.
 - (a) Trend Components`
 - (b) Cyclical Components
 - (c) Seasonal Components
 - (d) Irregular Components
- (iv) The model $Y = T+S+C+I$, that assures the time series value at time t is the sum of the forum time series components T , S , C and I ; is referred to as:
 - (a) Moving average model
 - (b) Multiplicative model
 - (c) Additive model
 - (d) Forecast model
- (v) Types of models used in time series analysis.
 - a) Additive
 - b) Multiplicative
 - c) Cyclic
 - d) Both (i) & (ii)

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) Seasonality is a component of time series
- (ii) Time-series analysis is used to detect patterns of change in statistical information over regular intervals of time
- (iii) Secular trends represent the long-term direction of a time series.
- (iv) The time series components that reflects a long-term, relatively smooth pattern on direction exhibited by a time series over a long time period is collect seasonal.
- (v) Time-series analysis helps us to analyze past trends, but it cannot aid us in future uncertainties.

Answer:

Multiple Choice Questions

- (i) b (ii) d (iii) a (iv) c (v) d

True/False

- (i) T (ii) T (iii) T (iv) F (v) T

Lesson 2: Measurement of Trend

Lesson Objectives:

After completing this lesson we would be able to:

- Explain the reasons for studying trend;
- Understand how to find trend;
- Explain different methods of finding trend;
- Understand merits & demerits of the methods.

Introduction

It should be kept in mind that not all time series are best represented in a mode. For this reason the performance of time series analysis is to operate on the data in such a way as to bring out separately each of the components present.

Trend Analysis

There are three reasons to study trends:

1. The study of trend allows us to describe a historical pattern.
2. Studying trend permits us to project past patterns, or trends, into the future.
3. By studying trend we can eliminate this component from the series.

Finding The Trend

Of the four components of a time series, secular trend represents the long-term direction of the series. One way to describe the trend component is to fit a line visually to set of points on a graph. Any given graph, however, is subject to slightly different interpretations by different individuals.

We can also fit a trend line by the method of least squares. In our discussion, we will concentrate on the method of least squares because visually fitting a line to a time series is not a completely dependable process. We shall also use method of moving average to find trend. Finally we shall describe the use of least squares technique for forecasting trend.

Method of Least Squares

This is the most widely used method and provides us with a mathematical device to obtain an objective fit to the trend of a given time series. This method is so called because a trend line computed by this method is such that the sum of the squares of the deviation between the original data, and the corresponding computed trend values is the minimum. This method can be used to fit either a straight-line trend or a parabolic trend.

A trend line computed by least square method is such that the sum of the squares of the deviation between the original data, and the corresponding computed trend values is the minimum

Straight Line Trend

The straight-line trend has an equation of the type: $Y = a + bX$

Where Y denotes the trend value of the dependent variable i.e. of the Y series, X is the independent variable i.e. time unit of X series, a and b are constants, a denoting the value of Y when $X = 0$ and b denoting the change in the value of Y for a unit change in X variable.

In order to determine the value of the constants a and b, the following normal equations are to be solved.

$$\sum Y = Na + b\sum X$$

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

Where N represents the number of years in the series.

Which gives

$$\text{Where } b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}, \quad \text{And} \quad a = \bar{Y} - b\bar{X}.$$

Using redefined data:

$$b = \frac{\sum xy}{\sum x^2}, \quad \text{and} \quad a = \bar{Y} - m\bar{X}$$

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

i.e. we redefine the origin from zero to the point of means $(\bar{X}, \bar{Y})$

Now, the estimated trend line denoted as $\hat{Y}$ and defined as $\hat{Y} = a + bx_i$

Coding Time

Normally, we measure the independent variable time in terms of weeks, months or years. Fortunately, we can convert these traditional measures of time to a form that simplifies the computation. To use coding here, we find the mean time and then subtract that value from each of the sample times.

A time series with odd and even number of elements. For odd number of elements we simply subtract the mean time from each element. For even number of elements when we find the mean and subtract it from each element, the fraction $\frac{1}{2}$ becomes part of the answer.

We need to consider two cases when we are coding time values. A time series with odd and even number of elements. For odd number of elements we simply subtract the mean time from each element. For even number of elements when we find the mean and subtract it from each element, the fraction $\frac{1}{2}$ becomes part of the answer. To simplify the coding process and to remove the $\frac{1}{2}$, we multiply each time element by 2. We will denote the “coded”, or translated, time with a lowercase x.

We have two reasons for this translation of time. First, it eliminates the need to square large numbers. Second, this method sets the mean year, $\bar{X}$, equal to zero.

Calculation of a and b with Coded data:

Now we can return to our calculations of the slope (b) and Y-intercept (a) to determine the best-fitting line.

$$\begin{aligned}
 \text{Since } b &= \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} \\
 &= \frac{\sum xY - n\bar{x}\bar{Y}}{\sum x^2 - n\bar{x}^2} \leftarrow \{ \text{Here; } x \text{ the coded variable substituted for } X \text{ and } \bar{x} \text{ substituted for } \bar{X} \} \\
 &= \frac{\sum xY}{\sum x^2} \text{ Since } \bar{x} = 0
 \end{aligned}$$

Therefore

Slope of the Trend for Coded Time Values

$$b = \frac{\sum xY}{\sum x^2}$$

$$\text{Farther } a = \bar{Y} - b\bar{X}$$

$$= \bar{Y} - b\bar{x} \leftarrow [\text{here } \bar{x} \text{ Substituted for } \bar{X}]$$

$$= \bar{Y}$$

Intercept of the Trend Line for Coded Time Values

$$a = \bar{Y}$$

Merits:

- i) The method of least squares is a mathematical method of measuring trend and is free from subjectiveness.
- ii) This method provides the line of best fit since it is this line from where the sum of positive and negative deviations is zero and the sum of square of deviations is the least.
- iii) This method enables us to compute the trend values for all the given time periods in the series.
- iv) The trend equation can be used to estimate the values of the variable for any given time period in future and forecasted values are quite reliable.
- v) This method is the only technique, which enables us to obtain the rate of growth per annum for yearly data in case of linear trend.

The trend equation can be used to estimate the values of the variable for any given time period in future and forecasted values are quite reliable.

Limitations:

- i) Fresh calculations become necessary if even a single new observation is added.
- ii) Calculations required in this method need basic computer knowledge.
- iii) Future predictions based on this method completely ignore the cyclical, seasonal and erratic fluctuations.

Parabolic Curves-Second Degree Parabola for Non-Linear Trend:

The second degree parabola is the simplest example of non linear trend.

A straight-line trend is a valid measure of trend for a series that tends to increase or decrease by a constant amount. It cannot describe the long term growth of an industry that expands by increments as the industry itself increase in size. In such cases we require a trend curve that will follow the tendency of a series throughout its course and will pass as nearly as possible through the centre of individual cycles. The second degree parabola is the simplest example of non linear trend. The equation of the second degree parabola is:

$$Y = a + bX + cX^2$$

Where 'a' is the value of Y at the origin, 'b' gives the slope at the point when $X=0$, and 'c' denotes the rate of change in the slope. The values of a, b and c can be obtained by solving the following three normal equations:

$$\sum Y = Na + b\sum X + c\sum X^2$$

$$\sum XY = a\sum X + b\sum X^2 + c\sum X^3$$

$$\sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4$$

When middle year is taken as the origin such that $\sum X = 0$ and $\sum X^3 = 0$. So for the Coded data above normal equations become:

$$\sum Y = Na + c\sum x^2 \dots\dots\dots(i)$$

$$\sum xY = b\sum x^2 \dots\dots\dots(ii)$$

$$\sum x^2Y = a\sum x^2 + c\sum x^4 \dots\dots\dots(iii)$$

The value of constant b can now be directly obtained from equation (ii) and now, the estimated trend line can be shown as

$$\hat{y} = a + bx + cx^2$$

The values of a and c can be obtained by solving equations (i) and (iii) simultaneously.

Example-11.1: A Natural Gas Company has supplied 18, 20, 21, 25, and 26, billion cubic feet of gas, respectively, for the years 1991 to 1995.

- (a) Find the liner estimating equation that best describes these data.
- (b) Find the trend values.

Solution:

Year X	$x = X - 1993$	Y	xY	x^2	$\hat{Y}$ Trend	$\frac{Y}{\hat{Y}} \times 100$	$\frac{Y - \hat{Y}}{\hat{Y}} \times 100$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1991	-2	18	-36	4	17.8	101.12	1.12
1992	-1	20	-20	1	19.9	100.50	0.50
1993	0	21	0	0	22.0	95.45	-4.55
1994	1	25	25	1	24.1	103.3	3.73
1995	2	26	52	4	26.2	99.24	-0.76
	0	110	21	10			

$$(a) \quad a = \bar{Y} = \frac{110}{5} = 22, \quad b = \frac{\sum xY}{\sum x^2} = \frac{21}{10} = 2.1$$

So $\hat{Y} = 22 + 2.1x$ (where 1993 = 0 and x units = 1 year)

- (b) Trend values are obtained putting values of X from column 2 in $\hat{Y} = 22 + 2.1x$
- (c) Trend values are given in column 6.

Example 11.2

The number of faculty-owned personal computers at the University of Ohio increased dramatically between 1990 and 1995:

Year	1990	1991	1992	1993	1994	1995
Number of PCs	50	110	350	1020	1950	3710

- a) Develop a linear estimating equation that best describes these data.
- b) Develop a second-degree estimating equation that best describes these data.
- c) Estimate the number of PCs that will be in use at the university in 1999, using both equations.
- d) If there are 8,000 faculty members at the university, which equation is the better predictor? Why?

Solution:

Year	x	Y	xY	x^2	x^2Y	x^4
1990	-5	50	-250	25	1250	625
1991	-3	110	-330	9	990	81
1992	-1	350	-350	1	350	1
1993	1	1020	1020	1	1020	1
1994	3	1950	5850	9	17550	81
1995	5	3710	18550	25	92750	625
	0	7190	24490	70	113910	1414

- i) Subtracting 1992.5 from each year and multiplying by 2 we get the coded (x) values. i. e., $X = 2(X - 1992.5)$.

now a

$$\bar{Y} = \frac{7190}{6} = 1198.333 \quad b = \frac{\sum xY}{\sum x^2} = \frac{2490}{70} = 349.8571$$

So $\hat{Y} = 1198.3333 + 349.8571x$ (Where 1992.5 = 0 and X units = 0.5 year)

ii) For second degree the equations are:

$$\sum Y = na + c\sum x^2 \quad \sum x^2 Y = a\sum x^2 + c\sum x^4$$

$$7190 = 6a + 70c \quad 113910 = 70a + 1414c$$

Which gives $a = 611.8750$, $c = 50.2679$

$$\text{So } \hat{Y} = 611.8750 + 349.8571X + 50.2679X^2$$

iii) Linear forecast: Since $x = 13$ then $\hat{Y} = 5746$ PCs
Second degree equation forecast: $\hat{Y} = 3655$ PCs.

- a. Neither is very good. The linear trend missed the acceleration in the rate of faculty PC acquisition. The second-degree trend assumed the acceleration would continue, ignoring the fact that there are only 8,000 faculty members.

Method of Moving average

The moving average we initially compute the simple total for a specified number of items of data.

