

BASIC STATISTICAL METHODS IN ECONOMICS

Edited By
Dr Pavitar Parkash Singh

Printed by
USI PUBLICATIONS
2/31, Nehru Enclave, Kalkaji Ext.,
New Delhi-110019
for
Lovely Professional University
Phagwara

SYLLABUS

Basic Statistical Methods in Economics

Objectives:

The course aims to equip the students with statistical tools and concepts that help in decision making. The emphasis is on their application in business.

Sr. No.	Content
1	Definition of Statistics: Importance and scope of statistics and its limitations, Types of data collection: Primary and Secondary: Methods of collecting Primary data, Classification and Tabulation of data: Frequency and cumulative frequency distribution
2	Central Tendency: Mean, Median and Mode and their Properties, Application of Mean, Median and Mode
3	Dispersion: Meaning and characteristics. Absolute and relative measures of dispersion including Range, Quartile deviation, Percentile, Mean deviation, Standard deviation, Skewness and Kurtosis: Karl Pearson, Bowley, Kelly's methods
4	Correlation: Definition, types and its application for Economists, Correlation: Scatter Diagram Method, Karl Pearson's coefficient of correlation, Rank correlation method
5	Linear Regression Analysis: Introduction and lines of Regression, Coefficient of regression method simple, Correlation analysis vs. Regression Analysis

CONTENTS

Unit 1:	Definition of Statistics, Importance and Scope of Statistics and its Limitations *Dilfraz Singh, Lovely Professional University*	1
Unit 2:	Types of Data Collection: Primary and Secondary, Methods of Collecting Primary Data *Pavitar Parkash Singh, Lovely Professional University*	8
Unit 3:	Classification and Tabulation of Data: Frequency and Cumulative Frequency Distribution *Pavitar Parkash Singh, Lovely Professional University*	17
Unit 4:	Central Tendency: Mean, Median and Mode and their Properties *Hitesh Jhanji, Lovely Professional University*	37
Unit 5:	Application of Mean, Median and Mode *Dilfraz Singh, Lovely Professional University*	48
Unit 6:	Dispersion: Meaning and Characteristics, Absolute and Relative Measures of Dispersion including Range, Quartile Deviation and Percentile *Pavitar Parkash Singh, Lovely Professional University*	72
Unit 7:	Mean Deviation and Standard Deviation *Dilfraz Singh, Lovely Professional University*	91
Unit 8:	Skewness and Kurtosis: Karl Pearson, Bowley, Kelly's Methods *Pavitar Parkash Singh, Lovely Professional University*	116
Unit 9:	Correlation: Definition, Types and its Application for Economists *Hitesh Jhanji, Lovely Professional University*	128
Unit 10:	Correlation: Scatter Diagram Method, Karl Pearson's Coefficient of Correlation *Pavitar Parkash Singh, Lovely Professional University*	147
Unit 11:	Rank Correlation Method *Dilfraz Singh, Lovely Professional University*	167
Unit 12:	Linear Regression Analysis: Introduction and Lines of Regression *Dilfraz Singh, Lovely Professional University*	176
Unit 13:	Coefficient of Simple Regression Method *Pavitar Parkash Singh, Lovely Professional University*	187

Dilfraz Singh, Lovely Professional University

Notes

Unit 1: Definition of Statistics, Importance and Scope of Statistics and Its Limitations

CONTENTS

Objectives

Introduction

1.1 Definition of Statistics

1.2 Importance and Scope of Statistics

1.3 Limitations of Statistics

1.4 Summary

1.5 Key-Words

1.6 Review Questions

1.7 Further Readings

Objectives

After reading this unit students will be able to:

- Know the Definition of Statistics.
- Discuss the Importance and Scope of Statistics.
- Explain the Limitations of Statistics

Introduction

In general sense the word Statistics means facts and figures of a particular phenomenon—Under reference in numerical numbers. In the traditional period the scope of Statistics was very much limited to the collection of facts and figures pertaining to the age-wise and sets-wise distribution of population, wealth etc. But now-a-days we can say that Statistics constitutes an integral part of every scientific and economic inquiry: Social and economic studies without Statistics are useless. Statistics thus play a vital role and as **Tippet** has rightly remarked, "It affects everybody and touches life at many points."

1.1 Definition of Statistics

It has been observed that the word 'Statistics' comes from Latin word 'Status' which means Political State. It has also been believed that the word Statistics comes from Italian word 'Stato'. This word was used in the fifteenth century for the 'State' in actual practice these words were used for Political State or Stateman's art.

Now-a-days Statisticians use statistics both in singular and plural sense. In the singular sense the term Statistics is associated with "A body of methods for making decisions when there is uncertainty arising from incompleteness or the instability of the information available for making such decisions." In its plural sense Statistics refers to numerical Statements of facts such as per capita income, population etc. Thus, some authorities have defined Statistics as Statical data (Plural sense) whereas other as Statistical method (Singular sense). According to **Oxford Concise Dictionary,** "Statistics—(as treated as plural): numerical facts, systematically collected, as Statistics of population, crime etc. (treated as singular): Science of collecting, classifying and using Statistics."

Notes

Definition

The definition of Statistics can be divided into the following two heads:

(A) In Plural Sense,

(B) In Singular Sense.

(A) In Plural Sense: The following are the definitions of Statistics in Plural Sense:

According to **H. Secrist—***"By Statistics we mean aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."*

In the words of **L. R. Connor—***"Statistics are measurements, enumeration or estimater of natural or social phenomena systematically arranged so as to exhibit their interrelations."*

According to **Yule & Kendall—***"By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes."*

In the opinion of **A. L. Bowley—***"Statistics are numerical statement of facts in any department of enquiry placed in relation to each other."*

According to **Webster—***"Classified facts, representing 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."*

On the basis of the above definitions the following characteristics are there in Statistics:

1. **Statistics are aggregate of facts:** Single and unconnected figures are not Statistics. A single age of 22 years or 37 years is not Statistics but a series relating to the ages of a group of people would be called Statistics. Likewise single figure relating to birth, death, sale, etc. cannot be called Statistics but aggregates of such figures would be Statistics because they can be studied in relation to each other and are capable of comparison.
2. **Statistics are affected to a marked extent by multiplicity of causes:** Usually facts and figures are affected, to a considerable extent, by a number of factors operating together. For example —Statistics of prices are affected by conditions of demand, supply, imports, exports, currency circulation, etc. and various other factors.
3. **Statistics are numerically expressed:** Qualitative expression like good, bad, young, old etc. do not form a part of statisical study unless numerical equivalent is assigned to such expression. If it is said that the production of rice per acre in 1997 was 30 quintals and in the year 2002 it was 50 quintals, we shall be making Statistical statements.
4. **Statistics are enumerated according to reasonable standard of accuracy:** Facts and figures relating to any subject can be derived in two way, example—by actual counting and measurement or by estimates. Estimates cannot be as accurate and precise as actual measurements. For example—If the heights of a group of people are being measured, it is right if the measurements are correct to a centimetre but if are measuring the distance from Agra to Gwalior, a difference of a few kilometres even, can be easily ignored.
5. **Collected in a systematic manner:** If Statistics are collected in a haphazard manner, it might fail to give the accurate result. It is, therefore, essential that statistics must be collected in a systematic manner so that they may *Conform* to reasonable standard of accuracy.
6. **Collected for a pre-determined purpose:** Statistical data are collected and processed for a definite and pre-determined purpose. In general, no data are collected without a pre-determine purpose.
7. **Placed in relation to each other:** The Statistics should be comparable. If they are not comparable, they lose part of their value and thus the efforts in collecting them may not prove to be as useful as the requirements may be. It is necessary that the figures which are collected should be a homogeneous so as to make them comparable and more useful.

   On the basis of the above description it may be said that numerical data cannot be called Statistics hence "All Statistics are numerical statements of facts but all numerical statements of facts may not essentially be Statistics."

(B) In Singular Sense: The following are the definitions of Statistics in Singular Sense. **Lovitt** defines the science as, "*That which deals with the collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomena.*"

Notes

According to **King,** "*The Science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection or estimation.*
According to **Croxton** and **Cowden,** "*Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data.*"
A. L. Bowley tried to define Statistics in this group also. He was of the opinion that, "*Statistics is the science of measurement of social organism, regarded as a whole in all its manifestations.*"
According to ***Seligman,*** "*Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting of numerical data collected to throw some light on any sphere of enquiry.*"
On the basis of the above definitions it may be said that Statistics are numerical statements of facts capable of analysis and interpretation and science of Statistics are a study of principles and the method used in the correction, presentation and analysis of numerical data in any sphere of enquiry.

1.2 Importance and Scope of Statistics

Scope of Statistics

The scope of statistics are concerned with the new dimensions in the definition of statistics. In other words we can say — Are statistics a science or an art or both ? Science is concerned with the systematised body of knowledge. It shows the relationship between cause and effects. So far as art is concerned, it refers to the skill of collecting and handling of data to draw logical inference and arrive at certain results. Statistics may be used as a science and as an art. In this regard the following definitions may be given:
As per **Netter** and **Wanerman —** "*Statistics methods are mathematical techniques used to facilitate the interpretation of numerical data secured from groups of individuals.*"
In the words of **Paden** and **Lindquist —** "*Statistical methods are mathematical techniques used to facilitate the interpretation of numerical data secured from groups of individuals.*"
According to **Kaney** and **Keeping —** "*Statistics has usually meant the science and art concerned with the collection, presentation and analysis of quantitative data so that intelligent judgement may be made upon them.*"
According to **Anderson** and **Bancraft —** "*Statistics is the science and art of the development and application of the most effective method of collecting, tabulating and interpreting quantitative data in such a manner that the fallibility of conclusions and estimates may be assessed by means of inductive reasoning based on mathematics of probability.*"


Did u know? "Statistics are the straw out of which, I like every other economist have to make the bricks."

Importance Or Significance of Statistics

The importance of statistics is now being felt in almost every field of study. In fact, it is difficult to mention a subject which does not have any relation with the science of statistics. As a matter of fact statistical methods are common ways of thinking and hence are used by all types of persons. Suppose a person wants to purchase a car and he goes through the price list of various companies and makers, to arrive at a decision, what he really aims at is to have an idea about the average level and the range within which the prices vary, though he may not know a word about these terms. No doubt to say that statistical methods are so closely connected with the human actions and behaviour that all human activity may be explained by statistical method. The importance of statistics can be shown in the following heads:
Importance in Economics: In the study of economics the use of statistical methods are of great importance. Most of the economic principles and doctrines are based on the study of a large number of units and their analysis. By statistical analysis we can study the ways in which people spend their

Notes

income over food, rent, clothing, entertainment, education etc. For example, if the law of demand is to be analysed then we have to make an idea about the effect of price changes on demand both for an individual and for a market. For this purpose a large number of data and figures would be collected. On the basis of the available information, the demand schedule can be prepared and then the law of demand can be formulated. We thus find that in the field of economics, the use of statistics is indispensable.

Importance in Business Management: Business managers are required to make decision in the face of uncertainty. Modern statistical tools of collection, classification, tabulation, analysis of interpretation of data have been found to aid in making wise decisions at various levels of managerial functions. These tools are relied upon in arriving at correct decision in all these aspects —sales forcasting, price situations, credit position, quality control, inventory control, investment planning, tax planning are some of the areas where statistical techniques help the business management in present and future planning. On the basis of the above discussion it is clear that the use of statistical data and techniques is indispensable in almost all the branches of business management.

Role of Statistics in planning: Today planning cannot be formulated without statistics. The problems like over production, unemployment, low rate of capital formation etc. which are the major characteristics of developing countries can be understood with the help of statistical data. National Sample Survey Scheme was started to collect statistical data for use in planning in India. Economic planning is done to achieve pre-determined objectives and goals. They have to be expressed in quantitative terms. We, thus, find that in the field of economic planning the use of statistics is indispensable.

Statistics in Commerce: Statistics plays a very important role in the development of commerce. The statistical data on some macro variables like income, investment, profits etc. are used for the compilation of national income. Economic barometer are the gifts of statistical methods and businessmen all over the world make extensive use of them. The increasing application of the statistical data and the statistical techniques in accountancy and auditing are supported by the inclusion of a compulsory paper on statistics both in the Charted Accountant's and the Cost Accountant's examinations. Various branches of commerce utilise the services of statistics in different forms. Cost Accounting is entirely statistical in its outlook and it is with the help of this technique that the manufacturers and the producers are in a position to decide about the prices of various commodities. We, thus, find that the science of statistics is of great importance to commerce.