We can also use the idea of a moving average to find the trend of the data. To find the moving average we initially compute the simple total for a specified number of items of data (For quarterly data we use four quarter total and for monthly data we use 12 month total). We then recalculate the total having dropped the initial item of data and added the subsequent item of data. This eliminates any seasonal variation as regards high or low values. By averaging these totals helps eliminate any irregular component.

If the period of moving average is even, then the moving total and moving average which are placed at the center of the time span from which they are computed fall between two time periods, and thus do not coincide with an original time period. In order to synchronize moving averages and original data, we adopt a process called centering. This process consists of taking a two period moving average of the moving averages.

Since in practice we are often presented with monthly or quarterly data over a number of years for which no obvious business cycle is present i.e. actual data have only three of the four components, i.e. $Y = T+S+I$ or $Y = T \times S \times I$. The process of moving average described above removes seasonal variation and irregular component, thus we are left with trend values.

Merits and Limitations

Merits

- i) Of all the mathematical methods of fitting a trend, this method is the simplest.
- ii) The method is flexible so that even if a few more observations are to be added; the entire calculations are not changed.
- iii) If the period of the moving average happens to coincide with the period of the cycle, the cyclical fluctuations are automatically eliminated.
- iv) The shape of the curve in case of moving average method is determined by the data rather than the statisticians' choice of mathematical function.

If the period of the moving average happens to coincide with the period of the cycle, the cyclical fluctuations are automatically eliminated.

Limitations

- i) Trend values cannot be computed for all the time points. In our example with 4 quarter moving averages, we cannot compute trend values for the first two and the last two quarters.
- ii) It is difficult to decide the period of moving average since there is no hard and fast rule for the purpose.
- iii) Moving average cannot be used in forecasting, as it is not represented by any mathematical function.
- iv) When the trend is not linear, the moving average lies either above or below the true sweep of the data.
- v) The moving average method gives a correct pattern of the general long-term tendency of the data under the following conditions:
 - a) The trend is linear or approximately so,
 - b) The cyclical variations affecting the data are regular both in period and amplitude.

Example 11.3: Find trend using a four-quarter centered moving average for the following 12 data on sales value of a product given over the periods 1995-97(Assume additive model).

Quarter Year1	1	2	3	4
1995	87.5	73.2	64.8	88.5
1996	90.3	76.0	69.2	94.7
1997	93.9	78.4	72.0	100.3

Solution:

Trend values are given in column 5

(1) Year and Quarter		(2) Sales Value (Y)	(3) 4 quarter moving total	(4) 4 quarter moving average	(5) Centred 4 quarter moving average (Trend T)	(6) S + I =Y-T
1995	1	87.5				
	2	73.2				
	3	64.8	314.0	78.5		
			316.8	79.2	78.9	-14.1
	4	88.5	319.6	79.9	79.6	8.9
1996	1	90.3	324.0	81.0	80.5	9.8
	2	76.0	330.2	82.6	81.8	-5.8
	3	69.2	333.8	83.5	83.1	-13.9
	4	94.7	336.2	84.1	83.8	10.9
1997	1	93.9	339.0	84.8	84.5	9.4
	2	78.4	344.6	86.2	85.5	-7.1
	3	72.0				
	4	100.3				

Least Squares Line for forecasting Trend

Trend values we obtained by the method of moving average can be plotted in the scatter plot against corresponding time and by linking the successive points by a sequence of straight lines we can estimate and forecast the trend values. We can also fit a single least squares line (regression line) to the trend values obtained by the method of moving average. This regression line, once calculated, can then be used for prediction (forecasting) of future trend values.

For finding the regression line we use the dependent variable Y standing for the trend value and the independent variable X standing for time. Using the original data the formula for the least squares regression line is: $\hat{Y} = a + bX$,

$$\text{Where } b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

$$\text{And } a = \bar{Y} - b\bar{X}.$$

Using redefined data:

$$b = \frac{\sum xy}{\sum x^2}, \quad a = \bar{Y} - m\bar{X}$$

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

i.e. we redefine the origin from zero to the point of means $(\bar{X}, \bar{Y})$

Example 11.4: Using the trend values of Example 11.3 (Column-5) and denoting 3rd quarter of 1995 as 1 estimate the trend line using Method of Least Squares.

Solution:

X (quarter data)	Y (trend value)	XY	X ²	Y ²
1	78.9	78.9	1	6225.2
2	79.6	159.2	4	6336.2
3	80.5	241.5	9	6480.3
4	81.8	327.2	16	6691.2
5	83.1	415.5	25	6905.6
6	83.8	502.8	36	7022.4
7	84.5	591.5	49	7140.3
8	85.5	684.0	64	7310.3
$\sum X = 36$	$\sum Y = 657.7$	$\sum XY = 3000.6$	$\sum X^2 = 204$	$\sum Y^2 = 54111.5$

$$\bar{X} = \frac{36}{8} = 4.5 \quad \bar{Y} = \frac{657.7}{8} = 82.2$$

Using original data, the formula for the regression (least squares) line is:

$$\hat{Y} = a + bX$$

$$\text{Where } b = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

$$a = \bar{Y} - b\bar{X}$$

Hence $b = 0.975$ and $a = 77.8$

$$\hat{Y}_T = 0.975X + 77.8$$

Where $\hat{Y}_T$ = trend value

X = quarterly time period (Q2, 1995 = 0)

We can now use this regression line to predict or forecast future trend values provided this past relationship holds into the future.

Exponential Smoothing forecasting

'Exponential smoothing' implies a learning process, whereby future forecasts are continually revised (smoothed) in the light of previous experience.

The idea of exponential smoothing is that the forecast at time t for the next time period ($t+1$) should take into account the observed error in the forecast made for t in the previous time period ($t-1$).

Clearly such 'exponential smoothing' implies a learning process, whereby future forecasts are continually revised (smoothed) in the light of previous experience. Strictly speaking, this approach is most appropriate when there is little or no trend (T) in the data and little or no seasonal variation (S). It is best used for short-term forecasting.

We can express this approach as follows:

Next forecast = Previous forecast + some proportion of the previous forecasting error.

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t)$$

Where Y_t = actual data (observed outcome) in time period t .

$\hat{Y}_t, \hat{Y}_{t+1}$ = forecast data for next time period (t or $t+1$) made in previous time period ($t-1$ or t respectively)

α = smoothing constant.

Putting in another way

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha(E_t)$$

Where E_t = error term = $Y_t - \hat{Y}_t$

The value assigned to α , the smoothing constant, can vary between zero and one. i.e. $0 \leq \alpha \leq 1$

We can usefully consider the extreme values of α by way of illustration.

- When $\alpha = 0$, no adjustment is made for the previous forecasting error. The next forecast is then assumed to be the same as the previous forecast.
- When $\alpha = 1$, full adjustment is made for the previous forecasting error. The next forecast is then the previous forecast $\pm$ the entire amount of any previous forecasting error.

In practice α can be derived experimentally, based on the average size of the error term in previous forecasts.

In practice α can be derived experimentally, based on the average size of the error term in previous forecasts. Values between 0.1 and 0.3 are typically assigned to α .

Self-Assessment Question:

Short questions

1. Describe the method of least squares for finding trend with its merits & demerits.
2. Describe method of moving average for finding trend with its merits and demerits.
3. Discuss the use of parabolic curves for non-linear trend.
4. State the idea of exponential smoothing for forecasting.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) The method of moving average is used:
(a) To plot a series (b) To exponentiate a series
(c) To smooth a series (d) In regression analysis
- (ii) In selecting an appropriate forecasting model, the following approaches are suggested:
(a) Perform a residual analysis
(b) Measure the size of the forecasting error
(c) Use the principle of parsimony
(d) All of the above
- (iii) Suppose that $\hat{Y} = 10 + 3x$ describes well an annual time series for 1987-1993. If the actual value of $\hat{Y}$ for 1990 is 8, what is the percent of trend for 1990?
a) 125% b) 112.5% c) 90% d) 80%
- (iv) If a time series has an even number of years, and we use coding, then each coded interval is equal to:
a) 1 year b) 2 years c) 1 month d) 6 months
- (v) Which of the following methods should not be used for short-term forecasts into the future?
(a) Exponential smoothing (b) Moving average
(c) Linear trend model (d) Autoregressive modeling

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The time series component that implies a long-term upward on downward pattern is called the trend component.
- (ii) When coding time values, we subtract from each value the smallest time value in the series; hence, the code of the smallest value is zero.

- (iii) When using least squares method to determine a second-degree equation of best fit, the values of four numerical constants must be determined.
- (iv) Centering is needed when there is an odd number of data.
- (v) One of the basic tools for creating a trend-based forecasting model is regression analysis.

Answer:

Multiple Choice Questions

- (i) c (ii) d (iii) c (iv) b (v) b

True/False

- (i) T (ii) T (iii) T (iv) F (v) T

Lesson 3: Seasonal Variations

Lesson Objectives:

After completing this lesson, we will be able to:

- Understand the reasons for studying seasonal variation;
- Know how to find seasonal variation;
- Understand different methods of finding seasonal variation;
- Explains merits & demerits of the methods;
- Understand uses of seasonal index.

Introduction

Seasonal variation is defined as repetitive and predictable movement around the trend line in one year or less. In order to detect seasonal variation, time intervals must be measured in small units, such as days, weeks, months, or quarters.

Reasons for studying seasonal variation

- i) We can establish the pattern of past changes. This gives us a way to compare two time intervals that would otherwise be too dissimilar.
- ii) It is useful to project past patterns into the future. In the case of long-range decisions, secular-trend analysis may be adequate. But for short-run decisions, the ability to predict seasonal fluctuations is often essential. For example, let us consider a wholesale food chain that wants to maintain adequate stock of all items. The ability to predict short-range patterns, such as the demand for candy at Pohela Baishakh, or Ice creams in the summer is useful to the management of the chain.
- iii) Once we have established the seasonal pattern that exists, we can eliminate its effects from the time series. This adjustment allows us to calculate the cyclical variation. When we eliminate the effect of seasonal variation from a time-series, we have deseasonalized time series.

Seasonal variation is useful to project past patterns into

When we eliminate the effect of seasonal variation from a time-series, we have deseasonalized time series.

Measurement of Seasonal Variations

The following methods are used for measuring the seasonal variations:

- i) Method of Simple Averages.
- ii) Ratio to Moving Average Method.
- iii) Ratio to Trend Method.
- iv) Link Relatives Method.

(s) Method of Simple Averages

This is the simplest method of obtaining seasonal indices. The steps of this method are:

- i) The unadjusted data are arranged by years and months (or quarters).
- ii) The values for each month or quarter are averaged for all the years.
- iii) Total of the monthly or quarterly averages are to be calculated.
- iv) Finally monthly or quarterly indices are computed in the following manner.

$$\text{Monthly Indices} = \frac{\text{Average of the Month} \times 1200}{\text{Total of the Monthly Averages}}$$

$$\text{Quarterly Indices} = \frac{\text{Average of the Quarter} \times 400}{\text{Total of the Quarterly Averages}}$$

$$\text{Weekly Indices} = \frac{\text{Average of the Week} \times 700}{\text{Total of the Weekly Averages}}$$

Example 11.5 A restaurant manager wishes to improve customer service and employee scheduling based on the daily levels of customers in the past 4 weeks. The numbers of customers served in the restaurant during the period were as follows

		Mon	Tue	Wed	Thu	Fri	Sat	Sun
Week	1	345	310	385	416	597	706	653
	2	418	333	400	515	664	761	702
	3	393	387	311	535	625	711	598
	4	406	412	377	444	650	803	822

Determine the seasonal (daily) indices for these data by the method of simple averages.