Utility of Bankers, Brokers, Insurance Companies etc: Bankers, Stock Exchange Brokers, insurance companies, investors and public utility concerns all make extensive use of statistical data and technique. A banker has to make a statistical study of business cycles to forecast a probable boom. On the basis of this study a banker decides about the amount of reserves that should be kept.

Statistics are important from the view-point of stock exchange, brokers and investors. They have to be conversant with the prevailing money rate at various centres and have to study their future trends. Likewise insurance companies cannot carry on their business in the absence of statistical data relating to life tables and premium rates. As a matter of fact insurance has been one of the basic branches of commerce and business which has been making use of statistics.

Importance to State: Statistics are very important to a State as statistics help in administration. In all the fields where the State has to keep accurate records and information, statistical systems are adopted. For example, for making the economic plan the State has to collect data or information, it has to estimate the figures of National Income to find out the real position of the country. For this purpose, the state needs statistics for carrying on these works. The state also needs data about the roads, transport and communication, financial affairs, internal and external trade etc.

Importance to Research: Statistical methods and techniques happen to be useful in gathering the public opinion on various problems facing the society. In the field of Industry and Commerce statisticians carry on different types of researches. No researcher, without the use of statistics, can fulfil his targets. Today, the study of statistical method is not only useful but necessary for research. To conclude, statistical methods and techniques have been used in almost all the spheres. Statistical methods are essential to understand the effect to determining the factors of economic development

Notes

in the past, what psychological and sociological factor need to be developed for economic development and for the success of a plan.

1.3 Limitations of Statistics

Though statistics is an important instrument of quantitative method and research in social sciences, physical science and life sciences, it suffers from a number of limitations. The following are the main limitations of statistics:

(1) **Absence of uniformity:** In any statistical inquiry the data obtained are heterogeneous in nature. Statistical methods alone cannot bring in perfect uniformity. Generally results obtained need not be uniform and hence will serve no purpose.

(2) **Statistics does not study individuals:** Statistics deals only with aggregate of facts. Hence, single figures, however important they might be, cannot be taken up within the purview of statistics. For example, the marks obtained by X student of a class are not the subject-matter of statistics but the average marks has statistical relevance.

(3) **Statistical results speak about only average:** Prof. A. L. Bowley has rightly remarked that Statistics is a science of average. It implies that statistical results are true only on average. For example, if we say that per capita income in India is Rs. 12,000 per annum, it does not mean that the per capita income of the members of the Birla's family and the income of the poor fellows who sleep in the slum area are equal. Therefore, averages give only contradicting results.

(4) **Statistics can be misused:** Statistics is misused very often in the sense that a corrupt man can always prove all that he wants to do by using false statistics. In the words of **W. I. King,** "One of the shortcomings of statistics is that they do not, bear on their face the label of their quality."

(5) **Laws are not stable:** The statistical laws are obtained on the basis of information available at one stage need not be true at another stage. The basic data changes and hence the basic laws governing them also change. Moreover, what is applicable to India need not be true in Japan.

(6) **Statistics cannot be applied to qualitative statistics:** The Statistic studies cannot be applied to qualitative attributes like good, bad, beautiful etc. For a whole sum coverage, the statistical tools must be applicable for quantitative and qualitative data.

Self-Assessment

1. Fill in the blanks:

(*i*) The word statistics is used in ..... senses namely ..... and ..... .

(*ii*) The word statistics refers either ..... information or to a method of dealing with ..... information.

(*iii*) Any collection of related observations is called as .....

(*iv*) Applied statistics is divided into two groups, they are ..... and .....

(*v*) All the rules of procedures and general principles which are applicable to all kinds of groups of data are studied under .....

1.4 Summary

- In the traditional period the scope of Statistics was very much limited to the collection of facts and figures pertaining to the age-wise and sets-wise distribution of population, wealth etc. But now-a-days we can say that Statistics constitutes an integral part of every scientific and economic inquiry: Social and economic studies without Statistics are useless. Statistics thus play a vital role and as **Tippet** has rightly remarked, "It affects everybody and touches life at many points."
- Now-a-days Statisticians use statistics both in singular and plural sense. In the singular sense the term Statistics is associated with "A body of methods for making decisions when there is uncertainty arising from incompleteness or the instability of the information available for making such decisions." In its plural sense Statistics refers to numerical Statements of facts such as per capita income, population etc. Thus, some authorities have defined Statistics as Statical data (Plural sense) whereas other as Statistical method (Singular sense). According to **Oxford Concise**

Notes

Dictionary, "Statistics—(as treated as plural): numerical facts, systematically collected, as Statistics of population, crime etc. (treated as singular): Science of collecting, classifying and using Statistics."

- If Statistics are collected in a haphazard manner, it might fail to give the accurate result. It is, therefore, essential that statistics must be collected in a systematic manner so that they may *Conform* to reasonable standard of accuracy.
- The Statistics should be comparable. If they are not comparable, they lose part of their value and thus the efforts in collecting them may not prove to be as useful as the requirements may be. It is necessary that the figures which are collected should be a homogeneous so as to make them comparable and more useful.
- On the basis of the above description it may be said that numerical data cannot be called Statistics hence "All Statistics are numerical statements of facts but all numerical statements of facts may not essentially be Statistics."
- The scope of statistics are concerned with the new dimensions in the definition of statistics. In other words we can say—Are statistics a science or an art or both ? Science is concerned with the systematised body of knowledge. It shows the relationship between cause and effects. So far as art is concerned, it refers to the skill of collecting and handling of data to draw logical inference and arrive at certain results. Statistics may be used as a science and as an art. In this regard the following definitions may be given:
- The importance of statistics is now being felt in almost every field of study. In fact, it is difficult to mention a subject which does not have any relation with the science of statistics. **Alfred Marshall** had mentioned that, "Statistics are the straw out of which, I like every other economist have to make the bricks." As a matter of fact statistical methods are common ways of thinking and hence are used by all types of persons. Suppose a person wants to purchase a car and he goes through the price list of various companies and makers, to arrive at a decision, what he really aims at is to have an idea about the average level and the range within which the prices vary, though he may not know a word about these terms. No doubt to say that statistical methods are so closely connected with the human actions and behaviour that all human activity may be explained by statistical method. The importance of statistics can be shown in the following heads:
- In the study of economics the use of statistical methods are of great importance. Most of the economic principles and doctrines are based on the study of a large number of units and their analysis. By statistical analysis we can study the ways in which people spend their income over food, rent, clothing, entertainment, education etc. For example, if the law of demand is to be analysed then we have to make an idea about the effect of price changes on demand both for an individual and for a market. For this purpose a large number of data and figures would be collected. On the basis of the available information, the demand schedule can be prepared and then the law of demand can be formulated. We thus find that in the field of economics, the use of statistics is indispensable.
- Today planning cannot be formulated without statistics. The problems like over production, unemployment, low rate of capital formation etc. which are the major characteristics of developing countries can be understood with the help of statistical data. National Sample Survey Scheme was started to collect statistical data for use in planning in India. Economic planning is done to achieve pre-determined objectives and goals. They have to be expressed in quantitative terms. We, thus, find that in the field of economic planning the use of statistics is indispensable.
- Bankers, Stock Exchange Brokers, insurance companies, investors and public utility concerns all make extensive use of statistical data and technique. A banker has to make a statistical study of business cycles to forecast a probable boom. On the basis of this study a banker decides about the amount of reserves that should be kept.
- Statistics are very important to a State as statistics help in administration. In all the fields where the State has to keep accurate records and information, statistical systems are adopted. For example, for making the economic plan the State has to collect data or information, it has to estimate the figures of National Income to find out the real position of the country. For this

Notes

purpose, the state needs statistics for carrying on these works. The state also needs data about the roads, transport and communication, financial affairs, internal and external trade etc.

- To conclude, statistical methods and techniques have been used in almost all the spheres. Statistical methods are essential to understand the effect to determining the factors of economic development in the past, what psychological and sociological factor need to be developed for economic development and for the success of a plan.
- In any statistical inquiry the data obtained are heterogeneous in nature. Statistical methods alone cannot bring in perfect uniformity. Generally results obtained need not be uniform and hence will serve no purpose.
- Prof. A. L. Bowley has rightly remarked that Statistics is a science of average. It implies that statistical results are true only on average. For example, if we say that per capita income in India is Rs. 12,000 per annum, it does not mean that the per capita income of the members of the Birla's family and the income of the poor fellows who sleep in the slum area are equal. Therefore, averages give only contradicting results.

1.5 Key-Words

1. Statistics : Statistics is the study of the collection, organization, analysis, interpretation, and presentation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments.
2. Statistical method : A method of analyzing or representing statistical data; a procedure for calculating a statistic.

1.6 Review Questions

1. Define statistics and explain its importance.
2. Examine the important definitions of statistics. Which in your opinion, is the best ?
3. Discuss the scope of the study of this science.
4. Explain the use of statistics for economic analyses and planning.
5. Discuss the limitations of statistics.

Answers: Self-Assessment

1. (*i*) Two, singular, plural (*ii*) Quantitative, qualitative
   (*iii*) Data (*iv*) Descriptive and inductive statistics
   (*v*) Statistical methods.

1.7 Further Readings



Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 2: Types of Data Collection: Primary and Secondary, Methods of Collecting Primary Data

CONTENTS

Objectives
Introduction
2.1 Primary Data and Secondary Data
2.2 Methods of Collecting Primary Data
2.3 Summary
2.4 Key-Words
2.5 Review Questions
2.6 Further Readings

Objectives

After reading this unit students will be able to:

- Define Primary and Secondary Data.
- Discuss the Methods of Collecting Primary Data.

Introduction

On the basis of the method and source by which the data is collected the data is classified into two types: (1) Primary Data, (2) Secondary Data. Primary data are collected by the investigator first hand, for the first time and are, therefore, original. Secondary data, on the other hand, which have already been collected by some other people or agency may be for some other type of enquiry, in this way it can be seen that secondary data are primary for some. Primary data are in the shape of raw material whereas secondary data are in the shape of finished products.

2.1 Primary Data and Secondary Data

Pavitar Parkash Singh, Lovely Professional University

Primary Data

Primary data are those data which are collected directly from the individual respondents for the first time by the investigator for certain purpose of study. They are but the raw materials for an investigation. Primary data are original in character in the sense that they have been recorded as they occurred without having being grounded at all. They simply relate to the collection of original statistical information. They are also current and fresh.

For example, the data collected by National Sample Survey Organization (NSSO) and Central Statistical Organization (CSO) for various surveys are primary in character. If an experiment is conducted to know the effect of a drug on the patients, the observations taken on each patient constitute the primary data. Primary data are collected when fresh data are needed and also when no other statistical information are available.

Notes

Merits

Primary data are of the following merits:

1. They are more accurate.
2. They are reliable.
3. They are suitable for sampling inquiry.
4. They are original in character.
5. They are the latest or current information.

Demerits

Primary data are of the following demerits:

1. They are inconvenient.
2. They are expensive which involves a great amount of planning and supervision.
3. They consume more time and energy.
4. They may prove inaccurate if the enumerators are not trained well.
5. Large number of investigators are needed.
6. There may be personal bias and prejudices.
7. If the respondents are non-responsive, the quality of the collected information suffers a lot.

Secondary Data

Secondary data are those data which have already been collected by somebody for others for some other purposes. They are in the finished form of the investigation.

For example, if the statistical data given in different population census years are again processed to obtain trends of population growth, sex ratio, mortality rate, etc. it is termed secondary data.

Secondary data are collected when adequate and authentic statistical information are already available and when there is waste of time and money to collect fresh statistical information.

Merits

Secondary data are of the following merits:

1. They require only a minimum cost for the collection.
2. They can be used for a quick survey or investigation.
3. They save time, money and energy.
4. The errors occurred can be easily eliminated by the primary investigator.

Demerits

Secondary data are of the following demerits:

1. They may be outdated information.
2. They may have no accuracy of data.
3. They are only secondary in character.
4. In some cases, all particulars may not be available.



Did u know? Secondary data are secondary in character in the sense that those statistical information which have already been processed to a certain extent for a certain purpose. They are expressed in totals, averages or percentages.

Notes

Sources of Secondary Data

The main sources of secondary data are of two categories:

Published sources

Unpublished sources

Published Sources

The various government agencies, international bodies and local agencies generally publish sources of data which are secondary in character. The following are some of the important published sources of secondary data:

1. Official publications brought out by the Central, State and Local Governments such as Pay Commission Reports, Indian Population Census Reports, etc.
2. Official publications brought out by the international bodies like International Monetary Fund (IMF), International Bank for Reconstruction and Development (IBRD), United Nations Organization (UNO), International Labour Organization (ILO), etc.
3. Reserve Bank of India's publications of Bulletin, namely "RBI Bulletin" and the official records of the nationalized commercial banks and the State Bank of India.
4. Publications of Trade Unions, Chambers of Commerce and Industry such as Federation of Indian Chambers of Commerce and Industry, Trade Bulletins issued by Stock Exchanges and large business houses.
5. Publications brought out by individual research scholars, Research centres, reports submitted by economists and statistical organizations.
6. Publications brought out by the National Sample Survey Organization (NSSO) and Central Statistical Organization (CSO).

Unpublished Sources

The following are some of the unpublished sources of data which are secondary in character:

1. Records of government offices and business concerns such as account books.
2. Research undertaken, by research institutions, scholars, site.
3. Unpublished sources of data are also available with Trade Unions, Chambers of Commerce, Labour Bureaus, etc.
4. Central Bureau of Investigation (CBI) Records.
5. District Collectorate Office Records.

Precautions in Using Secondary Data

Secondary data should be used only after careful examination of the data. It is because of the fact that the secondary data may sometimes be unsuitable, inadequate, inaccurate, unreliable and incomparable to serve the purpose of the the present investigation or inquiry.

The following are some of the precautions in using secondary data:

1. The suitability of the data available for the purpose of the present inquiry should be ascertained.
2. The adequacy of the data available for the present analysis should be ascertained.
3. The degree of accuracy desired and actually achieved should also be taken into consideration.
4. The degree of reliability of secondary data is to be assessed from the source, the compiler and his capacity to obtain current statistics for the purpose of interpretation.
5. The question of comparability of the data over a period of time should be assured.
6. The secondary data should be used only when the scope and object of the present inquiry are commensurate with that of the original inquiry.

Notes

7. The definition of units in which data are expressed should be kept stable in the present inquiry also.
8. While using secondary data, general information should also be obtained regarding the type of investigators, editors and tabulators employed in the primary data collection method.
9. The sources and the methods used should be clearly known.

2.2 Methods of Collecting Primary Data

There are various methods of collecting primary data. They are:

(1) Direct Personal Investigation: As the name itself suggests, in this method information is collected personally from the sources concerned. Under this, the investigator personally interviews everyone who is in a position to supply the information he requires. The investigator must possess the following qualities so that genuine data may be collected. *Firstly*, the investigator must be totally unbiased. *Secondly*, the investigator must have the tact of sustaining calm and unflustered enquiring environment so that the truth can be revealed. *Thirdly*, the investigator must have a sociable and pleasing personality so that the interviewers do not run away from him. Fourthly, the interviewer should be neutral in matters of clan, sex, race, religion etc.

Merits of this method: (i) Original data are collected, (ii) Correct and required information is gathered, (iii) The personal presence of the investigator helps in keeping the flexibility in the enquiry depending upon the type of respondent, (iv) Problem of no-answers is solved to a great extent *i.e.*, reluctance of response due to not understanding the question etc. can be avoided. (v) Uniformity in data collection is maintained.

Demerits of this method: (i) Requires a lot of time and personal presence of investigator every time, (ii) The method is very costly due the above reason, (iii) Cannot be used ideally in cases of wide investigating field, (iv) Too much dependence on the skills of investigator, (v) Not suitable when respondents are reluctant to reveal the truth when approached directly.

This method of direct personal investigation is suitable only for intensive investigations.

(2) Indirect Oral Investigation: Sometimes, the requirement of information is such that people are reluctant to answer when approached directly, and sometimes it is not possible for the investigator to be present personally with each respondent, it such is case, method of indirect oral in investigation can be used to collect the primary data. However, in order to ensure genuine data, it is essential that only those persons should be interviewed who: (i) possess full knowledge, (ii) are capable of expressing themselves, (iii) are not prejudiced, (iv) are rational.

In this method, a small list of questions is prepared and is put to different people (known as witnesses) and their answers are recorded. There should be no motivation to give colours to the facts.

Merits of this Method: (i) A wide field may be brought under investigation, (ii) This method saves time and labour and hence is less costly, (iii) There is no need to depend too much on the personal skills of the investigator, (iv) It is easier to deal with all aspects of the problem.

Demerits of this Method: (i) Great care and vigilance is needed in assessing the correct value of information collected, (ii) Due allowance needs to be made for the conscious and unconscious bias of the persons giving information, (iii) There is a possibility that the witnesses colour the information according to their interests, (iv) Dependence of local correspondents increases when the field of enquiry is wide.

(3) Local Reports: This method is generally used by the news-papers, periodicals, news channels etc. The government also collects information about prices, agricultural production etc. by this method.

Merits of this Method: (i) A very wide geographical area can be covered, (ii) Information on specific issues can be obtained, (iii) Regular flow of information over a long period of time can be obtained.

Notes

Demerits of this Method: (i) Reliability of the information is questionable, (ii) Correspondent's personal bias may come in, (iii) Same information with different attitudes may look different, (iv) Chances of errors are many, as the correspondents are not personally interested in the problem.

(4) Schedules and Questionnaires Method: This method is usually used by private individuals, research workers, non-official institutions and even the Government. In this method, a list of questions is prepared and information is collected about the problem. This can be done by —

(a) distributing questionnaires to the persons knowing about the information.

(b) sending the information by post or e-mail to the people from whom information is to be collected.

(c) using enumerators to send the questionnaire to the people, who also help them in filling the questionnaire. This method is adopted when there is a possibility of language problem or the respondent is illiterate or when there is probability of avoidance of answering by the respondents.

Merits of this Method: (i) Wide area of investigation can be covered, (ii) The Method is simple and cheap, (iii) Can be used with least expenses for geographically dispersed respondents, (iv) Original data is collected, (v) Information is given by the respondents themselves, hence the data is free from the bias of the investigation.

Demerits of this Method: (i) Possibility of no-response is quite high, (ii) This method can be used successfully where the respondents are educated, (iii) Information given by the respondents may be false, (iv) Clarification of the questions, supplementary and complimentary questions etc. is not possible, hence the method is inflexible.

From the above, it is quite clear that none of the methods is free from one or the other drawback. In fact, the method to be chosen depends upon the nature of investigation, object and scope of enquiry, budget made for the purpose of data collection, degree of accuracy desired and the time within which the data has to be collected.

Questionnaire Method Or Essentials of a Good Questionnaire

The questionnaire method is the method in which primary data is collected by distributing a list of questions related to the probe to those who are supposed to have knowledge about the problem. In this way, it can be said that, the success of the Statistical enquiry in this case depends, to a large extent on the questionnaire.

Preparation of questionnaire is a highly specialised job. Although, there are no hard and fast set rules to prepare a questionnaire. However, a few broad principles should be followed in order to have a good questionnaire. They are:

1. The questionnaire should be started with a covering letter which should be written in a polite language requesting the respondents to answer to the best of their knowledge. The letter should emphasize the need and usefulness of the information that is being collected. The letter should also ensure that the information obtained from them shall be strictly used only for the said purpose and the information and name of the respondents shall be kept confidential.

   The covering letter may also accompany some small gift etc. to create the acceptance among the respondents so that there is a greater chance of getting a response. Moreover, the letter may also give a promise, that if the respondents so desire, a copy of the results of the survey may be sent to them. This would increase the credibility of the investigator/investigating institution.

2. The questionnaire should not be very long. Unnecessary details in the form of separate questions must be avoided. Although, there is no hard and fast rule about what should be the number of questions in a questionnaire. Much shall depend upon the problem undertaken. However, efforts should be made to frame only relevant questions otherwise, the respondents feel bored or feel answering them to be a waste of time, and correct information will be a casualty.

Notes

3. The questions should be framed in a simple way and in easy language. They should be capable of a straight answer. As far as possible, the questions should be capable of objective answers. A set of possible answers may be accompanied with each question so that the respondents feel easy to give the answers. Questions with 'yes' or 'no' answers are also useful.
4. Asking personal questions should be avoided because it is quite likely that these questions are not answered correctly. For example, perks received, income tax paid etc.
5. Questions which tend to hurt the sentiments of the respondents must be avoided. For example, private-life litigation, indebtness, etc.

   In both these cases it is quite likely that there will be either no-response or the response shall be false.
6. Corroboratory questions should be incorporated in a good questionnaire. These are the questions which are meant for cross-checking the answers given by the respondents for earlier question.
7. Unless and until it is very essential, questions whose answering requires calculations should not be asked. For example, what per cent of your income is spent on your children's schooling will need a series of calculation. And it is quite likely that the respondents does not give an accurate answer.
8. The questionnaire should be attractive and impressive. There should be sufficient space for answering the questions, the quality of paper used and printing on the paper should be good. It always helps if it is so.

Which method is the best in Collecting Primary Data ?

None of the methods can be termed to be best or worst. Following considerations are taken into account while selecting the method which should be used to collect primary data —

(1) **The nature of investigation:** If it is essential to establish personal contacts, 'direct personal investigation' will be appropriate. But if the number of respondents is large and they are educated also, questionnaire method shall be better. But if the area covered is very wide and information is to be gathered on a number of subjects, using enumerators shall be better.

(2) **Object and scope of enquiry:** If the scope of enquiry is limited and is of confidential nature, 'direct personal investigation' should be done. But if the scope extends to a number of subjects, use of questionnaire or enumerators can be made.

(3) **Budget:** If financial resources are strong, personal investigation can be carried out. But if a wide survey is to be done with limited financial resources, questionnaire method should be chosen.

(4) **Degree of accuracy desired:** Highest degree of accuracy is achieved from direct personal investigation and the accuracy is least in case of information collected from correspondents. Now, on the basis of budget and other above requirements, the method can be chosen.

(5) **Time factor:** A large amount of information can be obtained in minimum time by using enumerators and/or correspondents. If there is long-time available, 'direct personal investigation' may be done.

Self-Assessment

1. Fill in the blanks:

(*i*) Secondary data may be ..... or .....

(*ii*) Before finalising a questionnaire ..... is done.

(*iii*) ..... method is the cheapest method of collecting primary data.

(*iv*) In ..... method, much depends upon the skills of the investigator.

(*v*) Least accuracy of information is likely in case of information from the ..... .

Notes

2.3 Summary

- Primary data are those data which are collected directly from the individual respondents for the first time by the investigator for certain purpose of study. They are but the raw materials for an investigation.
- Primary data are original in character in the sense that they have been recorded as they occurred without having being grounded at all. They simply relate to the collection of original statistical information. They are also current and fresh.
- Secondary data are those data which have already been collected by somebody for others for some other purposes. They are in the finished form of the investigation.
- Secondary data are secondary in character in the sense that those statistical information which have already been processed to a certain extent for a certain purpose. They are expressed in totals, averages or percentages.
- Secondary data are collected when adequate and authentic statistical information are already available and when there is waste of time and money to collect fresh statistical information.
- Secondary data should be used only after careful examination of the data. It is because of the fact that the secondary data may sometimes be unsuitable, inadequate, inaccurate, unreliable and incomparable to serve the purpose of the the present investigation or inquiry.
- While using secondary data, general information should also be obtained regarding the type of investigators, editors and tabulators employed in the primary data collection method.
- As the name itself suggests, in this method information is collected personally from the sources concerned. Under this, the investigator personally interviews everyone who is in a position to supply the information he requires. The investigator must possess the following qualities so that genuine data may be collected. *Firstly*, the investigator must be totally unbiased. *Secondly*, the investigator must have the tact of sustaining calm and unflustered enquiring environment so that the truth can be revealed. *Thirdly*, the investigator must have a sociable and pleasing personality so that the interviewers do not run away from him. Fourthly, the interviewer should be neutral in matters of clan, sex, race, religion etc.
- Sometimes, the requirement of information is such that people are reluctant to answer when approached directly, and sometimes it is not possible for the investigator to be present personally with each respondent, it such is case, method of indirect oral in investigation can be used to collect the primary data.
- This method is usually used by private individuals, research workers, non-official institutions and even the Government. In this method, a list of questions is prepared and information is collected about the problem.
- Using enumerators to send the questionnaire to the people, who also help them in filling the questionnaire. This method is adopted when there is a possibility of language problem or the respondent is illiterate or when there is probability of avoidance of answering by the respondents.
- The questionnaire method is the method in which primary data is collected by distributing a list of questions related to the probe to those who are supposed to have knowledge about the problem. In this way, it can be said that, the success of the Statistical enquiry in this case depends, to a large extent on the questionnaire.
- The questionnaire should be started with a covering letter which should be written in a polite language requesting the respondents to answer to the best of their knowledge. The letter should emphasize the need and usefulness of the information that is being collected. The letter should also ensure that the information obtained from them shall be strictly used only for the said purpose and the information and name of the respondents shall be kept confidential.