Solution:

The daily seasonal indices are given at the last row.

		Mon	Tue	Wed	Thu	Fri	Sat	Sun	Average
Week	1	345	310	385	416	597	706	653	
	2	418	333	400	515	664	761	702	
	3	393	387	311	535	625	711	598	
	4	406	412	377	444	650	803	822	
Average		390.5	360.5	368.25	447.25	634	745.25	693.75	519.93
Daily Seasonal Index		75.11	69.34	70.83	86.02	121.44	143.34	133.43	

Merits and Limitation

This is the simplest method of measuring seasonality in a time series. However, this method assumes that data do not contain any trend and cyclical fluctuations at all or their influence in the time series data is not quite significant. But this assumption is not true since most economic time series have trend and therefore, the seasonal index computed by this method is really an index of trends and seasons. In order to obtain any meaningful seasonal indices, it is necessary that first of all the trend effects are eliminated from the given values.

The seasonal index computed by this method is really an index of trends and seasons

(ii) Ratio to moving average method

In reality we deal with time series data having only three components T, S and I. Now for an additive model having the trend values (T) we can find (S+I) by subtracting the trend values (T) from the actual observations (Y)

i.e. $Y - T = S + I$ for additive model.

In case of multiplicative model the ratio $\frac{Y}{T}$ gives the value of the product $S \times I$. Now by averaging the values of (S+I) or ($S \times I$) against particular point of time we can remove I, leaving us with S. Finally seasonal indices are computed by using the adjusting constant. These indices are used to remove the effects of seasonality from a time series.

Example 11.6: Find the seasonal indices and the deseasonalized values using data of Example 11.3

Solution :

Arranging (S+I) values of column 6 in the solution table of example 11.3 against the corresponding quarters we get:

Table 2:

	Q-1	Q-2	Q-3	Q-4
1995			-14.1	8.9
1996	9.8	-5.8	-13.9	10.9
1997	9.4	-7.1		
Total	19.2	-12.9	-28.0	19.8
Average (S)	9.6	-6.45	-14	9.9

The average process of each quarter gives the seasonal variation for that quarter. But the average now is unadjusted and the net value is -0.95. (Computed as $9.6 - 6.45 - 14 + 9.9 = -0.95$)

Strictly speaking the plus and minus values should cancel out. Now adding $0.95 \div 4 = 0.24$ to each of the four quarterly values for S, we compensate -0.95 and the results are the adjusted S i.e. the seasonal variation component.

	Q-1	Q-2	Q-3	Q-4
Unadjusted S (-0.95 net)	+9.6	-6.45	-14.0	+9.9
Adjusted * S (zero net)	+9.48	-6.21	-13.76	+10.14

Since we are using additive model adding or subtracting seasonal indices with the original values the deseasonalised data are obtained.

Deseasonalised data:

		Y (original data)	S (adjusted seasonal variation)	Y-S (deseasonalised data)
1995	Q1	87.5	+9.84	77.66
	Q2	73.2	-6.21	79.41
	Q3	64.8	-13.76	78.56
	Q4	88.5	+10.14	78.36
1996	Q1	90.3	+9.84	80.46
	Q2	76.0	-6.21	82.21
	Q3	69.2	-13.76	82.96
	Q4	94.7	+10.14	84.56
1997	Q1	93.9	+9.84	84.06
	Q2	78.4	-6.21	84.61
	Q3	72.0	-13.76	85.76
	Q4	100.3	+10.14	90.16

(iii) Ratio to Trend Method:

Under this method, the trend for each monthly/quarter is calculated by the method of least squares. The trend so computed is eliminated from the original data by dividing the latter with the former. Then an average is computed for the ratio for each month or quarter. This process of averaging eliminates cyclical and random fluctuations, leaving only the seasonal component. This method involves the following steps:

- The trend for each month or quarter is computed.
- Each original figure is expressed as a percentage of the corresponding trend figures.
- The percentage for each month or quarter is averaged.
- Seasonal index is calculated expressing these percentages as percentage of their own averages.
- These indices are adjusted so that their mean value is 100.

Example 11.7: Calculate seasonal index for the following data by the ratio to trend method:

Year	1 st Quarter	2 nd Quarter	3 rd Quarter	4 th Quarter
1974	72	86	80	70
1975	76	70	82	76
1976	74	66	84	88
1977	76	74	84	86
1978	78	74	86	82

Solution :

For calculating seasonal index by ratio to trend method, first the trend for yearly data will be obtained and then these trend values will be converted into quarterly trend values. For this, average of each years quarterly values will be obtained and then assuming the mean to be Y_s , trend values will be obtained by the method of least squares.

Year	Yearly totals	Average of Quarterly values (Y)	Deviation from 1976 or (x)	xy	x^2
1974	308	77	-2	-154	4
1975	304	76	-1	-76	1
1976	312	78	0	0	0
1977	320	80	1	80	1
1978	320	80	2	160	4
N=5		$\Sigma Y = 391$	$\Sigma X = 0$	$\Sigma XY = 10$	$\Sigma X^2 = 10$

$$Y_c = a + bx$$

$$a = \frac{\sum y}{N} = \frac{391}{5} = 78.2$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{10}{10} = 1$$

$$\text{Quarterly Increase} = \frac{1}{4} = 0.25$$

Quarterly Trend Values

Year	1 st Quarter	2 nd Quarter	3 rd Quarter	4 th Quarter
1974	77.825	78.075	78.325	78.575
1975	78.825	79.075	79.325	79.575
1976	79.825	80.075	80.325	80.575
1977	80.825	81.075	81.325	81.575
1978	81.825	82.075	82.325	82.575

Given Quarterly values as percentage of Trend Values $\frac{Q}{T} \times 100$

This method provides ratio to trend values for each month or quarter for which data are available and is thus superior to ratio to moving average method where loss of data occurs in the beginning and end of the series.

Year	1 st Quarter	2 nd Quarter	3 rd Quarter	4 th Quarter
1974	92.515	110.150	102.138	89.086
1975	96.416	88.525	103.372	95.507
1976	92.702	82.423	104.575	109.215
1977	94.030	91.273	103.289	105.424
1978	95.325	90.161	104.464	99.303
Total	470.988	462.530	517.838	498.535
Average	94.198	92.506	103.568	99.707
Seasonal Index adjusted	96.6	94.9	106.2	102.3

Now total of Averages = $94.198 + 92.506 + 103.568 + 99.707 = 389.979$

Since the total of averages is less than 400, multiplying each average by $400/389.979$ makes an adjustment and adjusted seasonal indices are obtained.

Merits

- It is simple to compute and easy to understand.
- This method provides ratio to trend values for each month or quarter for which data are available and is thus superior to ratio to moving average method where loss of data occurs in the beginning and end of the series. Thus is a distinct advantage, especially when the period covered by the time series is very short.

Limitations

If there are pronounce cyclical swings in the series, the trend will never follow the actual data as closely as a 12-monthly or 4 quarterly moving average does. Thus a seasonal index calculated by ratio to moving average method may be less biased than the one calculated by ratio to trend method.

(iv) Link Relative Method

This method is also known as Pearson's method of constructing seasonal indices. It involves the following steps:

- Conversion of the original data into link relatives by dividing the value for each month (or quarter) by the value for the preceding month (or quarter) and multiplying it by 100. For example:

$$\text{Link Relative for February} : \frac{\text{Value for February}}{\text{Value for January}} \times 100$$

- Averaging these link relatives for each month (or quarter), the average being taken over the given number of years.
- Converting these average link relatives into chain relatives on the base of the first month (or quarter) taken as 100 by using the following formula:

$$\text{Chain Relative for any month} = (\text{L.R for that month} \times \text{C.R. for preceding month}) / 100$$

- iv) Calculating the chain relative of the first month (or quarter) on the base of the last month (or quarter). Usually this will not be equal to 100 due to the effect of long-term secular trend. It is, therefore, necessary to correct these chain relatives.
- v) For correction, the chain relative of the first month (or quarter), calculated by the first method is deducted from the chain relative of the first month (or quarter) calculated by the second method. The difference is divided by 12 (in case of monthly indices) or by 4 (in case of quarterly indices). The resulting figure multiplied by 1,2,3 and so on is deducted respectively from the chain relatives of the 1st 2nd, 3rd (and so on), months or quarters. These are corrected chain relatives.
- vi) The corrected chain relatives are expressed as percentages of their averages. These provide the required seasonal indices by the method of link relatives.

The chain relative of the first month (or quarter), calculated by the first method is deducted from the chain relative of the first month (or quarter) calculated by the second method.

Uses of the Seasonal Index

The seasonal indices are used to remove the effects of seasonality from a time-series. This is called deseasonalizing a time series. Before we can identify either the trend or cyclical components of a time series, we must eliminate seasonal variation. To deseasonalize a time series, we divide (add/subtract in the case of additive model) each of the actual values in the series by the appropriate seasonal index expressed as a fraction of 100. Once we have removed the seasonal variation, we can compute a deseasonalized trend line, which we can then project into the future.

Either the trend or cyclical components of a time series, we must eliminate seasonal variation

Self-Assessment Question:

Short questions

1. Mention the reasons for studying seasonal variation.
2. Describe the methods for measuring seasonal variation.
3. Describe the merits & demerits of the different methods of measuring seasonal variation.
4. Mention the uses of seasonal indices.

Multiple-Choice Question:

1. Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:

- (i) Which of the following statements about moving average is not true?
 - (a) It can be used to smooth a series
 - (b) It gives equal weights to all values in the computation
 - (c) It is simpler than the method of exponential smoothing.
 - (d) It gives greater weight to more recent data.
- (ii) The overall upward or downward pattern of the data in an annual time series will be contained in which following component:
 - (a) Trend (b) Cyclical (c) Irregular (d) Seasonal

- (iii) For a given year, if an adjusted seasonal index for some period is greater than 100, then the following must be true:
- (a) The adjusted index for some other period is >100
 - (b) The adjusted index for some other period is <100
 - (c) The adjusted index for some other period is $=100$
 - (d) (i) and (ii) but not (iii).
- (iv) When a time series appears to be increasing at an increasing rate, such that percentage difference from observation to observation is constant, the appropriate model to fit is the:
- (a) Linear trend
 - (b) Quadratic trend
 - (c) Exponential trend
 - (d) None of the above
- (v) A company has developed a linear trend regression model based on 16 quarters of data. The independent variable is the measure of time ($t = 1$ to 16, where quarters 1 is winters quarter, 2 is spring etc.). The company has also developed seasonal indexes for each quarter as follows:

Winter	Spring	Summer	Fall
1.20	1.00	0.70	1.10

The linear trend forecast equation is : $\hat{y} = 120 + 56t$.

Given this information, what is the seasonally adjusted forecast for period 19?

- (a) 1064
- (b) 1184
- (c) 828.80
- (d) None of the above

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) The time series component that reflects a long-term, relatively smooth pattern or direction exhibited by a time series over long time period is called seasonal.
- (ii) The repetitive movement around a trend line in a 2-year period is best described as a seasonal variation.
- (iii) Once seasonal indices are computed for a time series, the series can be deseasonalized so that only the trend component remains.
- (iv) We calculate the three period moving average for a time series for all time periods except the first period.
- (v) Seasonal variation is a repetitive and predictable variation around the trend line within a year.