Notes

- The covering letter may also accompany some small gift etc. to create the acceptance among the respondents so that there is a greater chance of getting a response. Moreover, the letter may also give a promise, that if the respondents so desire, a copy of the results of the survey may be sent to them. This would increase the credibility of the investigator/investigating institution.
- The questions should be framed in a simple way and in easy language. They should be capable of a straight answer. As far as possible, the questions should be capable of objective answers. A set of possible answers may be accompanied with each question so that the respondents feel easy to give the answers. Questions with 'yes' or 'no' answers are also useful.
- The questionnaire should be attractive and impressive. There should be sufficient space for answering the questions, the quality of paper used and printing on the paper should be good. It always helps if it is so.
- If it is essential to establish personal contacts, 'direct personal investigation' will be appropriate. But if the number of respondents is large and they are educated also, questionnaire method shall be better. But if the area covered is very wide and information is to be gathered on a number of subjects, using enumerators shall be better.
- Highest degree of accuracy is achieved from direct personal investigation and the accuracy is least in case of information collected from correspondents. Now, on the basis of budget and other above requirements, the method can be chosen.

2.4 Key-Words

1. Primary Data : Primary research consists of a collection of original primary data. It is often undertaken after the researcher has gained some insight into the issue by reviewing secondary research or by analyzing previously collected primary data. It can be accomplished through various methods, including questionnaires and telephone interviews in market research, or experiments and direct observations in the physical sciences, amongst others.
2. Secondary Data : Secondary data, is data collected by someone other than the user. Common sources of secondary data for social science include censuses, organisational records and data collected through qualitative methodologies or qualitative research. Primary data, by contrast, are collected by the investigator conducting the research.
3. Secondary Data : analysis saves time that would otherwise be spent collecting data and, particularly in the case of quantitative data, provides larger and higher-quality databases that would be unfeasible for any individual researcher to collect on their own. In addition, analysts of social and economic change consider secondary data essential, since it is impossible to conduct a new survey that can adequately capture past change and/or developments.

2.5 Review Questions

1. What preliminary steps ought to be taken by a statistician before starting on with the task of collection of data ?
2. To make collection of data smooth and result-oriented, it is essential for a statistician to carry out some preliminary steps. Justify the statement giving details about these steps.
3. Describe the questionnaire method of collecting Primary data. State the essentials of a good questionnaire.
4. What are the essentials of a good questionnaire ?
5. What are the various methods of collecting statistical data ? Which of these is most suitable and why ?

Notes

Answers: Self-Assessment

1. (*i*) published and unpublished, (*ii*) pre-testing,
   (*iii*) Questionnaire, (*iv*) Direct personal investigation,
   (*v*) correspondents.

2.6 Further Readings


Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Pavitar Parkash Singh, Lovely Professional University

Notes

Unit 3: Classification and Tabulation of Data: Frequency and Cumulative Frequency Distribution

CONTENTS

Objectives
Introduction
3.1 Classification
3.2 Tabulation of Data
3.3 Frequency Distribution
3.4 Cumulative Frequency Distribution
3.5 Summary
3.6 Key-Words
3.7 Review Questions
3.8 Further Readings

Objectives

After reading this unit students will be able to:

- Describe the Classification and Tabulation of Data.
- Understand Frequency Distribution.
- Explain Cumulative Frequency Distribution.

Introduction

The data collection leaves an investigator with a large mass of information. But it is the weakness of human mind that it fails to assimilate a lot of things or information at a time. To remove this difficulty and to make the large mass of data useful to its fullest, classification and tabulation of data is done. By doing so the data are presented in condensed form which helps in making comparisons, analysis and interpretations. Moreover, classification and tabulation segregates the likes from the unlikes. The heterogeneity is removed. The data are classified into classes and sub-classes according to their characteristics. This process is called **classification**. The classified data are presented in prescise and systematic tables. This process is called **tabulation**. By these two processes, the data collected are made simple, easy to understand and carry out analysis and interpretations.

3.1 Classification

Meaning and Definition of Classification

Classification may be defined as the process of arranging the available data in various groups or classes in accordance with their resemblances and similarities and keeping in view some common features and objectives of study. Thus, through classification, an effort is made to achieve homogeneity of the collected information. While classifying, the units with common characteristics are placed together and in this way the whole data is divided into a number of classes and sub-classes. It may be argued that the data collected is as per the requirement, and it is in general homogeneous in nature, then how does classification help ? For example, if a study on 500 students is to be carried out then, the data is homogeneous as it is about students. But this information on 500 students may be classified in terms of different hostels and universities they are coming from, different areas they come from, different subjects they have opted for, and so on. Only by carrying out the one, classification, will the invesitgator be in a position to compare, analyse and interpret the above data.

Notes

Definition

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

From the above definitions it may be said that a group or class has to be determined on the nature of the data and the purpose for which it is going to be used. For example, the data on household may be classified on the basis of age, income, education, occupation, expenditure etc.



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

Features of Classification

On the basis of the above dicussion, the chief features of classification can be summarised as under— (i) Classification may be according to attributes, characteristics or measures, (ii) The basis of classification is unity in diversity, (iii) Classification may be actual as notional.

Chief Objects of Classification

Classification helps the investigator and the investigation in a number of ways. The following are the objects of classification. These objects also suggest the *importance of classification*.

(1) **Classification presents the facts in simple form:** The process of classification helps in arranging the data in such a way that the large mass of irrelevant looking data becomes simple and easy to understand, avoiding unnecessary details, making logical sense.

(2) **Classification points out similarities and dissimilarities clearly:** Since classification is done on the basis of characteristics and similarity of data, it helps the investigator in poining out clearly the similarities and dissimilarities so that they can be easily grasped.

(3) **Classification facilitates comparison:** By classification of data, comparison becomes easier, inferences can be drawn logically and confidently and facts can be located with much ease.

(4) **Classification brings out relationship:** The cause—and effect relationship can be located with the help of classification.

(5) **Classification prepares basis for tabulation:** The importance of classification also lies in the fact that it prepares the ground on the basis of which tabulation can be done.

Methods of Classification

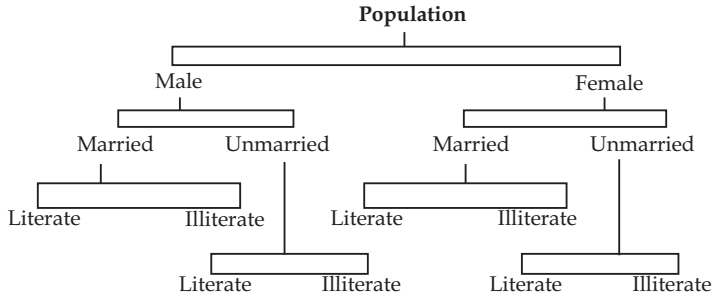
The methods of classification are divided broadly into four types:

(I) **Qualitative Classification:** Here, classification is done in accordance with the attributes or characteristics of the data. Such classification is generally done where data cannot be measured. Under this classification method, the presence or absence of an attribute is the basis of classification. Qualitative classification can be done in two ways:

(1) **Two-fold or Dichomous Classification:** This type of classification is based on the presence or absence of an attribute and the data gets classified in two groups/classes—*one*, possessing that attribute and *two*, not possessing that attribute. For example, on the basis of marital status, the data can be divided into *two* classes, one, married and two, unmarried. On the basis of literacy there can be two classes, one, literate other non-literate.

(2) **Manifold Classification:** Here, the bases of classification are manifold, *i.e.*, more than one attribute. In this method, classes/groups are further classified into sub-classes and sub-groups. For example, population/sample is first classified on the basis of sex, then for each sex (male or female) marital status forms two sub-classes then further these sub-classes are classified as per their literacy state. This can be explained in simple was as below:

Notes



(II) Quantitative Classification: In this method the data are classified on the basis of variables which can be measured for example, age, income, height etc. This kind of classification is done in the form of statistical series. For example, 100 students can be classified in terms of mark obtained by them. This is shown as below:

Marks Obtained	Number of Students
0-20	15
20-40	25
40-60	35
60-80	15
80-100	10

(III) Geographical Classification: In this case, data are classified on the basis of place or location. For example, population is shown on the basis of various states, or students are classified on the basis of the place they belong to etc. Series which are arranged on the basis of place or location are called spatial series.

Name of State	Per capita Income
Punjab	331
Kerala	310
Madhya Pradesh	206
J & K	216
Haryana	320

(IV) Chronological Classification: This is done by arranging the data with respect to time, such series are known as time series. For example,

Year	Sales made (in Rs. crore) by Company 'X'
1995	11,800
1996	12,600
1997	11,200
1998	16,800
1999	17,000
2000	18,200
2001	19,100
2002	20,000
2003	21,000

Notes

2004	21,800
2005	22,200
2006	23,000
2007	25,200
2008	24,600
2009	26,000

On the basis of the above, it may be concluded that depending upon the nature of data, and requirement of the investigation and objectives of study, classification of data facilitates the investigator to compare, analyse and interpret the data, thus helping in using the data scientifically.

3.2 Tabulation of Data

The process of presenting data in the tabular form is termed as tabulation. As per **L. R. Connor**, "Tabulation involves the orderly and systematic presentation of numerical data in a form designed to elucidate the problem under consideration.

Importance of Tabulation

(1) **Simplifies the complex data:** The process of tabulation eliminates unnecessary details and present the complex data concisely in rows and columns. This helps in simplifying the complex data which becomes more meaningful and better understood.

(2) **Presents facts in minimum space:** A large number of facts can be condensed in one table in a much better way than otherwise.

(3) **Facilitates comparison:** Data when depicted in rows and columns, facilitates comparison, and the problem can be better understood.

(4) **Depicts data characteristics:** The important characteristics of data are brought about by the process of tabulation as it is presented concisely but clearly.

(5) **Depicts trends and pattern of data:** Data, in the form of tables, helps in understanding the trends and patterns lying within the figures without much effort. This facilitates better understanding of the problem under study.

(6) **Helps in making references:** Data can be stored perfectly in the form of tables which can be easily identified by its head and footnotes. This can be used for future studies.

(7) **Facilitates statistical analysis:** It is only possible after tabulation, that the data can be subjected to statistical analysis and interpretation. Measures of correlation, regression dispersion etc can be easily calculated when the data is in tabular form.

The above points form the advantages of tabulations to the investigator and investigation as well.

Limitations of Tabulation

Although tabulation is an essential activity in the process of statistical analysis, it is not absolutely free of limitations. The limitations are:

(1) A table does not present any description about the figures expressed. For those who are not familiar, it is not easy to understand facts with the help of tables.

(2) Specialised knowledge is essential to understand a table. It is not a layman's cup of tea.

(3) A table does not lay emphasis on any section of particular importance.

It is because of these limitations that tables are only complementary to textual report. A table only accompanies a text, facilitating better understanding in a concise way.

Essential Parts of a Table

Notes

In order to be complete and most informative, a table should have the following parts:

(1) **Table number:** This is the number by which, the table can be identified and can be used for reference in future. The table number can be given at the top or at the bottom.

(2) **Title of the table:** A title is a brief statement indicating about the nature of the data and the time-span to which the data relate. The geographical distribution of the data, if any, should also be indicated in the title. The title should be in dark letters in comparison to other heads and sub-heads in the table.

(3) **Captions:** A caption is the main heading of the vertical columns. A number of small sub-headings are followed by the caption. Caption should be given in unambiguous language and should be placed at the middle of the column (as shown in the table given below).

(4) **Stubs:** The stubs are the headings of the horizontal rows and are written on the extreme left of the table. The number of stubs depend upon the nature of the data.

(5) **Head note:** The head note refers to the data contained in the major part of the table, and it is placed below the title of the table. It is generally put in brackets. For example, (in percentage) or (in *kg*.) etc.

(6) **Footnote:** They are given at the bottom of the table and are used to clarify about heading, title, stubs and caption etc. It may also be used to give further explanation about the data, or certain terms or figures used in the table. Footnotes are also used to describe the source of the data, if the data is secondary in nature. For example, "the figures in bracket show the per cent rise over previous year" or "the profit above is after tax" etc are some of the information which is given in footnotes.

(7) **Body of the table:** This is the most important and inevitable part of the table which contains the statistical data which have to be presented. The data is arranged in captions and stubs.

Table No.

Title of the Table

Head note

Sub-heading	⟵ Caption ⟶			
	Col. head	Col. ead	Col. head	Col. head
↑ Stub-entires ↓	B	ODY		

Footnotes

Source.

From the above, it becomes very clear that it is only with the help of tabulation, that statistical enquiry can be scientifically carried out. It helps the investigation and the investigator.

A table presents the statistical data in a systematic way in rows and columns which concisely explain the numerical facts. Tabulation is nothing but the process of preparing a table. The preparation of table is a specialised activity and is done through a set of rules.

Rules for Tabulation

Although there are no hard and fast rules regarding preparation of tables as pointed out by Bowley, common sense and experience are a prerequisite while tabulation, however, some general rules may prove to be handy while carrying out tabulation process. According to **Harry Jerome**, "A good statistical table is not mere careless grouping of columns and rows of figures: it is a triumph of

Notes

ingenuity and technique, a masterpiece of economy of space combined with a maximum of clearly presented information. According to **W. M. Harpen**, "The construction of a table is in many ways a work of art." The rules regarding tabulation are not hard and fast, but prove as a guide in tabulation. These rules are divided into two groups:

(A) Rules relating to Table Structure: This includes the rules explaining the preparation of structure of the table. This group includes the following rules:

(1) Table must always have a table number, so that it can be easily identified or can be referred to whenever required.

(2) The table must always have a relevant title, which should be clear, concise and self-explanatory. The title should explain about: (a) the subject area, (b) data or period to which the data belong to, (c) basis and principles used in classification of data, (d) the field to which the data relate to.

(3) The stub and caption should be clear and as brief as possible. Columns should be numbered.

(4) Neat and tidy appearance must be given to the table which can be done by providing proper ruling and spacing as is necessary. If the table continues to the next page, no bottom line must be drawn, as it would indicate the end of the table. Major and minor items must be given space according to their relative importance. Coloured inks, heavy printed titles or sub-titles, thick and thin ruling etc must be used to clarify a complex table.

(5) Use of averages, sub-totals, totals etc must be made if the data so require. In case, it is required, the table should contain sub-totals for each separate classification of data and a general total for all combined classes. For example, data about cost break-up of a particular production process shall require the above. But the data on annual percentage rise in bonus of employees may not require use of these sub-totals and totals.

(6) The body of the table must be as comprehensive as possible, consistent with the purpose. Unnecessary details must be avoided and, items in 'miscellaneous or unclassified columns must be least.

(7) The items in the body of the table must be arranged in some systematic order. Depending upon the type of data and purpose of enquiry, data may be arranged: (a) alphabetically, (b) geographically, (c) progressively, (d) chronologically, (e) ascending or descending order, (f) conventionally or (g) in order of importance etc.

(8) If further clarification about some figures, sub-heads etc is required, it must be given in the footnotes. The important limitation of data can also be specified here if the need is felt by the person who is tabulating.

(9) The source of the data must be specified if the data is secondary.

(10) The units of measurement, if common, must be indicated in the head note, otherwise under each heading and sub-heading.

(B) General Rules: Other than the rules relating to table structure, there are certain rules which should be followed while tabulation, so that, the tabulation can be accomplished successfully. They are:

(1) Table should be precise and easy to understand.

(2) If the data are very large they should not be crowded in a single table. However, it is essential that each table is complete in itself.

(3) The table should suit the size of the paper. The width of the columns should be pre-decided giving due consideration to this.

(4) Those columns whose data are likely to be compared should be preferably kept side-by-side.

(5) Percentages, averages, totals etc must be kept close to the data.

(6) The figures must be approximated to one or two decimals. This must be specified in the footnote.

Notes

(7) 'Zero' quantity must be indicated separately, and in case of unavailability of a particular figure 'NA' (not available) must be indicated clearly. 'Zero' is not equivalent to 'Not available'.

(8) Abbreviations should be avoided. But if the need so arises, it must be clarified in footnotes.

(9) The tabulation should be explicit. Words like 'etc'. must not be used.

(10) The table should be of manageable size.



Notes Tabulation is an art which requires common sense in planning a table and viewing the proposed table from the point of view of the user or the other person. Tabulation is done keeping in mind the purpose of statistical investigation. The rules of tabulation act as guides in preparing a good table.

Seriations of Data

Quantitative classification data is done through seriation of data. If two variable quantities are arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other, the result is formation of a statistical series. For example, marks obtained by a class of students. Here, there are two elements one, the variable (marks) and *two*, the frequency (of students). The number of times a particular variable has repeated is noted down and the total is the frequency of that class. For example, the marks obtained by 25 students out of 10 in a particular subject is as follows:

Marks Obtained

2	4	8	6	8
4	10	6	4	4
8	2	4	6	10
2	6	10	2	6
6	2	6	2	2

On counting how many students obtained 2, 4, 6, 8 and 10 we get the frequencies:

Marks	Tallies	Frequency
2	𝍸 ||	7
4	𝍸	5
6	𝍸 ||	7
8	|||	3
10	|||	3
Total		**25**

Notes

So the discrete data of marks obtained by 25 students is:

Marks Obtained	Number of Students
2	7
4	5
6	7
8	3
10	3

Formation of Continuous Frequency Distribution

In continuous frequency distribution data are divided into class intervals instead of individual values (as is done in case of discrete frequency distribution) class intervals can be formulated like marks from 0 to 10, 10 – 20 and so on. Here the magnitude of class interval is 10 marks. (20 – 10 = 10). But it can be lower, for example, 0 – 5, 5 – 10 and so on or higher, like 0 – 20, 20 – 40 and so on.

Suppose, the marks obtained of 100 students is to be given in continuous series, it can be done as below:

Marks Obtained	Number of Students
0–10	14
10–20	26
20–30	30
30–40	20
40–50	10
	100

This means 14 students obtained marks between '0' (zero) to 10. The marks may lie in any fraction 1/4, $3\frac{3}{4}$, $9\frac{3}{4}$, 9.99 or 10, and likewise. These are called exclusive class intervals.

Sometimes, the data is given as below:

Marks Obtained	Number of Students
10–19	4
20–29	6
30–39	10
40–49	10
	30

This is called inclusive class arrangement. In this case, a question arises as to in which class the student getting 9.5 or 29.5 must be placed ? In such a case to ensure continuity following adjustments must be made:

Marks obtained	Number of Students
9.5–19.5	4
19.5–29.5	6
29.5–39.5	10
39.5–49.5	10
	30

Notes

Difference between Classification and Tabulation

Some important points or differences between classification and tabulation are:

(1) Although both are necessary in statistical investigations, classification is done first and it forms the basis for tabulation.

(2) Tabulation is a mechanical function of classification. But classification is not a mechanical function.

(3) In classification, data are divided/classified in different classes as per similarities and dissimilarities. Under tabulation this classified data is put in rows and columns.

(4) Classification involves analysis of data. Tabulation is the process of presenting the data.

Types of Tables

Following are the important types of tables:

(1) One-way Table: This supplies information about only one characteristic. For example, marks obtained by 100 students can be illustrated in one-way table as below:

Marks Obtained	Number of Students
0–10	14
10–20	26
20–30	30
30–40	20
40–50	10
	100

(2) Two-way Table: If the information of two related characteristics is to be given, it is done by two-way tables. Suppose, in the above example, the students are to be classified in males and females the data can be re-written as below:

Marks Obtained	Number of Students		
	Male	Female	Total
0–10	4	10	14
10–20	16	10	26
20–30	15	15	30
30–40	12	8	20
40–50	6	4	10
Total	**53**	**47**	**100**

(3) Three-way Table: In the above example, the male and female students can be further divided into hostellers and day-scholars, thus providing information about three different characteristics. This is done by a three-way table. The format of the three-way table in this case is given below:

Notes

Number of Students									
Marks		**Males**			**Females**			**Total**	
	Hostellers	Day-scholars	Total	Hostellers	Day-scholars	Total	Hostellers	Day-scholars	Total
0–10	3	1	4	6	4	10	9	5	14
10–20	10	6	16	6	4	10	16	10	26
20–30	9	6	15	10	5	15	19	11	30
30–40	5	7	12	3	5	8	8	12	20
40–50	2	4	6	1	3	4	3	7	10
Total	**29**	**24**	**53**	**26**	**21**	**47**	**55**	**45**	**100**

(4) Higher-order Tables: Tables giving information about more than three characteristics can also be made. They are called higher-order tables. Suppose, in the above example, marital status of the students is also to be informed, the table will have to be made as below:

Marks	**Number of Students**								
	Males			**Females**			**Total**		
	Married	Unmarried	Total	Married	Unmarried	Total	Married	Unmarried	Total
Hostellers									
0–10									
10–20									
20–30									
30–40									
40–50									
Day-Scholars									
0–10									
10–20									
20–30									
30–40									
40–50									
Total									
0–10									
10–20									
20–30									
30–40									
40–50									

(5) General Purpose Tables: They provide information for general use and reference. These are also called repository or reference tables.

Notes

(6) Special Purpose Tables: They are formulated to present some specific information relating to some specific subject under study. Such tables are also called text or summary tables.

Machine Tabulation

Tables are nor prepared with the help of machines which may be either hand-operated or are operated with electricity. Use of 'needle sorting' is one such machine for tabulation. Similarly 'Punch Cards' are also used. The work by these machines is more fast, easy and accurate.

Advantages of machine tabulation are:

(1) Greater accuracy.

(2) Less time required.

(3) Large-scale data can also be handled easily.

(4) Complex procedures like fitting trend lines etc becomes very easy with the help of computer.

(5) No work monotony can be avoided.

(6) Lowers cost by avoiding manual labour.

(7) The results are obtained without much waiting.

3.3 Frequency Distribution

The most important method of organising and summarising statistical data is by constructing a frequency distribution table. In this method, classification is done according to quantitative magnitude. The items are classified into groups or classes according to their increasing order in terms of magnitude and the number of items falling into each group is determined and indicated.

We shall discuss later questions such as how the classes are to be formed and how many classes are to be taken. We consider now how a frequency distribution table is to be constructed in the case of a discrete variable by taking a particular example.

Example 1 (a): Suppose that the marks secured by 60 students of a class are as follows:

46, 67, 23, 5, 12,	53, 38, 58, 26, 43,
36, 63, 26, 48, 76,	45, 66, 74, 16, 86,
56, 31, 58, 90, 32,	43, 36, 66, 46, 58,
36, 59, 54, 48, 21,	36, 64, 58, 45, 76,
58, 84, 68, 65, 59,	74, 48, 64, 58, 50,
46, 53, 64, 57, 65,	58, 95, 56, 66, 44.

Statistical Methods

Construct a frequency distribution table.

Marks obtained are divided into 10 groups or intervals as follows:

Marks below 10, between 11 and 20, between 21 and 30, and so on, between 91 and 100. Represent each mark by a tally (/), for example, corresponding to the mark 46 we put a tally (/) in the group 41 to 50; similarly we continue putting tallies for each mark. We continue upto four tallies and the fifth tally is put crosswise (\) so that it becomes clear at once that the lot contains five tallies, *i.e.* there are five marks. A gap is left after a lot of five tallies, before starting again to mark the tallies after each lot. The number of tallies in a class or group indicates the number of marks falling under that group. This number is known as the frequency of that group or corresponding to that class interval. Proceeding in this way, we get the following frequency table.

Notes

Table 1: Frequency distribution of marks secured by 60 students.

Class interval	Tally	Frequency (No. of students securing marks which fall in the class interval)
0 to 10	|	1
11 to 20	||	2
21 to 30	||||	4
31 to 40	𝍸 ||	7
41 to 50	𝍸 𝍸 ||	12
51 to 60	𝍸 𝍸 𝍸	15
61 to 70	𝍸 𝍸 |	11
71 to 80	||||	4
81 to 90	|||	3
91and above	|	1
Total		**60**

We shall now consider construction of a frequency distribution table of a continuous variable.