Answer:

Multiple Choice Questions

- (i) d (ii) a (iii) a (iv) c (v) c

True/False

- (i) F (ii) F (iii) T (iv) F (v) T

Lesson 4: Cyclical Variation

Lesson Objectives:

After completing this lesson you will be able to:

- Describe cyclical variation
- Apply the techniques of measuring of the cyclical variation
- Apply the techniques the measuring of the irregular variation.

Introduction

Business cycles are perhaps the most important types of fluctuation in business data. In this unit we discuss various methods used for measuring cyclical and variations.

Cyclical Variation

Cyclical variation is the component of a time series that tends to oscillate above and below secular trend line for periods longer than one year.

Measure of Cyclical Variations

Cyclical components are the most difficult to measure. This is so because successive cycles vary widely in timing, pattern and amplitude and the cyclical rhythm is usually intermingled with irregular factor. As a result, it is impossible to construct meaningful typical cyclical indices or curves similar to those developed for trend and seasonal variation. The methods used for cyclical variations are:

- i) Residual Method
- ii) Relative cyclical residual Method

(i) Residual Method

Of all these methods, the residual method and relative cyclical residual methods are the most commonly used methods of estimating cyclical variations. These methods consist of eliminating seasonal variations and trend, thus obtaining the cyclical and irregular movements. This method involves the following steps:

- a) By an appropriate method, the trend values (T) and seasonal indices (S) are obtained.
- b) The original values are divided by the respective trend values and multiplied by 100 for each datum. This is ratio to trend in percentage.

$$\frac{Y}{T} \times 100 \text{ or } \frac{T \times S \times C \times I}{T} \times 100 = S \times C \times I$$

- c) The ratio to trend figures are divided by seasonal indices to obtain of cyclical the product and irregular variations:

$$\frac{S \times C \times I}{S} = C \times I$$

- d) If, the series is not influenced by random variations, then the figures obtained by step (c) will give cyclical variations. Otherwise, the data are smoothed by the method of moving

These methods consist of eliminating seasonal variations and trend, thus obtaining the cyclical and irregular movements.

This will eliminate the random variations, leaving only the cyclical variations.

average by taking an appropriate time period. This will eliminate the random variations, leaving only the cyclical variations. Cyclical movements are usually represented as percentages and are thus termed cyclical relatives.

Annual time series data contains only three components secular-trend, Cyclical and irregular components. After determining trend value ($\hat{Y}$) we can measure the cyclical variation as a percent of the trend.

$$\text{i.e. Cyclical variation} = \frac{Y}{\hat{Y}} \times 100$$

Where Y = Actual time series value.

In this method we assume that the cyclical component explains most of the variation, which are relatively unpredictable, we can't forecast any specific patterns of cyclical variation using this method.

(ii) Relative Cyclical Residual Method

For annual time series data or deseasonalized data in this method, the percentage deviation from the trend is found for each value and cyclical

$$\begin{aligned} \text{variation} &= \frac{Y - \hat{Y}}{\hat{Y}} \times 100 \\ &= \frac{Y}{\hat{Y}} \times 100 - 100 \\ &= \text{Percent trend} - 100 \end{aligned}$$

Both the measures of cyclical variation are percentages of trend and these are the values either above or below trend line.

Example-11.8: Using the data and the calculated trend values of example 11.1 answer the following:

- Calculate the percent of trend for these data.
- Calculate the relative cyclical residual for these data.
- In which years does the largest fluctuation from trend occur, and is it the same for both methods?

Solution:

Year	x	Y	xY	x ²	$\hat{Y}$ Trend	$\frac{Y}{\hat{Y}} \times 100$	$\frac{Y - \hat{Y}}{\hat{Y}} \times 100$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1991	-2	18	-36	4	17.8	101.12	1.12
1992	-1	20	-20	1	19.9	100.50	0.50
1993	0	21	0	0	22.0	95.45	-4.55
1994	1	25	25	1	24.1	103.3	3.73
1995	2	26	52	4	26.2	99.24	-0.76
	0	110	21	10			

- Percent of trend is given in column (7).
- Relative cyclical residual is given in column (8).
- Largest fluctuation (by both methods) was in 1993.

Irregular Variation

After eliminating trend, cyclical and seasonal variation from a time series, we are left with the unpredictable factor called irregular variation. Typically irregular variation occurs over short intervals and follows a random pattern. Irregular variation is very important but is not explainable mathematically. In most cases, irregular variation is difficult if not impossible to predict and we never attempt to “fit a line” to account for irregular variation. Often we will find irregular variation acknowledged with a footnote or a comment on a graph.

Typically irregular variation occurs over short intervals and follows a random pattern

Measurement of Irregular Variation

If T, S and C divide the original data, we get I for multiplicative model.

$$\frac{T \times S \times C \times I}{T \times S \times C} = I$$

In practice, however the cycle itself is so erratic and intermingled with irregular movements that it is impossible to separate the two, therefore, in practice, trend and seasonal variations are measured directly while cyclical and irregular variations are left together after the other elements have been removed.

Trend and seasonal variations are measured directly while cyclical & irregular variations are left together after the other elements have been removed.

Self-Assessment Question:

Short questions

1. What do you mean by cyclical variation? Cite examples of business data having cyclical variation.
2. Describe the measures of cyclical variation.
3. Define irregular component of a time series. How can you measure this component?
4. Why don't we project irregular variations into the future?

Multiple-Choice Question:

1. **Select the best response for each of the following items and put a tick mark (✓) the corresponding letter:**

- (i) Which time series component with take the most historical data to identify?
(a) Trend (b) Seasonal (c) Cyclical (d) Random
- (ii) A method used to deal with cyclical variation when the cyclical component does not explain most the variation left unexplained by the trend component is:
(a) Spearman analysis (b) Specific analysis
(c) Second-degree analysis (d) Relative cyclical residual

- (iii) If the percent of trend for a particular year in a time series is greater than 100% then for this year:
 - (a) The actual time series value lies above the trend line and the relative cyclical residual is positive.
 - (b) The actual time series value lies below the trend line and the relative cyclical residual is positive.
 - (c) The actual time series value lies above the trend line and the relative cyclical residual is negative.
 - (d) The actual time series value lies above the trend line and the relative cyclical residual is positive.
- (iv) Cyclical variation oscillates.
 - a) Below the trend line b) Above the trend line
 - c) Along the trend line d) None of the above case
- (v) A time series for the years 1985 – 1996 had the following relative cyclical residuals, in chronological order: -1%, -2%, 1%, 2%, -1%, -2%, 1%, 2%, -1%, -2%, 1%, 2%. The relative cyclical residual for 1997 should be:
 - a) 3% b) -1% c) -2%
 - d) Cannot be determined from information given.

2. Write “T” if the statement is true and “F” if the statement is false:

- (i) When using the residual method, we assume that the cyclical component explains most of the variation left unexplained by the trend component.
- (ii) In order to identify a cyclical component in time series data, one year of weekly data should be sufficient.
- (iii) The relative cyclical residual can be computed for an entry in a time series by subtracting 10 from the percent of trend for that entry.
- (iv) The percent of trend should not be used for predicting future cyclical variations.
- (v) In a recent meeting, a manager indicated that sales tend to be higher during October, November and December and lower in the spring. In making this statement, she is indicating that sales for the company are cyclical.

Answer:

Multiple Choice Questions

- (i) c (ii) c (iii) a (iv) b (v) d

True/False

- (i) F (ii) F (iii) T (iv) T (v) F

Exercise

1. (a) What is Business Forecasting? Explain clearly its role and limitations.
(b) How does analysis of time series helps in making business forecasting?
2. Explain clearly the different components into which a time series may be analyzed. Explain any method in isolating trend values in a time series.
3. Explain clearly the meaning of Time series Analysis. Mention its important components. Explain these components with examples, indicating the importance of each component in business.
4. (a) How seasonal variations are accounted for in the analysis of Time Series?
(b) What are the common methods is use for eliminating seasonality from a time series data?
5. Critically examine the various methods that are used for business forecasting. Why is time series considered to be an effective tool for forecasting analysis? Explain.
6. Explain the following terms in the study to time series:
(i) Secular trend (ii) Seasonal variation (iii) Cyclical fluctuations
7. Explain the method of moving averages is estimating the trend of a time series. What are the disadvantages in using this method?
8. The following series related to the profits of a commercial concern for 8 years.

Year	Profit (Rs.)	Year	Profit (Rs.)
1989	15,420	1993	26,120
1990	14,420	1994	31,950
1991	15,520	1995	35,360
1992	21,020	1996	35,670

9. Find out the trend values for the following time series of steel production by the method of moving average using 5 point time period for your purpose. State briefly the procedure that would have adopted if you were to choose a 4-point time period. How does one choose the proper 'period of the moving average'?

Year	Production (in tunes)	Year	Production (in tones)	Year	Production (in tones)
1985	351	1979	410	1991	502
1986	366	1980	420	1992	540
1987	361	1981	450	1993	557
1988	362	1982	500	1994	571
1989	400	1983	518	1995	586
1990	419	1984	455	1996	612

10. A company estimates its sales for a particular year to be Rs. 24,00,000. The season indicates for sales are as follows:

Months	Seasonal Index	Months	Seasonal Index
January	75	July	102
February	80	August	104
March	98	September	100
April	128	October	102
May	137	November	82
June	119	December	73

Using the information, calculate estimates of monthly sales of the company. Assume that there is no trend.

11. Table below shows the power and top speeds of different brands of sports cars:

Brand	:	A	B	C	D	E	F	G	H	I	J	K	L
Power X (kw)	:	70	63	72	60	66	70	74	65	62	67	65	68
Speed Y (kw/h)	:	155	150	180	135	156	168	178	160	132	145	139	152

- Find the best linear relationship that fits the given data.
- Estimate the spread of a car that has a power of 65 kw. and find a 95% confidence interval for this estimate.
- Determine how much of the variability in speed may be explained by the regression hypothesis.

Activity:

The number of people admitted to a Nursing Home per quarter is given in the following table:

	Spring	Summer	Fall	Winter
1992	29	30	41	43
1993	27	34	45	48
1994	33	36	46	51
1995	34	40	47	53

- Calculate the seasonal indices for these data (use a 4-quarter centered moving average).
- Deseasonalize these data using the indices from part (a).
- Find the least squares line that best describes the trend of the deseasonalized figures.