Example 1 (a): The heights of 50 students to the nearest centimetre are as given below:

151, 147, 145, 153, 156,	152, 159, 153, 157, 152,
144, 151, 157, 147, 150,	157, 153, 151, 149, 147,
151, 147, 155, 156, 151,	158, 149, 147, 153, 152,
149, 151, 153, 150, 152,	154, 150, 152, 149, 151,
151, 154, 155, 152, 154	152, 156, 155, 154, 150.

Construct a frequency distribution table.

We form the classes as follows: 145-146, 147-148, 149-150, 151-152, 153-154, 155-156, 157-158, 159-160 and construct the following frequency table:

Table 2: Frequency distribution of heights of 50 students

Class interval (Height in *cm*)	Tally	Frequency (Number of students having height)
145-146	||	2
147-148	𝍸	5
149-150	𝍸 |||	8
151-152	𝍸 𝍸 𝍸	15
153-154	𝍸 ||||	9
155-156	𝍸 |	6
157-158	||||	4
159-160	|	1
Total		**50**

Notes

We have given heights in *cms* in whole numbers or heights have been recorded to the nearest centimetre. Thus a height of 144.50 or more but less than 145.5 is recorded as 145; a height of 145.5 or more but less than 146.5 is recorded as 146 and so on. So the class 145-146 could also be indicated by 144.5-146.5 implying the class which includes any height greater than or equal to 144.5 but less than 146.5; the class 147-148 could be indicated by 146.5-148.5, meaning the class which includes any height greater than or equal to 146.5 but less than 148.5. Following this convention, the classes could be represented as: 144.5-146,5, 146.5-148.5, and so on. The above frequency distribution should finally be represented as follows.

Table 3: Frequency distribution of heights of 50 students

Heights (in *cm*)	Frequency (Number of students)
144.5-146.5	2
146.5-148.5	5
148.5-150.5	8
150.5-152.5	15
152.5-154.5	9
154.5-156.5	6
156.5-158.5	4
158.5-160.5	1
Total	**50**

Class intervals, Class limits and Class boundaries

The interval defining a class is known as a *class interval*. For Table: 2 145-146, 147-148, . . . are class intervals. The end numbers 145 and 146 of the class interval 145-146 are known as *class limits*; the smaller number 145 is the *lower class limit* and the larger number 146 is the *upper class limit*.

When we refer to the heights being recorded to the nearest centimetre and consider a height between 144.5.and 146.5 (greater or equal to 144.5 but less than 146.5) as falls in that class and the class is represented as 144.5-146.5, the end numbers are called *class boundaries*, the smaller number 144.5 is known as *lower class boundary* and the larger number 146.5 as *upper class boundary*. The difference between the upper and lower class boundaries is known as the *width* of the class. Here the width is 146.5–144.5 = 2 *cm* and is the same for all the classes. The common width is denoted *by c:* here $c = 2$ *cm*. Note that in certain cases, it may not be possible to have the same width for all the classes (specially the end classes).

Note also that the upper class boundary of a class *coincides* with the lower class boundary of the next class; there is no ambiguity: we have clearly indicated that an observation less than 146.5 will fall in the class 144.5-1465 and an observation equal to 146.5 will fall in the class 146.5-148.5.

3.4 Cumulative Frequency Distribution

Consider the number of all observations which are *less than the upper class boundary* of a given class interval; this number is the sum of the frequencies upto and including that class to which the upper class boundary corresponds. This sum is known as the *cumulative frequency* upto and including that class interval For example, consider Table 2; the cumulative frequency upto and including the class interval 145-146 is 2, that upto and including the next class interval 147-148 is 2 + 5 = 7, that upto and including the next class interval 149-150 is 2 + 5 + 8 = 15 and so on. This implies that two students have heights less than the upper class boundary of the class 145-146, seven students have heights less than the upper class boundary of the class 147-148 and so on. We can thus construct the cumulative frequency table as follows:

Notes

Table 4: Cumulative frequency (less than) table of heights of 50 students

Class (in *cm*) interval	Frequency	Cumulative Frequency (less than)
145-146	2	2
147-148	5	7
149-150	8	15
151-152	15	30
153-154	9	39
155-156	6	45
157-158	4	49
159-160	1	50
Total	**50**	

The cumulative frequency distribution is represented by joining the points obtained by plotting the cumulative frequencies along the vertical axis and the corresponding upper class boundaries along the *x-axis*. The corresponding polygon is known as cumulative frequency polygon (less than) or ogive. By joining the points by a freehand curve we get the cumulative frequency curve ("less than"). Similarly we can construct another cumulative frequency distribution ("more than" type) by considering the sum of frequencies greater than the lower class boundaries of the classes. For example, the total frequency greater than the lower class boundary 158.5 of the class 159-160 is one (1), while the total frequency greater than the lower class boundary 156.5 of the class 157-158 is 1 + 4 = 5, that of the class 155-156 is I + 4 + 6 = 11, and so on. Given below is Table 5 of cumulative frequency distribution ("more than") of the same distribution.

Table 5: Cumulative frequency (more than) table of heights 50 students

Class (in *cm*.) interval	Frequency	Cumulative frequency (more than)
145-146	2	50
147-148	5	48
149–150	8	43
151-152	15	35
153-154	9	20
155-156	6	11
157-158	4	5
159-160	1	1
Total	**50**	

The graph obtained by joining the points obtained by plotting the cumulative frequencies ("more than") along the vertical axis and the corresponding *lower* class boundaries along the *x*-axis is known as *cumulative frequency polygon* (greater than) or *ogive*. By joining the points by a free-hand curve, one gets the *cumulative frequency curve* ("more than" type).

Notes

These two curves are shown in Figure 1.

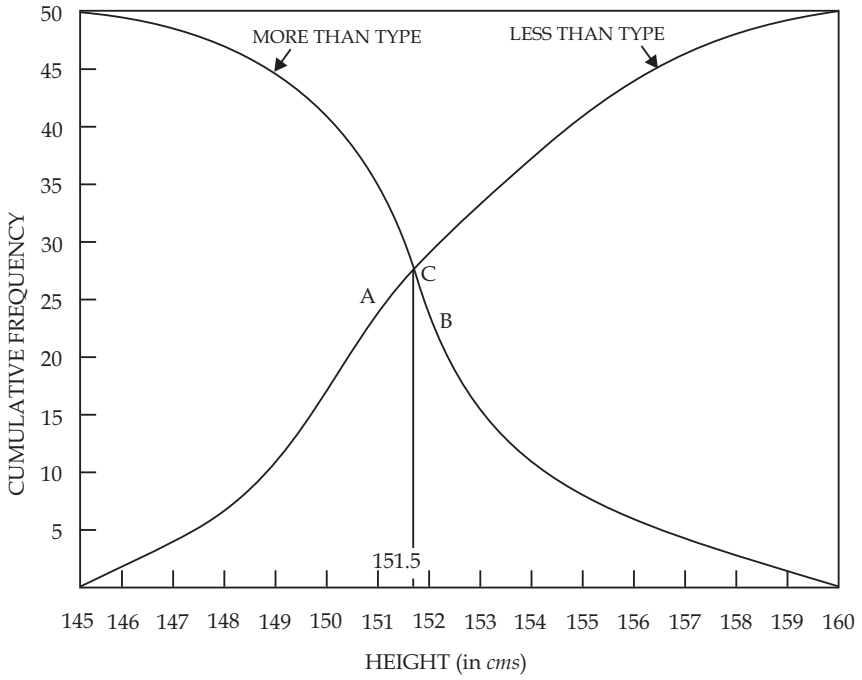


Figure 1: Cumulative frequency curves ('more than' and 'less than' types) of height 1 50 students

PRACTICAL QUESTIONS

1. Transform the following information in continuous series:

Weekly wages of 100 workers (in Rs.) of factory A								
880	230	270	280	960	940	930	860	990
820	240	240	550	990	950	860	820	360
960	390	260	540	1000	560	840	830	860
1020	480	270	260	1000	590	830	840	480
1040	460	300	270	1010	600	890	460	490
1060	330	360	300	1030	700	900	490	500
1040	360	370	400	1060	720	940	500	600
240	390	490	460	1070	760	960	460	670
260	780	500	440	460	790	990	360	680
290	670	560	990	480	800	1020	320	510

Solution: The lowest value is 1230 and highest is 1060. The difference in the highest and lowest value is 830. If we take a class interval of 100, nine classes would be formed. The first class should be 200-300 instead of 230-330.

Notes

Frequency Distribution

Wages	Tally bars	Frequency
200-300	𝍸 𝍸 |	13
300-400	𝍸 𝍸 |	11
400-500	𝍸 𝍸 𝍸 |	18
500-600	𝍸 𝍸	10
600-700	𝍸 |	6
700-800	𝍸	5
800-900	𝍸 𝍸 ||||	14
900-1000	𝍸 𝍸 ||	12
1000-1100	𝍸 𝍸 |	11
Total		**100**

2. If the class mid-points in a frequency distribution of age of a group of persons are 25, 32, 39, 46, 53 and 60, find: (i) the size of the class interval, (ii) the class boundaries and (iii) the class limits, assuming that the age quoted is the age completed on last birthday.

Solution: (i) The size of class interval

= Difference between the mid-values of any two consecutive classes

= 32 – 25 = 39 – 39.....60 – 53 = 7.

(ii) Since the magnitude of the class is 7 and the mid-values of the classes are 25, 32.....60, the corresponding class boundaries for different classes are obtained by adding and subtracting half the magnitude of the class interval *i.e.*, $\frac{7}{2}$ = 3·5 to/from the mid-values to obtain higher and lower class boundaries.

1st Class → 25 – 3·5, 25 + 3·5

2nd Class → 32 – 3·5, 32 + 3·5

3rd Class → 39 – 3·5, 39 + 3·5

4th Class → 46 – 3·5, 46 + 3·5

5th Class → 53 – 3·5, 53 + 3·5

6th Class → 60 – 3·5, 60 + 3·5.

Class Intervals

21·5 – 28·5

28·5 – 35·5

35·5 – 42·5

42·5 – 49·5

49·5 – 56·5

56·5 – 63·5

Notes

(iii) Assuming that the age quoted (X) is the age completed on last birthday then X will be a discreate variable which can take only integral values. Hence the given distribution can be expressed in an inclusive type classes with class interval of magnitude 7, as in the table given below.

Age (on last birthday)	Mid Values
22-28	25
29-35	32
36-42	39
43-49	46
50-56	53
56-63	60

3. Industrial finance in India showed great variations in respect of sources during the first, second and third plans. There were two main sources *viz.*, internal and external. The former had two sources — depreciation and free reserves surplus. The latter had three sources — capital issues, borrowings and 'other sources'. During the first plan, internal and external sources accounted for 62% and 38% of the total and in this depreciation fresh capital and 'other sources formed 29%, 7% and 10·6% respectively.

   During the second plan internal sources decreased by 17·3% compared to the first plan and depreciation was 24·5%. The external finance during the same period consisted of fresh capital 10·9% and borrowings 28·9%. Compared to the second plan, during the third plan, external finance decreased by 4·4% and borrowings and 'other sources' were 29·4% and 14·9% respectively. During the third plan, internal finance increased by 4·4% and free reserves and surplus formed 18·6%.

   Tabulate the above information with the above details as clearly as possible, observing the rules of tabulation.

Solution:

Table Showing Pattern of Industrial Finance (in per cent)

Plan	Sources						
	Internal			External			
	Depreciation	Free reserves and surplus	Total	Capital issues	Borrowings	Other	Total Sources
First	29·0	33·0	62·0	7·0	20·4	10·6	38·0
Second	243·5	20·2	44·7	10·9	28·9	15·5	55·3
Third	30·5	18·6	49·1	6·6	29·4	14·9	50·9

Self-Assessment

1. Fill in the blanks:

(*i*) Classification is the ..... step in tabulation.

(*ii*) When data are observed ..... the type of classification is known as chronological classification.

(*iii*) ..... classification refers to the classification of data according to some characteristics that can be measured.

Notes

(*iv*) A table is a systematic arrangement of statistical data in ..... .

(*v*) In collection and tabulation ..... is the chief requisite and experience the chief ..... .

(*vi*) The number of observations corresponding to a particular class is known as the ..... of that class.

3.5 Summary

- Moreover, classification and tabulation segregates the likes from the unlikes. The heterogeneity is removed. The data are classified into classes and sub-classes according to their characteristics. This process is called **classification**. The classified data are presented in prescise and systematic tables. This process is called **tabulation**. By these two processes, the data collected are made simple, easy to understand and carry out analysis and interpretations.
- Classification may be defined as the process of arranging the available data in various groups or classes in accordance with their resemblances and similarities and keeping in view some common features and objectives of study. Thus, through classification, an effort is made to achieve homogeneity of the collected information. While classifying, the units with common characteristics are placed together and in this way the whole data is divided into a number of classes and sub-classes.
- The process of classification helps in arranging the data in such a way that the large mass of irrelevant looking data becomes simple and easy to understand, avoiding unnecessary details, making logical sense.
- Classification is done in accordance with the attributes or characteristics of the data. Such classification is generally done where data cannot be measured. Under this classification method, the presence or absence of an attribute is the basis of classification.
- The bases of classification are manifold, *i.e.*, more than one attribute. In this method, classes/ groups are further classified into sub-classes and sub-groups. For example, population/sample is first classified on the basis of sex, then for each sex (male or female) marital status forms two sub-classes then further these sub-classes are classified as per their literacy state.
- The process of tabulation eliminates unnecessary details and present the complex data concisely in rows and columns. This helps in simplifying the complex data which becomes more meaningful and better understood.
- It is only possible after tabulation, that the data can be subjected to statistical analysis and interpretation. Measures of correlation, regression dispersion etc can be easily calculated when the data is in tabular form.
- A title is a brief statement indicating about the nature of the data and the time-span to which the data relate. The geographical distribution of the data, if any, should also be indicated in the title. The title should be in dark letters in comparison to other heads and sub-heads in the table.
- Footnotes are also used to describe the source of the data, if the data is secondary in nature. For example, "the figures in bracket show the per cent rise over previous year" or "the profit above is after tax" etc are some of the information which is given in footnotes.
- A table presents the statistical data in a systematic way in rows and columns which concisely explain the numerical facts. Tabulation is nothing but the process of preparing a table. The preparation of table is a specialised activity and is done through a set of rules.
- Neat and tidy appearance must be given to the table which can be done by providing proper ruling and spacing as is necessary. If the table continues to the next page, no bottom line must be drawn, as it would indicate the end of the table. Major and minor items must be given space according to their relative importance. Coloured inks, heavy printed titles or sub-titles, thick and thin ruling etc must be used to clarify a complex table.

Notes

- The body of the table must be as comprehensive as possible, consistent with the purpose. Unnecessary details must be avoided and, items in 'miscellaneous or unclassified columns must be least.
- Tabulation is an art which requires common sense in planning a table and viewing the proposed table from the point of view of the user or the other person. Tabulation is done keeping in mind the purpose of statistical investigation. The rules of tabulation act as guides in preparing a good table.
- Quantitative classification data is done through seriation of data. If two variable quantities are arranged side by side so that the measurable differences in the one correspond to the measurable differences in the other, the result is formation of a statistical series.
- In continuous frequency distribution data are divided into class intervals instead of individual values (as is done in case of discrete frequency distribution) class intervals can be formulated like marks from 0 to 10, 10 – 20 and so on.
- Tables are nor prepared with the help of machines which may be either hand-operated or are operated with electricity. Use of 'needle sorting' is one such machine for tabulation. Similarly 'Punch Cards' are also used. The work by these machines is more fast, easy and accurate.
- The most important method of organising and summarising statistical data is by constructing a frequency distribution table. In this method, classification is done according to quantitative magnitude. The items are classified into groups or classes according to their increasing order in terms of magnitude and the number of items falling into each group is determined and indicated.
- Consider the number of all observations which are *less than the upper class boundary* of a given class interval; this number is the sum of the frequencies upto and including that class to which the upper class boundary corresponds.
- The cumulative frequency distribution is represented by joining the points obtained by plotting the cumulative frequencies along the vertical axis and the corresponding upper class boundaries along the *x-axis*. The corresponding polygon is known as cumulative frequency polygon (less than) or ogive. By joining the points by a freehand curve we get the cumulative frequency curve ("less than"). Similarly we can construct another cumulative frequency distribution ("more than" type) by considering the sum of frequencies greater than the lower class boundaries of the classes.

3.6 Key-Words

1. Tabulation of data : The process of placing classified data into tabular form is known as tabulation. A table is a symmetric arrangement of statistical data in rows and columns. Rows are horizontal arrangements whereas columns are vertical arrangements. It may be simple, double or complex depending upon the type of classification.
2. Frequency distribution : In statistics, a frequency distribution is an arrangement of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

3.7 Review Questions

1. What do you mean by Tabulation ? What are the objectives and advantages of tabulation?
2. What is the frequency distribution ? Explain how it is formed from raw data.
3. Describe the importance of classification and tabulation in statistical analysis.
4. Describe the various points to be considered in the construction of a frequency table.
5. What are the different parts of a table ? What points should be taken into account while preparing a table ?

Notes

Answers: Self-Assessment

1. (*i*) first (*ii*) over a period of the time
   (*iii*) Quantitative (*iv*) columns and rows
   (*v*) common sense, teacher (*vi*) frequency

3.8 Further Readings


Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 4: Central Tendency: Mean, Median and Mode and their Properties

CONTENTS

Objectives
Introduction
4.1 Meaning and Definition of Central Tendency
4.2 Mean, Median and Mode and their Properties
4.3 Summary
4.4 Key-Words
4.5 Review Questions
4.6 Further Readings

Objectives

After reading this unit students will be able to:

- Know the Meaning and Definition of Central Tendency.
- Discuss the Mean, Median and Mode and their Properties.

Introduction

For statistical analysis, condensation of data is essential so that the complexity of data is reduced and is made comparable. This can be done by finding the central tendencies of the data or the averages. By this, the large mass of data gets reduced to one figure each and thus comparison becomes much easier. For example, if a comparison of student's results in two different colleges with 200 students each, is to be made, it seems to be impossible to draw any conclusion looking at the results of these 400 students. But if, each of these series is represented by a single figure, comparison becomes very easy. This figure is the one which represents the whole series, and so it neither is the highest nor the lowest value rather, it is the value where most of the items of the series cluster or are nearer. Such figures present the central tendency of the series and are called Measures of central tendency or Averages. Its value lies between maximum and minimum.

4.1 Meaning and Definition of Central Tendency

Measures of central tendency or averages reduce the large number of observations to one figure. Actually the measures of central tendency describe the tendency of items of group around the middle in a frequency distributions of numerical values.

Definitions

According to *L. J. Kaplan* — "One of the most widely used set of summary figures is known as measures of location, which are often referred to as averages, central tendency or central location. The purpose for computing an average value for a set of observations is to obtain a single value which is representative of all the items and which the mind can grasp simply and quickly. The single value is the point of location around which the individual items cluster."

According to *G.P. Watkins,* "average is a representative figure which is gist, if not the substance of statistics."

In the words of *Croxton* and *Cowden,* "An average value is single value within the range of the data that is used to represent all the values in the series."

Notes

Characteristics of a Good Average

The above discussion reveals that an average or the value of central tendency is a representative figure. Therefore, a good average would be the one which has the capability of representing the data most efficiently and effectively. For this, certain are the characteristics of the average so that it can prove to be good. These essential characteristics for an average to prove to be good are:

(1) **It should be rigidly defined:** According to Prof. Yule and Kendall, the average should be defined rigidly so that there is only one possible interpretation and is not subject to observers' own interpretation and bias. For this, the average should be defined in terms of alzebraic formula.

(2) **It should be based on all the observations:** In order to make the data representative it is very essential that it is based on all the observations.

(3) **It should be capable of further alzebraic treatment:** For the average to be good, it is essential that it is capable of further alzebraic treatment, otherwise its use will become very limited. For example, in the absence of this quality, the combined average of two or more series from their individual averages will not be calculated. This would hinder the possibility to study the average relationship of various parts of a variable, if it is expressed as the sum of two or more variables.

(4) **It should not be affected by fluctuations of sampling:** If two independent sample studies are made in any particular field, their averages obtained, should not differ from each other ideally. Practically, it is difficult to obtain no difference, but the average in which this difference, technically called as 'fluctuation of sampling' is least, is considered to be a better average.

(5) **It should be easy to compute:** An average should be capable of being calculated with reasonable ease and within reasonable time. If the time taken is long or the calculations are tedious and complicated, the average shall have only limited use.

(6) **It should be easy to understand:** A good average is the one which is easily understood by the common people. It should neither be abstract nor too mathematical; otherwise its use will again be restricted.

Types of Statistical Averages

The following are the main types of statistical averages:

(1) **Positional Averages:** These include —

(a) Median (represented by M)

(b) Mode (represented by Mo).

(2) **Mathematical Averages:** These include —

(a) Arithmatic Average or Mean (represented by $\bar{X}$)

(b) Geometric Mean (represented by 'G.M.')

(c) Harmonic Mean (represented by 'H.M.')

(d) Quadratic Mean (represented by 'Q.M.')

(3) **Commercial Averages:**

(a) Moving Average.

(b) Progressive Average.

(c) Composite Average.

Measures or types of Central tendency or averages can be shown as in Figure 1.

Notes

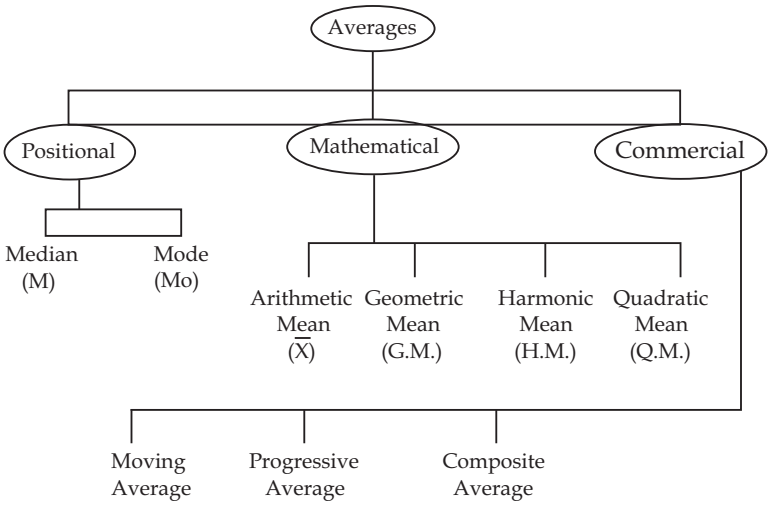


Figure: 1

Symbolically, the above may be shown as:

Mode = Z or M_O

Median = M

A. A. or Mean or $\bar{X}$

Geometric Mean = *g* or G.M.

Harmonic Mean = *h* or H.M.

Quadratic Mean = Q.M.

4.2 Mean, Median and Mode and their Properties

Arithmetic Average or Mean

Arithmetic mean is the most widely used method of calculated averages, so much so that when only 'mean' is indicated it is assumed to be arithmetic mean universally. It is obtained by adding up all the observations and dividing it by number of observations.

Merits of Arithmetic Mean

The merits of Arithmetic Mean are:

(1) Simple to understand,

(2) Easy to compute,

(3) Capable of further mathematical treatment,

(4) Calculated on the basis of all the items of the series,

(5) It gives the value which balances the either side,

(6) Can be calculated even if some values of the series are missing,

(7) It is least affected by fluctuations in sampling.

Demerits of Arithmetic Mean

The demerits of Arithmetic Mean are:

(1) Extreme items have disproportionate effect. For example, average marks obtained of five students are:

Notes

$$\frac{50+10+10+10+10}{5} = \frac{90}{5} = 18.$$

Whereas in reality 4 out of 5 students failed. Therefore, '18' marks cannot be termed as representative.

(2) When data is vast, the calculations become tedious.

(3) In case of open end classes, mean can only be calculated by making some assumptions.

(4) A.M. is not representative if series is asymmetrical.

Purpose or Objectives of Averaging

Central tendency or average is the value by which the data can be represented. The purpose or objectives of calculating this representative figure are —

(1) To present the most important features of a mass of complex data.

(2) To facilitate comparing one set of data with others, so that conclusions can be drawn quickly.