References

- Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2011). *Essentials of statistics for business and economics* (6th ed.). South-Western, Cengage Learning.
- Aczel, A. D., & Sounderpandian, J. (2008). *Complete business statistics* (7th ed.). McGraw-Hill/Irwin.
- Berenson, M. L., & Levine, D. M. (1996). *Basic business statistics: Concepts and applications* (6th ed.). Prentice Hall.
- Bhuyan, K. C. (2004). *Methods of statistics*. Sahitya Prokashani. India
- Black, K. (2010). *Business statistics for contemporary decision making* (6th ed.). John Wiley & Sons, Inc.
- Buglear, J. (2001). *Stats means business: A guide to business statistics*. Elsevier Butterworth-Heinemann.
- Byron, L. N. (1973). *Statistics for business*. Science Research Associates, Inc.
- Cochran, W. G. (1950). *Sampling techniques*. John Wiley.
- Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). John Wiley & Sons.
- Cynthia, F. (2007). *Business statistics for competitive advantage with Excel 2007: Basics, model building, and cases*. Springer Science+Business Media, LLC.
- Das, N. G. (1975). *Statistical methods*. Manasi Press.
- Dubois, E. N. (1964). *Essential methods in business statistics*. McGraw-Hill.
- Duncan, A. J. (1953). *Quality control and industrial statistics* (Parts III & IV). Richard D. Irwin.
- Evans, J. R. (2013). *Statistics, data analysis, and decision modeling* (5th ed.). Pearson.
- Fraser, C. (2007). *Business statistics for competitive advantage with Excel 2007*. Springer.
- Freund, J. E., Williams, F. J., & Perles, B. M. (1993). *Elementary business statistics* (6th ed.). Prentice Hall.
- Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed.). Marcel Dekker.

- Grant, E. L. (1964). *Statistical quality control* (Parts I–IV). McGraw-Hill.
- Groebner, D. F., Shannon, P. W., & Fry, P. C. (2018). *Business statistics: A decision-making approach*. Pearson Education Limited.
- Gupta, S. P. (2006). *Statistical methods*. Sultan Chand & Sons.
- Gupta, S. P., & Gupta, M. P. (2004). *Business statistics*. Sultan Chand & Sons.
- Huchendorf, S. C., Porter, D. C., & Schur, P. J. (2017). *Business statistics in practice: Using modeling, data, and analytics* (8th ed.). McGraw-Hill Education.
- Kazmier, L. J. (2004). *Business statistics* (4th ed.). McGraw-Hill.
- Kendall, M. G., & Stuart, A. (1966). *The advanced theory of statistics* (Vol. 3). Charles Griffin.
- Levine, D. M., Szabat, K. A., & Stephan, D. F. (2020). *Business statistics*. Pearson.
- Marchal, W. C., Lind, D. A., & Wathen, S. A. (2006). *Basic statistics for business and economics* (5th ed., International ed.). McGraw-Hill/Irwin.
- McClave, J. T., & Benson, P. G. (1994). *Statistics for business and economics* (6th ed.). Prentice Hall.
- Mill, T. C. (1990). *Time series for economics*. Cambridge University Press.
- Newbold, P. (2012). *Statistics for business and economics* (8th ed., Global ed.). Pearson Education.
- Raj, D. (1968). *Sampling theory*. McGraw-Hill.
- Rowntree, D. (1984). *Probability*. Charles Scribner's Sons.
- Sharpe, N. R., De Veaux, R. D., & Velleman, P. F. (2019). *Business statistics* (4th ed.). Pearson Education, Inc.
- Shavelson, R. J. (1988). *Statistical reasoning for the behavioral sciences* (2nd ed.). Allyn & Bacon Inc.
- Triola, M. F. (2021). *Elementary statistics* (14th ed.). Pearson.

Appendix

Statistical Tables

- I. Logarithms
- II. Antilogarithms
- III. Powers, Roots and Reciprocals
- IV. Binomial Coefficients
- V. Values of e^{-m}
- VI. Ordinates (Y) of the Standard Normal Curve at Z
- VII. Areas under the Standard Normal Distribution
- VIII. Critical Values of χ^2
- IX. Critical Values of t
- X. 5% Points of F-distribution
- XI. 1% Points of F-distribution
- XII. Control Charts Constants
- XIII. Random Numbers

I. LOGARITHMS																			
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	5	9	13	17	21	26	30	34	38
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0775	4	8	12	16	20	23	27	31	35
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	11	14	18	21	25	28	32
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	19	22	25	28
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	9	11	14	17	20	23	26
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	6	8	11	14	16	19	22	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	3	5	8	10	13	15	18	20	23
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	17	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
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	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	4183	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	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	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	5729	5740	5752	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	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	6	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	5	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8

II. LOGARITHMS													
	0	1	2	3	4	5	6	7	8	9	1 2 3	4 5 6	7 8 9
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1 2 3	3 4 5	6 7 8
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1 2 3	3 4 5	6 7 8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1 2 2	3 4 5	6 7 7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1 2 2	3 4 5	6 6 7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1 2 2	3 4 5	6 6 7
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1 2 2	3 4 5	5 6 7
56	7582	7490	7497	7505	7513	7520	7528	7536	7543	7551	1 2 2	3 4 5	5 6 7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1 2 2	3 4 5	5 6 7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1 1 2	3 4 4	5 6 7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1 1 2	3 4 4	5 6 7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1 1 2	3 4 4	5 5 6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1 1 2	3 4 4	5 5 6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1 1 2	3 3 4	5 5 6
63	7996	8003	8007	8014	8021	8028	8035	8041	8048	8055	1 1 2	3 3 4	5 5 6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1 1 2	3 3 4	5 5 6
65	8129	8136	8142	8149	8156	8162	8269	8176	8182	8189	1 1 2	3 3 4	5 5 6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1 1 2	3 3 4	5 5 6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1 1 2	3 3 4	5 5 6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1 1 2	3 3 4	4 5 6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1 1 2	2 3 4	4 5 6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1 1 2	2 3 4	4 5 6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1 1 2	2 3 4	4 5 5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1 1 2	2 3 4	4 5 5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1 1 2	2 3 4	4 5 5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1 1 2	2 3 4	4 5 5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1 1 2	2 3 3	4 5 5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1 1 2	2 3 3	4 5 5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1 1 2	2 3 3	4 4 5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1 1 2	2 3 3	4 4 5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1 1 2	2 3 3	4 4 5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1 1 2	2 3 3	4 4 5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1 1 2	2 3 3	4 4 5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1 1 2	2 3 3	4 4 5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1 1 2	2 3 3	4 4 5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1 1 2	2 3 3	4 4 5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1 1 2	2 3 3	4 4 5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1 1 2	2 3 3	4 4 5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0 1 1	2 2 3	3 4 4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0 1 1	2 2 3	3 4 4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0 1 1	2 2 3	3 4 4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0 1 1	2 2 3	3 4 4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0 1 1	2 2 3	3 4 4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0 1 1	2 2 3	3 4 4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0 1 1	2 2 3	3 4 4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0 1 1	2 2 3	3 4 4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0 1 1	2 2 3	3 4 4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0 1 1	2 2 3	3 4 4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0 1 1	2 2 3	3 4 4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0 1 1	2 2 3	3 4 4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0 1 1	2 2 3	3 3 4

I. ANTILOGARITHMS													
	0	1	2	3	4	5	6	7	8	9	123	456	789
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	0 0 1	1 1 1	2 2 2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	0 0 1	1 1 1	2 2 2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	0 0 1	1 1 1	2 2 2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	0 0 1	1 1 1	2 2 2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	0 1 1	1 1 2	2 2 2
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	0 1 1	1 1 2	2 2 2
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	0 1 1	1 1 2	2 2 2
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	0 1 1	1 1 2	2 2 2
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	0 1 1	1 1 2	2 2 3
.09	1230	1233	1236	1139	1242	1245	1247	1250	1253	1256	0 1 1	1 1 2	2 2 3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	0 1 1	1 1 2	2 2 3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	0 1 1	1 2 2	2 2 3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0 1 1	1 2 2	2 2 3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0 1 1	1 2 2	2 3 3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0 1 1	1 2 2	2 3 3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	0 1 1	1 2 2	2 3 3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	0 1 1	1 2 2	2 3 3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	0 1 1	1 2 2	2 3 3
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	0 1 1	1 2 2	2 3 3
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	0 1 1	1 2 2	3 3 3
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	0 1 1	1 2 2	3 3 3
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	0 1 1	2 2 2	3 3 3
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	0 1 1	2 2 2	3 3 3
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	0 1 1	2 2 2	3 3 4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	0 1 1	2 2 2	3 3 4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0 1 1	2 2 2	3 3 4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0 1 1	2 2 3	3 3 4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0 1 1	2 2 3	3 3 4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	0 1 1	2 2 3	3 4 4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	0 1 1	2 2 3	3 4 4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	0 1 1	2 2 3	3 4 4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	0 1 1	2 2 3	3 4 4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	0 1 1	2 2 3	3 4 4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	0 1 1	2 2 3	3 4 4
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	1 1 2	2 3 3	4 4 5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	1 1 2	2 3 3	4 4 5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	1 1 2	2 3 3	4 4 5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	1 1 2	2 3 3	4 4 5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	1 1 2	2 3 3	4 4 5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	1 1 2	2 3 3	4 5 5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	1 1 2	2 3 4	4 5 5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	1 1 2	2 3 4	4 5 5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	1 1 2	2 3 4	4 5 6
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	1 1 2	3 3 4	4 5 6
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	1 1 2	3 3 4	4 5 6
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	1 1 2	3 3 4	5 5 6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	1 1 2	3 3 4	5 5 6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	1 1 2	3 3 4	5 5 6
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	1 1 2	3 4 4	5 6 6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	1 1 2	3 4 4	5 6 6

II. ANTILOGARITHMS																								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9					
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7					
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7					
.52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7					
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7					
.54	3467	3475	3483	3491	3499	3508	3516	3524	3532	3540	1	2	2	3	4	5	6	6	7					
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7					
.56	3631	3639	3648	3656	3664	3673	3681	3690	3698	3707	1	2	3	3	4	5	6	7	8					
.57	3715	3724	3733	3741	3750	3758	3767	3776	3784	3793	1	2	3	3	4	5	6	7	8					
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8					
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	5	6	7	8					
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8					
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9					
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9					
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9					
.64	4365	4375	4385	4395	4406	4416	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9					
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9					
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10					
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10					
.68	4786	4797	4808	4819	4831	4842	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10					
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10					
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11					
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11					
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11					
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11					
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12					
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12					
.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	6223	6237	6252	6266	6281	6295	1	3	4	6	7	9	10	11	13					
.80	6310	6324	6339	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	7295	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	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	6	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	8128	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	10	12	14	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	14	16	18					
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19					
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19					
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20					
.98	9550	9572	9594	9616	9638	9661	9683	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					

III. POWERS, ROOTS AND RECIPROCAL								
n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\sqrt{10n}$	$\sqrt[3]{10n}$	$\sqrt[3]{100n}$	$\frac{1}{n}$
1	1	1	1	1	3.162	2.154	4.642	1
2	4	8	1.414	1.260	4.472	2.714	5.848	.5000
3	9	27	1.732	1.442	5.477	3.107	6.694	.3333
4	16	64	2	1.587	6.325	3.420	7.638	.2500
5	25	125	2.236	1.710	7.701	3.684	7.937	.2000
6	36	216	2.449	1.817	7.746	3.915	8.434	.1667
7	49	343	2.646	1.913	8.361	4.121	8.879	.1429
8	64	512	2.828	2.000	8.944	4.309	9.283	.1250
9	81	729	3.000	2.080	9.487	4.481	9.655	.1111
10	100	1000	3.162	2.154	10.0	4.642	10.000	.1000
11	121	1331	3.317	2.224	10.488	4.791	10.323	.09091
12	144	1728	3.464	2.289	10.954	4.932	10.627	.08333
13	169	2197	3.606	2.351	11.402	5.066	10.914	.07692
14	196	2744	3.742	2.410	11.832	5.192	11.187	.07143
15	225	3375	3.873	2.466	12.247	5.313	11.447	.06667
16	256	4096	4.000	2.520	12.649	5.429	11.696	.06250
17	289	5913	4.123	2.571	13.038	5.540	11.935	.05882
18	324	5832	4.243	2.621	13.416	5.646	12.164	.05556
19	361	6859	4.359	2.668	13.784	5.749	12.386	.05263
20	400	8000	4.472	2.714	14.142	5.848	12.599	.0500
21	441	9261	4.583	2.759	14.491	5.944	12.806	.04762
22	484	10648	4.690	2.802	14.832	6.037	13.006	.04545
23	529	12167	4.796	2.844	15.166	6.127	13.200	.04348
24	576	13824	4.899	2.884	15.492	6.214	13.389	.04167
25	625	15625	5.000	2.924	15.811	6.300	13.572	.0400
26	676	17576	5.099	2.962	16.125	6.383	13.751	.03846
27	729	19683	5.196	3.000	16.432	6.463	13.925	.03704
28	784	21952	5.292	3.037	16.733	6.542	14.095	.03571
29	841	24389	5.385	3.072	17.029	6.619	14.260	.03448
30	900	27000	5.477	3.107	17.321	6.694	14.422	.03333
31	961	29791	5.568	3.141	17.607	6.768	14.581	.03226
32	1024	32768	5.657	3.175	17.889	6.840	14.736	.03125
33	1089	35937	5.745	3.208	18.166	6.910	14.888	.03030
34	1156	39304	5.831	3.240	18.439	6.980	15.037	.02941
35	1225	42875	5.916	3.271	18.708	7.047	15.183	.02857
36	1296	46656	6.000	3.302	18.974	7.114	15.326	.02778
37	1369	50653	6.083	3.332	19.235	7.179	15.467	.02703
38	1444	54872	6.164	3.362	19.494	7.243	15.605	.02632
39	1521	59319	6.245	3.391	19.748	7.306	15.741	.02564
40	1600	64000	6.325	3.420	20.00	7.368	15.874	.0250
41	1681	68921	6.403	3.448	20.248	7.429	16.005	.02439
42	1764	74088	6.481	3.476	20.494	7.489	16.134	.02381
43	1849	79507	6.557	3.503	20.736	7.548	16.261	.02326
44	1936	85184	6.633	3.530	20.976	7.606	16.386	.02273
45	2025	91125	6.708	3.557	21.213	7.663	16.510	.02222
46	2116	97336	6.782	3.583	21.448	7.719	16.631	.02174
47	2209	103823	6.856	3.609	21.679	7.775	16.751	.02128
48	2304	110592	6.928	3.634	21.909	7.830	16.869	.02083
49	2401	117649	7.000	3.659	22.136	7.884	16.985	.02041
50	2500	125000	7.071	3.684	22.361	7.937	17.100	.020

III. POWERS, ROOTS AND RECIPROCAL								
n	n^2	n^3	$\sqrt{n}$	$\sqrt[3]{n}$	$\sqrt{10n}$	$\sqrt[3]{10n}$	$\sqrt[3]{100n}$	$\frac{1}{n}$
51	2601	132651	7.141	3.708	22.583	7.990	17.213	.01961
52	2704	140608	7.211	3.733	22.804	8.041	17.325	.01923
53	2809	148877	7.280	3.756	23.022	8.093	17.435	.01887
54	2916	157464	7.348	3.780	23.238	8.143	17.544	.01852
55	3025	166375	7.416	3.803	23.452	8.193	17.652	.01818
56	3136	175616	7.483	3.826	23.664	8.243	17.758	.01786
57	3249	185193	7.550	3.849	23.875	8.291	17.863	.01754
58	3364	195112	7.616	3.871	24.083	8.340	17.967	.01724
59	3481	205379	7.681	3.893	24.290	8.387	18.070	.01695
60	3600	216000	7.746	3.915	24.495	8.434	18.171	.01667
61	3721	226981	7.810	3.936	24.698	8.481	18.272	.01639
62	3844	238328	7.874	3.958	24.900	8.527	18.371	.01613
63	3969	250047	7.937	3.979	25.100	8.573	18.469	.01587
64	4096	262144	8.000	4.000	25.298	8.618	18.566	.01562
65	4225	274625	8.062	4.021	25.495	8.662	18.663	.01538
66	4356	287496	8.124	4.041	25.690	8.707	18.758	.01515
67	4489	300763	8.185	4.062	25.884	8.750	18.852	.01493
68	4624	314432	8.246	4.082	26.077	8.794	18.945	.01471
69	4761	328509	8.307	4.102	26.268	8.837	19.038	.01449
70	4900	343000	8.367	4.121	26.458	8.879	19.129	.01429
71	5041	357011	8.426	4.141	26.646	8.921	19.220	.01408
72	5184	373248	8.485	4.160	26.833	8.963	19.310	.01389
73	5329	389017	8.544	4.179	27.019	9.004	19.399	.01370
74	5476	405224	8.602	4.198	27.203	9.045	19.487	.01351
75	5625	421875	8.660	4.217	27.386	9.086	19.574	.01333
76	5776	438976	8.718	4.236	27.568	9.126	19.661	.01316
77	5929	456533	8.775	4.254	27.740	9.166	19.747	.01299
78	6084	474552	8.832	4.273	27.928	9.205	19.832	.01282
79	6241	493039	8.883	4.291	28.107	9.244	19.916	.01266
80	6400	512000	8.944	4.309	28.284	9.283	20.000	.01250
81	6561	531441	9.000	4.327	28.460	9.322	20.083	.01235
82	6724	551368	9.055	4.344	28.636	9.360	20.165	.01220
83	6889	571787	9.110	4.362	28.810	9.398	20.247	.01205
84	7056	592704	9.165	4.380	28.983	9.435	20.328	.01190
85	7225	614125	9.220	4.397	29.155	9.473	20.408	.01176
86	7396	636056	9.274	4.414	29.326	9.510	20.488	.01163
87	7569	658503	9.327	4.431	29.496	9.546	20.567	.01149
88	7744	681472	9.381	4.448	29.665	9.583	20.646	.01136
89	7921	704969	9.434	4.465	29.833	9.619	20.724	.01124
90	8100	729000	9.487	4.481	30.000	9.655	20.801	.01111
91	8281	753571	9.538	4.498	30.166	9.691	20.878	.01099
92	8464	778688	9.592	4.514	30.332	9.726	20.954	.01087
93	8649	804357	9.644	4.531	30.496	9.761	21.029	.01075
94	8836	830584	9.695	4.547	30.659	9.796	21.105	.01064
95	9025	857375	9.747	4.563	30.822	9.830	21.179	.01053
96	9216	884736	9.798	4.579	30.984	9.865	21.253	.01042
97	9409	912673	9.849	4.595	31.145	9.899	21.327	.01031
98	9604	941192	9.899	4.610	31.305	9.933	21.400	.01020
99	9801	970299	9.909	4.626	31.464	9.967	21.472	.01010
100	10000	1000000	10.000	4.642	31.623	10.000	21.544	.0100

IV. BINOMIAL COEFFICIENTS

n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2	1								
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1287	715	286	78
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184756

V. VALUES OF e^{-m} (For Computing Poisson Probabilities)
($0 < m < 1$)

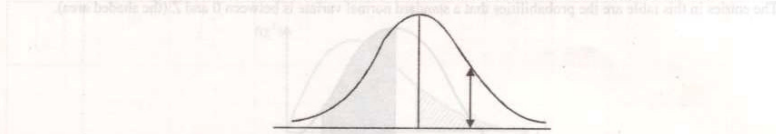
m	0	1	2	3	4	5	6	7	8	9
0.0	1.0000	.9900	.9802	.9704	.9608	.9512	.9418	.9324	.9231	.9139
0.1	0.9048	.8958	.8860	.8781	.8694	.8607	.8521	.8437	.8353	.8270
0.2	0.8187	.8106	.8025	.7945	.7866	.7788	.7711	.7634	.7558	.7483
0.3	0.7408	.7334	.7261	.7189	.7118	.7047	.6977	.6907	.6839	.6771
0.4	0.6703	.6636	.6570	.6505	.6440	.6376	.6313	.6250	.6188	.6126
0.5	0.6065	.6005	.5945	.5886	.5827	.5770	.5712	.5655	.5599	.5543
0.6	0.5488	.5434	.5379	.5326	.5278	.5220	.5160	.5117	.5066	.5016
0.7	0.4966	.4916	.4868	.4810	.4771	.4724	.4670	.4630	.4584	.4538
0.8	0.4493	.4449	.4404	.4360	.4317	.4274	.4232	.4190	.4148	.4107
0.9	0.4066	.4025	.3985	.3946	.3906	.3867	.3829	.3791	.3753	.3716

(m = 1, 2, 3, ..., 10)

m	1	2	3	4	5	6	7	8	9	10
e^{-m}	.36788	.13534	.04979	.01832	.006737	.002478	.00092	.000335	.000123	.000045

Note : To obtain values of e^{-m} for other values of m , use the laws of exponents.Example. $e^{-2.35} = (e^{-2.00})(e^{-0.35}) = (.13534)(.7047) = .095374$

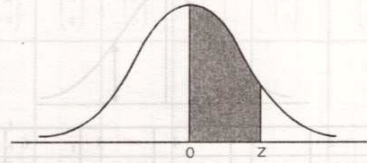
VI. ORDINATES (Y) OF THE STANDARD NORMAL CURVE AT z



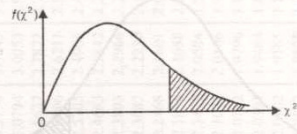
z	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	0.3989	0.3989	0.3988	0.3986	0.3984	0.3982	0.3980	0.3977	0.3973
0.1	0.3970	0.3965	0.3961	0.3956	0.3951	0.3945	0.3939	0.3932	0.3925	0.3918
0.2	0.3910	0.3902	0.3894	0.3885	0.3876	0.3867	0.3857	0.3847	0.3836	0.3825
0.3	0.3814	0.3802	0.3790	0.3778	0.3765	0.3752	0.3739	0.3725	0.3712	0.3697
0.4	0.3683	0.3668	0.3653	0.3637	0.3621	0.3605	0.3589	0.3572	0.3555	0.3538
0.5	0.3521	0.3503	0.3485	0.3467	0.3448	0.3429	0.3410	0.3391	0.3372	0.3352
0.6	0.3332	0.3312	0.3292	0.3271	0.3251	0.3230	0.3209	0.3187	0.3166	0.3144
0.7	0.3123	0.3101	0.3079	0.3056	0.3034	0.3011	0.2989	0.2966	0.2943	0.2920
0.8	0.2897	0.2874	0.2850	0.2827	0.2803	0.2780	0.2756	0.2732	0.2709	0.2685
0.9	0.2661	0.2637	0.2613	0.2589	0.2565	0.2541	0.2516	0.2492	0.2468	0.2444
1.0	0.2420	0.2396	0.2371	0.2347	0.2323	0.2299	0.2275	0.2251	0.2227	0.2203
1.1	0.2179	0.2155	0.2131	0.2107	0.2083	0.2059	0.2036	0.2012	0.1989	0.1965
1.2	0.1942	0.1919	0.1895	0.1872	0.1849	0.1826	0.1804	0.1781	0.1758	0.1736
1.3	0.1714	0.1691	0.1669	0.1647	0.1626	0.1604	0.1582	0.1561	0.1539	0.1518
1.4	0.1497	0.1476	0.1450	0.1435	0.1415	0.1394	0.1374	0.1354	0.1334	0.1315
1.5	0.1295	0.1276	0.1257	0.1238	0.1219	0.1200	0.1182	0.1163	0.1145	0.1127
1.6	0.1109	0.1092	0.1074	0.1057	0.1040	0.1023	0.1006	0.0989	0.0973	0.0957
1.7	0.0940	0.0925	0.0909	0.0893	0.0878	0.0863	0.0848	0.0833	0.0818	0.0804
1.8	0.0790	0.0775	0.0761	0.0748	0.0734	0.0721	0.0707	0.0694	0.0681	0.0669
1.9	0.0656	0.0644	0.0632	0.0620	0.0608	0.0596	0.0584	0.0573	0.0562	0.0551
2.0	0.0540	0.0529	0.0519	0.0508	0.0498	0.0488	0.0478	0.0468	0.0459	0.0449
2.1	0.0440	0.0431	0.0422	0.0413	0.0404	0.0396	0.0387	0.0379	0.0371	0.0363
2.2	0.0355	0.0347	0.0339	0.0332	0.0325	0.0317	0.0310	0.0303	0.0297	0.0290
2.3	0.0283	0.0277	0.0270	0.0264	0.0258	0.0252	0.0246	0.0241	0.0235	0.0229
2.4	0.0224	0.0219	0.0213	0.0208	0.0203	0.0198	0.0194	0.0189	0.0184	0.0180
2.5	0.0175	0.0171	0.0167	0.0163	0.0158	0.0154	0.0151	0.0147	0.0143	0.0139
2.6	0.0136	0.0132	0.0129	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.0107
2.7	0.0104	0.0101	0.0099	0.0096	0.0093	0.0091	0.0088	0.0086	0.0084	0.0081
2.8	0.0079	0.0077	0.0075	0.0073	0.0071	0.0069	0.0067	0.0065	0.0063	0.0061
2.9	0.0060	0.0058	0.0056	0.0055	0.0051	0.0051	0.0050	0.0048	0.0047	0.0046
3.0	0.0044	0.0043	0.0042	0.0040	0.0039	0.0038	0.0037	0.0036	0.0035	0.0034
3.1	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026	0.0025	0.0025
3.2	0.0024	0.0023	0.0022	0.0022	0.0021	0.0020	0.0020	0.0019	0.0018	0.0018
3.3	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0013	0.0013
3.4	0.0012	0.0012	0.0012	0.0011	0.0011	0.0010	0.0010	0.0010	0.0009	0.0009
3.5	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006
3.6	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0004
3.7	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003	0.0003	0.0003	0.0003
3.8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002	0.0002	0.0002	0.0002
3.9	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0001	0.0001

VII. AREAS UNDER THE STANDARD NORMAL DISTRIBUTION

The entries in this table are the probabilities that a standard normal variate is between 0 and Z (the shaded area).

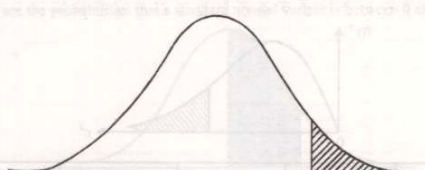


Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

VIII. CRITICAL VALUES OF CHI-SQUARE (χ^2)

α df.	.100	.050	.025	.010	.005	.001
1	2.71	3.84	5.02	6.63	7.88	10.8
2	4.61	5.99	7.38	9.21	10.6	13.8
3	6.25	7.81	9.35	11.3	12.8	16.3
4	7.78	9.49	11.1	13.3	14.9	18.5
5	9.24	11.1	12.8	15.1	16.7	20.5
6	10.6	12.6	14.4	16.8	18.5	22.5
7	12.0	14.1	16.0	18.5	20.3	24.3
8	13.4	15.5	17.5	20.1	22.0	26.1
9	14.7	16.9	19.0	21.7	23.6	27.9
10	16.0	18.3	20.5	23.2	25.2	29.6
11	17.3	19.7	21.9	24.7	26.8	31.3
12	18.5	21.0	23.3	26.2	28.3	32.9
13	19.8	22.4	24.7	27.7	29.8	34.5
14	21.1	23.7	26.1	29.1	31.3	36.1
15	22.3	25.0	27.5	30.6	32.8	37.7
16	23.5	26.3	28.8	32.0	34.3	39.3
17	24.8	27.6	30.2	33.4	35.7	40.8
18	26.0	28.9	31.5	34.8	37.2	42.3
19	27.2	30.1	32.9	36.2	38.6	43.8
20	28.4	31.4	34.2	37.6	40.0	45.3
21	29.6	32.7	35.5	38.9	41.4	46.8
22	30.8	33.9	36.8	40.3	42.8	48.3
23	32.0	35.2	38.1	41.6	44.2	49.7
24	33.2	36.4	39.4	43.0	45.6	51.2
25	34.4	37.7	40.6	44.3	46.9	52.6
26	35.6	38.9	41.9	45.6	48.3	54.1
27	36.7	40.1	43.2	47.0	49.6	55.5
28	37.9	41.3	44.5	48.3	51.0	56.9
29	39.1	42.6	45.7	49.6	52.3	58.3
30	40.3	43.8	47.0	50.9	53.7	59.7
35	46.1	49.8	53.2	57.3	60.3	66.6
40	51.8	55.8	59.3	63.7	66.8	73.4
45	57.5	61.7	65.4	70.0	73.2	80.1
50	63.2	67.5	71.4	76.2	79.5	86.7
55	68.8	73.3	77.4	82.3	85.7	93.2
60	74.4	79.1	83.3	88.4	92.0	99.6
65	80.0	84.8	89.2	94.4	98.1	106.0
70	85.5	90.5	95.0	100.4	104.0	112.3
75	91.1	96.2	100.8	106.4	110.3	118.6
80	96.6	101.9	106.6	112.3	116.3	124.8
85	102.1	107.5	112.4	118.2	122.3	131.0
90	107.6	113.1	118.1	124.1	128.3	137.2
95	113.0	118.8	123.9	130.0	134.2	143.3
100	118.5	124.3	129.6	135.8	140.2	149.4

IX. CRITICAL VALUES OF t



df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576

X. 5% POINTS OF FISHER'S F-DISTRIBUTION



$m \backslash n$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	α
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	250.09	252.20	254.32
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413	19.430	19.446	19.462	19.479	19.496
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8868	8.8452	8.8123	8.7855	8.7446	8.7029	8.6602	8.6166	8.5720	8.5265
4	7.7086	6.9443	6.5914	6.3883	6.2560	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117	5.8578	5.8025	5.7459	5.6878	5.6281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8753	4.8183	4.7725	4.7351	4.6777	4.6188	4.5581	4.4957	4.4314	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2066	4.1468	4.0990	4.0600	3.9999	3.9381	3.8742	3.8082	3.7398	3.6688
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747	3.5108	3.4445	3.3758	3.3043	3.2298
8	5.3177	4.4590	4.0662	3.8378	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2840	3.2184	3.1503	3.0794	3.0053	2.9276
9	5.1174	4.2565	3.8626	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.8637	2.7872	2.7077
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.6996	2.6211	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876	2.7186	2.6464	2.5705	2.4901	2.4045
12	4.7272	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6169	2.5436	2.4663	2.3842	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037	2.5331	2.4589	2.3803	2.2966	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6021	2.5342	2.4630	2.3879	2.3082	2.2230	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4035	2.3275	2.2468	2.1601	2.0658
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247	2.3522	2.2756	2.1938	2.1058	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807	2.3077	2.2304	2.1477	2.0584	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421	2.2686	2.1906	2.1071	2.0166	1.9168
19	4.3808	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080	2.2341	2.1555	2.0712	1.9796	1.8780
20	4.3513	3.4928	3.0984	2.8661	2.7100	2.5980	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0391	1.9464	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3661	2.3210	2.2504	2.1757	2.0960	2.0102	1.9165	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258	2.1508	2.0707	1.9842	1.8895	1.7831
23	4.2793	3.4221	3.0280	2.7955	2.6500	2.5377	2.4522	2.3848	2.3301	2.2847	2.2134	2.1382	2.0576	1.9706	1.8759	1.7685
24	4.2597	3.4028	3.0088	2.7763	2.6307	2.5182	2.4326	2.3651	2.3102	2.2647	2.1934	2.1182	2.0372	1.9500	1.8553	1.7479
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649	2.0889	2.0075	1.9192	1.8217	1.7110
26	4.2252	3.3690	2.9751	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479	2.0716	1.9898	1.9010	1.8027	1.6906
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323	2.0558	1.9736	1.8842	1.7851	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179	2.0411	1.9586	1.8687	1.7689	1.6541
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2782	2.2229	2.1768	2.1045	2.0275	1.9446	1.8543	1.7537	1.6377
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8409	1.7396	1.6223
40	4.0848	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9245	1.8389	1.7444	1.6373	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3688	2.2540	2.1665	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.6491	1.5343	1.3893
120	3.9201	3.0718	2.6802	2.4472	2.2900	2.1750	2.0867	2.0164	1.9588	1.9105	1.8337	1.7505	1.6587	1.5543	1.4290	1.2539
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.4591	1.3180	1.0000

XI. 1% POINTS OF FISHER'S F-DISTRIBUTION

$\frac{m}{n}$	1	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
1	4052.2	4999.5	5403.3	5624.6	5763.7	5859.0	5928.3	5981.6	6022.5	6055.8	6106.3	6157.3	6208.7	6260.7	6313.0	6366.0
2	98.503	90.000	84.166	79.249	75.637	72.911	70.752	69.356	68.388	67.729	67.052	66.472	65.949	65.466	65.019	64.591
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.132	27.052	26.980	26.916	26.859	26.805
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.454	14.374	14.302	14.238	14.180	14.126
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.963	9.883	9.812	9.748	9.690	9.636
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1016	7.9761	7.8741	7.783	7.700	7.625	7.558	7.496	7.438
7	12.246	9.5466	8.4513	7.8467	7.4604	7.1914	6.9928	6.8401	6.7188	6.6201	6.534	6.459	6.392	6.331	6.273	6.218
8	11.259	8.6491	7.5910	7.0060	6.6318	6.3707	6.1776	6.0289	5.9106	5.8143	5.730	5.655	5.589	5.529	5.473	5.419
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511	5.2565	5.174	5.099	5.032	4.971	4.914	4.859
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424	4.8492	4.767	4.691	4.622	4.560	4.501	4.444
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315	4.5393	4.457	4.381	4.311	4.248	4.188	4.130
12	9.3302	6.9266	5.9526	5.4119	5.0643	4.8206	4.6395	4.4984	4.3875	4.2961	4.214	4.138	4.067	4.003	3.942	3.882
13	9.0738	6.7010	5.7539	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911	4.1003	4.019	3.943	3.871	3.807	3.745	3.684
14	8.8616	6.5149	5.5939	5.0354	4.6916	4.4504	4.2719	4.1339	4.0229	3.9321	3.851	3.775	3.703	3.639	3.576	3.514
15	8.6831	6.3589	5.4170	4.8583	4.5145	4.2733	4.0948	3.9568	3.8458	3.7550	3.674	3.598	3.526	3.462	3.398	3.335
16	8.5310	6.2262	5.2922	4.7326	4.3888	4.1476	3.9691	3.8311	3.7201	3.6293	3.548	3.472	3.400	3.336	3.272	3.208
17	8.3997	6.1121	5.1850	4.6250	4.2812	4.0400	3.8615	3.7235	3.6125	3.5217	3.440	3.364	3.292	3.228	3.164	3.100
18	8.2854	6.0129	5.0919	4.5318	4.1880	3.9468	3.7683	3.6303	3.5193	3.4285	3.347	3.271	3.200	3.136	3.072	3.008
19	8.1850	5.9259	5.0103	4.4503	4.1065	3.8653	3.6868	3.5488	3.4378	3.3470	3.266	3.190	3.119	3.055	2.991	2.927
20	8.0960	5.8489	4.9382	4.3782	4.0344	3.7932	3.6147	3.4767	3.3657	3.2749	3.194	3.118	3.047	2.983	2.919	2.855
21	8.0166	5.7804	4.8740	4.3140	3.9702	3.7290	3.5505	3.4125	3.3015	3.2107	3.130	3.054	2.983	2.919	2.855	2.791
22	7.9454	5.7190	4.8166	4.2566	3.9128	3.6716	3.4931	3.3551	3.2441	3.1533	3.073	2.997	2.926	2.862	2.798	2.734
23	7.8811	5.6637	4.7649	4.2049	3.8611	3.6200	3.4415	3.3035	3.1925	3.1017	3.021	2.945	2.874	2.810	2.746	2.682
24	7.8229	5.6136	4.7181	4.1581	3.8143	3.5732	3.3947	3.2567	3.1457	3.0549	2.974	2.898	2.827	2.763	2.699	2.635
25	7.7698	5.5680	4.6755	4.1155	3.7717	3.5306	3.3521	3.2141	3.1031	3.0123	2.932	2.856	2.785	2.721	2.657	2.593
26	7.7213	5.5263	4.6366	4.0766	3.7328	3.4917	3.3132	3.1752	3.0642	2.9734	2.893	2.817	2.746	2.682	2.618	2.554
27	7.6767	5.4881	4.6009	4.0409	3.6971	3.4560	3.2775	3.1395	3.0285	2.9377	2.857	2.781	2.710	2.646	2.582	2.518
28	7.6356	5.4529	4.5681	4.0081	3.6643	3.4232	3.2447	3.1067	2.9957	2.9049	2.824	2.748	2.677	2.613	2.549	2.485
29	7.5976	5.4208	4.5381	3.9781	3.6343	3.3932	3.2147	3.0767	2.9657	2.8749	2.794	2.718	2.647	2.583	2.519	2.455
30	7.5625	5.3904	4.5097	3.9497	3.6059	3.3648	3.1863	3.0483	2.9373	2.8465	2.766	2.690	2.619	2.555	2.491	2.427
40	7.3141	5.1785	4.3126	3.7526	3.4088	3.1677	2.9892	2.8512	2.7402	2.6494	2.569	2.493	2.422	2.358	2.294	2.230
60	7.0771	4.9774	4.1259	3.5659	3.2221	2.9810	2.8025	2.6645	2.5535	2.4627	2.382	2.306	2.235	2.171	2.107	2.043
120	6.8510	4.7865	3.9493	3.3893	3.0455	2.8044	2.6259	2.4879	2.3769	2.2861	2.206	2.130	2.059	2.000	1.936	1.872
∞	6.6349	4.6052	3.7816	3.2216	2.8778	2.6367	2.4582	2.3202	2.2092	2.1184	2.038	1.962	1.891	1.832	1.768	1.704

For $m > 10$ interpolate using $60/m$.For $n > 30$ interpolate using $120/n$.

XII. FACTORS USEFUL IN THE CONSTRUCTION OF CONTROL CHARTS

Sample size	Mean-chart				Factors for central line				Factors for control limit				Range-chart			
	Factors for control limit				Factors for central line				Factors for control limit				Factors for central line			
	A	A ₁	A ₂	c ₄	B ₁	B ₂	B ₃	B ₄	d ₂	D ₁	D ₂	D ₃	D ₄	D ₅	D ₆	D ₇
2	2.121	3.760	1.881	0.6642	0	1.843	0	3.267	1.128	0	3.686	0	3.267	0	3.686	0
3	1.732	3.394	1.023	0.7236	0	1.858	0	2.566	1.693	0	4.358	0	2.575	0	4.358	0
4	1.500	2.880	0.729	0.7979	0	1.808	0	2.269	2.059	0	4.698	0	2.282	0	4.698	0
5	1.342	1.596	0.577	0.8407	0	1.756	0	2.089	2.326	0	4.918	0	2.115	0	4.918	0
6	1.225	1.410	0.483	0.8686	0.026	1.711	0.030	1.970	2.534	0	5.078	0	2.004	0	5.078	0
7	1.134	1.277	0.419	0.8882	0.105	1.672	0.118	1.888	2.704	2.205	5.203	0.076	1.924	0.076	5.203	0
8	1.061	1.175	0.073	0.9027	0.167	1.638	0.185	1.815	2.847	0.387	5.307	0.136	1.864	0.136	5.307	0
9	1.000	1.094	0.037	0.9139	0.219	1.609	0.239	1.761	2.970	0.546	5.394	0.184	1.816	0.184	5.394	0
10	0.949	1.028	0.308	0.9227	0.262	1.584	0.284	1.716	3.078	0.687	5.469	0.223	1.777	0.223	5.469	0
11	0.905	0.973	0.285	0.9300	0.299	1.561	0.321	1.679	3.173	0.812	5.534	0.256	1.744	0.256	5.534	0
12	0.866	0.925	0.256	0.9359	0.331	1.541	0.354	1.646	3.258	0.924	5.592	0.284	1.716	0.284	5.592	0
13	0.832	0.883	0.249	0.9410	0.359	1.523	0.382	1.618	3.336	1.026	5.646	0.308	1.692	0.308	5.646	0
14	0.802	0.848	0.235	0.9453	0.384	1.507	0.406	1.594	3.407	1.121	5.693	0.329	1.671	0.329	5.693	0
15	0.775	0.816	0.223	0.9490	0.406	1.492	0.428	1.572	3.472	1.207	5.737	0.348	1.652	0.348	5.737	0
16	0.750	0.788	0.212	0.9523	0.427	1.478	0.448	1.542	3.532	1.285	5.779	0.365	1.636	0.365	5.779	0
17	0.728	0.762	0.203	0.9551	0.445	1.465	0.466	1.534	3.588	1.359	5.817	0.379	1.621	0.379	5.817	0
18	0.707	0.738	0.196	0.9576	0.461	1.454	0.482	1.518	3.640	1.426	5.854	0.404	1.608	0.404	5.854	0
19	0.688	0.717	0.187	0.9599	0.477	1.443	0.497	1.503	3.689	1.490	5.888	0.404	1.596	0.404	5.888	0
20	0.671	0.697	0.180	0.9619	0.491	1.433	0.510	1.490	3.735	1.548	5.922	0.414	1.585	0.414	5.922	0
21	0.655	0.670	0.173	0.9638	0.504	1.424	0.523	1.477	3.778	1.606	5.950	0.425	1.575	0.425	5.950	0
22	0.640	0.662	0.167	0.9655	0.516	1.415	0.534	1.466	3.819	1.659	5.979	0.434	1.566	0.434	5.979	0
23	0.626	0.647	0.162	0.9670	0.527	1.407	0.545	1.455	3.858	1.710	6.006	0.443	1.557	0.443	6.006	0
24	0.612	0.632	0.157	0.9684	0.538	1.399	0.555	1.445	3.895	1.759	6.031	0.452	1.548	0.452	6.031	0
25	0.600	0.619	0.153	0.9696	0.548	1.392	0.565	1.435	3.931	1.804	6.058	0.459	1.541	0.459	6.058	0

XIV. RANDOM NUMBERS

58941	72711	39408	91620	27963	96478	21559	19246	88097	44026
02349	71389	45608	60947	60775	73181	43264	56895	04232	59604
89210	44546	27174	27499	53523	63110	57106	20865	91683	80688
11826	91326	29664	01603	23156	89223	43429	95353	44662	59433
69810	17100	35066	00815	01552	06392	31437	70385	45863	75971
81060	33449	68055	83844	90942	74857	52418	68723	47830	63010
56135	80647	51404	06626	10042	93629	37609	57215	08409	81906
57361	65304	93258	56760	63348	24949	11839	29793	37457	59377
24548	56415	61927	64416	29934	00755	09418	14230	62887	92683
66504	02036	02922	63569	17906	38076	32135	19096	96970	75917
45068	05520	56321	22693	35089	07694	04252	23791	60249	83010
99717	01542	72990	43413	59744	44595	71326	91382	45114	20245
05394	61840	83089	09224	78530	33996	49965	04851	18280	14039
38155	42661	02363	67625	34683	95372	74733	63558	09665	22610
04319	04318	99387	86874	12549	38369	54952	91579	26023	81076
18134	90062	10761	54548	49505	52685	63903	13193	33905	66936
32012	42710	34650	73236	66167	21788	03581	40699	10396	81827
78101	44392	53767	15220	66319	72953	14071	59148	95154	72852
23469	42846	94810	16151	08029	50554	03891	38313	34016	18671
35342	56119	97190	43635	84249	61254	80993	55431	90793	62603
65846	18076	12415	30193	42777	85611	57635	51362	79907	77364
22184	33998	87436	37430	45246	11400	20986	43996	73112	88474
83668	66236	79665	88312	93047	12088	86937	70794	01041	74867
90083	70696	13558	98995	58159	04700	90443	13168	31553	67891
97765	27552	49617	51734	20819	70198	67906	00880	82899	66065
49988	13176	94219	88698	41755	56216	66832	17748	04963	54859
78257	86249	46134	51865	09836	73966	65711	41699	11732	17173
30946	22210	79302	40300	08852	27528	84648	79589	95295	72895
19468	76358	69203	02760	28625	70476	76410	32988	10194	94917
30806	80857	84383	78450	26245	91763	73117	33047	03577	62599
42163	68332	98851	50252	56911	62693	73817	98693	18728	94741
39249	51463	95963	07929	66728	47761	81472	44806	15592	71357
88717	29289	77360	09030	39605	87507	85446	51257	89555	75520
16767	57345	42285	56670	88445	85799	76200	21795	38894	58070
77516	96648	51868	48140	13583	94911	13318	64741	64336	95103
87192	66483	55649	36764	86132	12463	28385	94242	32065	45233
74078	64120	04643	14351	71381	26133	68269	65145	28152	39087
94119	20108	78101	81276	00835	63835	87174	42446	08882	27067
62180	27453	18567	55524	86088	00069	59254	24654	77371	26409
56199	05993	71201	78852	65889	32719	13758	23937	90740	16866
04994	09879	70337	11861	69032	51915	23510	32050	52052	24004
21725	43827	78862	67699	01009	07050	73324	06732	27510	33761
24365	37661	18956	50064	39500	17450	18030	63124	48061	59412
14762	69734	89150	93126	17700	94400	76075	08317	27324	72723
28387	99781	52977	01657	92602	41043	05686	15650	29970	95877

Source: Extracted from Table of 105,00 Random Decimal Digits Statement No. 4914, File No. 261-A (Washington D.C. Interstate Commerce Commission, 1949)