(3) To help in understanding the picture of a complete group by means of sample data.

(4) To trace the mathematical relationship between different groups or classes.

(5) To help in the decision-making.

(6) To facilitate the process of experimentation and research.

Weighted Arithmetic Mean

Weighted arithmetic mean is the method of calculating a more representative central value and takes into consideration the relative importance of the various figures in the series. Whereas in simple arithmetic mean, equal weight or importance is given to each item. If the central value has to more representative and the data is such that few items are more important than other, the method of weighted arithmetic mean is used. This method is generally used in the following situations:

(1) When importance of all the items in the series is not equal.

(2) When the classes of the same group contain widely varying frequencies.

(3) Where there is a change either in the proportion of values or items or in the proportion of frequencies.

(4) When ratios, percentages or rates are being averaged.

(5) When it is desired to calculate the average of series from the average of its component parts.

The formula for the weighted arithmetic average is:

Direct Method : $\overline{X}W = \frac{W_1X_1 + W_2X_2 + ...W_nX_n}{W_1 + X_2 + ...W_n} = \frac{\Sigma WX}{\Sigma W}$

Short-cut Method : $\overline{X}W = A_w + \frac{\Sigma Wd}{\Sigma W}$

where $\overline{X}W$ represents the weighted arithmetic mean.

X represents the variable values *i.e.*, X_1, X_2 ... X_n.

W represents the weights attached to variable values *i.e.*, W_1, W_2, ... W_n, respectively.

ΣWd Sum of the product of the deviations from the assumed mean (AW) multiplied by the respective weights.

AW Assumed (weighted) mean.

Harmonic Mean

An average rate like kilometer per hour, per day items manufactured etc. are required to be found, harmonic mean is calculated. Harmonic mean is a type of average which has limited application only

Notes

that too in a restricted field. The harmonic mean of a series of values is the reciprocal of the arithmetic mean of the reciprocals of the individual values. Reciprocal tables are used with ease for this. The Harmonic Mean is less than the geometric mean of the same observations. The formula to calculate harmonic mean is:

$$\text{H.M.} = \frac{N}{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}+\ldots\frac{1}{X_n}}$$

or, reciprocal of —

$$\frac{\frac{1}{X_1}+\frac{1}{X_2}+\frac{1}{X_3}\ldots\frac{1}{X_n}}{N}$$

or, reciprocal of $\frac{(\Sigma \text{Reciprocal of } X)}{N}$

where X, X_1, X_2, ... X_n are the observations or the values of the series.

Merits of H.M.

(1) Harmonic mean is calculated by taking into account all the items of the series.

(2) In series with wide dispersion, this method is useful.

(3) It gives less weight to large items and more weight to small ones (because reciprocals are used).

(4) The method is useful in calculating rate.

(5) While calculating harmonic mean, the values get weight automatically and there is no need to assign weights specifically.

Demerits of H.M.

(1) It is difficult to calculate.

(2) Difficult to be understood by common man.

(3) Harmonic mean cannot be calculated if even one item of the series is missing.

(4) The value of harmonic mean obtained may be a value which is no item in the given series.

Geometric Mean

Geometric mean of '*n*' numbers is defined as the n^{th} root of the product of '*n*' numbers. Symbolically,

$$\text{G.M.} = \sqrt[n]{(X_1)(X_2).(X_3)\ldots(X)_n}$$

where X_1 X_2 ... X_n are the various values of the series.

'*n*' is the number of items. To make the calculations of finding out n^{th} root simpler, logarithms are used. G.M. by using logs is found thus:

$$\text{G.M.} = \text{Antilog of } \frac{\Sigma \log X}{N}.$$

Where, measuring the ratios of change is required, the geometric mean are used. It is also most suitable when large weights have to be given to small items and small weights to large items, which is usually required to study economic and social phenomena.

Merits

(1) Based on all the values of the series.

(2) Capable of further alzebraic treatment.

Notes

(3) It is not much affected by the extreme values in the series.

(4) Capable of being applicable to study social and economic phenomena.

Demerits

(1) If any value is '0' (zero), G.M. cannot be calculated (because (X_1), (X_2) ... (X_n) is to be found. If any of these is zero, the multiplication result will be zero and interpretation would become impossible.

(2) Knowledge of logarithms is essential. Therefore it is found difficult to compute by a non-mathematics background person.

(3) Difficult to locate.

(4) G.M. cannot be calculated if even one value of the series is not available.

(5) The value of G.M. obtained may not be there in the series, therefore, it cannot be termed as the true representative of the data.

Mathematical Properties of the Arithmetic Mean

The following are a few important mathematical properties of the arithmetic mean:

1. The sum of the deviations of the items from the arithmetic mean (taking signs into account) is always zero, *i.e.*, $\Sigma(X - \bar{X}) = 0$*. This would be clear from the following example:

X	$(X - \bar{X})$
10	– 20
20	– 10
30	0
40	+ 10
50	+ 20
$\Sigma X = 150$	$\Sigma(X - \bar{X}) = 0$

Here $\bar{X} = \frac{\Sigma X}{N} = \frac{150}{5} = 30$. When the sum of the deviations from the actual mean, *i.e.*, 30, is taken it comes out to be zero. It is because of this property that the mean is characterised as a *point of balance, i.e.*, the sum of the positive deviations from it is equal to the sum of the negative deviations from it.

2. The sum of the squared deviations of the items from arithmetic mean is minimum, that is less than the sum of the squared deviations of the items from any other value. For example, if the items are 2, 3, 4, 5 and 6 their squared deviation shall be:

X	$(X - \bar{X})$	$(X - \bar{X})^2$
2	– 2	4
3	– 1	1
4	0	0
5	+ 1	1
6	+ 2	4
$\Sigma X = 20$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 = 10$

Notes

The sum of the squared deviations is equal to 10 in the above case. If the deviations are taken from any other value the sum of the squared deviations would be greater than 10. This is known as least squares property of the arithmetic mean.

3. The standard error of arithmetic mean is less than that of any other measure of central tendency.

4. Since $\bar{X} = \frac{\Sigma X}{N}$

$N\bar{X} = \Sigma X$.

In other words, if we replace each item in the series by the mean, then the sum of these substitutions will be equal to the sum of the individual items. For example, in the discussion of first property $\Sigma X = 150$ and the arithmetic mean = 30. If for each item we substitute 30, we get the same total *i.e.*, 150 = [30 + 30 + 30 + 30 + 30].

* Algebraically the property $\Sigma(X - \bar{X}) = 0$ is derived from the fact that $N\bar{X} = \Sigma X$

This property is of great practical value. For example, if we know the average wage in a factory, say, Rs. 200, and the number of workers employed, say, 50, we can compute total wage bill from the relation $N\bar{X} = \Sigma X$. The total wage bill in this case would be 200 × 50 = 10,000 which is equal to ΣX.

5. If we have the arithmetic mean and number of items of two or more than two related groups, we can compute combined average of these groups by applying the following formula:

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

$\bar{X}_{12}$ = combined mean of the two groups

$\bar{X}_1$ = arithmetic mean of first group

$\bar{X}_2$ = arithmetic mean of second group

N_1 = number of items in the first group

N_2 = number of items in the second group.

6. If the given observations on X be changed to observation on Y, where Y = $a + bX$, then $\bar{Y} = a + b\bar{X}$.

The following example shall illustrate the application of the above formula:

Example 1 : A factory employs 100 workers of whom 60 work in the first shift and 40 work in the second shift. The average wage of all the 100 workers is Rs. 38. If the average wage of 60 workers of the first shift is Rs. 40, find the average wage of the remaining 40 workers of the second shift.

Solution : Total no. of employees = 100

No. of employees in the first shift, *i.e.*, $N_1 = 60$

No. of employees in the second shift, *i.e.*, $N_2 = 40$

$\bar{X}_{12} = 38, \bar{X}_1 = 40$

$$\bar{X}_{12} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2}$$

Notes

$$38 = \frac{60(40) + 40\overline{X}_2}{100}$$

$$3800 = 2400 + 40\overline{X}_2$$

$$40\overline{X}_2 = 1400$$

$$\therefore \quad \overline{X}_2 = \frac{1400}{40} = 35$$

Hence the wage of the remaining 40 workers in the second shift is Rs. 35.

If we have to find out the combined mean of the three series, the above formula can be extended as follows:

$$\overline{X}_{123} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2 + N_3\overline{X}_3}{N_1 + N_2 + N_3}$$

Median

Median is one of the measures of central tendency. It is a positional average.

Definition of Median

Median may be defined as 'the middlemost or central value of the series when the values are arranged in ascending or descending order of magnitude'.

If there is an odd number of an item, then the median is found out by taking the middle most items of the series only after arranging the data in order of magnitude.

For example, if daily wages of 7 workers are Rs. 127, Rs. 167, Rs. 154, Rs. 177, Rs. 135, Rs. 160 and Rs. 157 and we wish to know the median wage, the wages must be arranged either in ascending order or descending order of magnitude and the 4^{th} reading will be the median wage.

Arranged in order of magnitude, the wages might be

Rs. 127 135 154 157 160 168 177

and the median wage is Rs. 157, *i.e.* the 4^{th} item.

If there is an even number of items, then the median is half-way between the two middle ones and it is found by taking the average of these two items. For example, if the marks secured by 10 students in an examination are 75, 48, 63, 89, 100, 55, 35, 28, 93 and 79 and we wish to know the median mark, the marks must be arranged either in ascending or descending order of magnitude and then the average of the value of the 5^{th} item and the 6^{th} item will be the median mark.

Arranged in order of magnitude, the marks might be 28, 35, 48, 55, 63, 75, 79, 89, 93, 100 and the median marks is Rs. $\frac{63+75}{2} = \frac{138}{2} = 69$, *i.e.* the average of the 5^{th} item and the 6^{th} item.

Properties of Median

Median is of the following properties:

1. Median may be the middlemost value of the series when the values are arranged in order of magnitude.
2. Median is influenced by the position of items in the array but not by the size of the items.
3. The value of the median of a series may or may not coincide with the value of an existing item.
4. The median cannot readily be located unless the data have been put into an array or into a frequency distribution.

Notes

5. The median of the sum or difference or pairs of corresponding items into two series is not equal to the sum or difference of the medians of the two series.

Mode

The term 'mode' has come from French in which it means 'to be in fashion'. As a statistical language, mode is the value that occurs most frequently in a statistical distribution. Thus 'Mode' is the most representative average and is a position of greatest concentration of values. It has great value conceptually. It is what the doctor means when he describes that a desease of cold and fever usually takes a week to get cured. Similarly, average size of shirt/shoes sold, average family income etc. also cannot be most frequently occurring value.

According to *Tate,* "The mode may be defined as the item which occurs most frequently in a statistical series."

In the words of *Garrett,* "Mode is that single measure or score which occurs most frequently."

Merits

The merits are as follows:

(1) Easy to understand,

(2) Simple to calculate and locate,

(3) Quantitative data in ranking is possible, mode is very useful,

(4) It is the actual value that is in the series,

(5) Mode remains unaffected by dispersion of series,

(6) Not affected by extreme items,

(7) Can be calculated even if extreme values are not known.

Demerits

The demerits are as follows:

(1) Mode cannot be subject to further Mathematical treatment, because it is not obtained from any alzebraic calculations,

(2) It is quite likely that there is no mode for a series,

(3) Cannot be used if relative importance of items have to be considered,

(4) Choice of grouping has a considerable influence on the value of the mode.



Did u know? Harmonic mean is a type of average which has limited application only that too in a restricted field.

Properties of mode

Assuming definedness, and for simplicity uniqueness, the following are some of the most interesting properties.

- All three measures have the following property: If the random variable (or each value from the sample) is subjected to the linear or affine transformation which repalces X by $aX + b$, so are the mean, median and mode.
- However, if there is an arbitrary monotonic transformation, only the median follows; for example, if X is replaced by exp (X), the median changes from m to exp (m) but the mean and mode won't.
- Except for extremely samples, the mode is insentive to "outliers" (such as occasional, rare, false experimental readings). The median is also very robust in the presence of outliers, while the mean is rather sensitive.

Notes

- In continuous unimodal distributions the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, median $\approx$ (2 × mean + mode)/3. This rule, due to Karl Pearson, often applies to slightly non-symmetric distributions that resemble a normal distribution, but it is not always true and in general the three statistics can appear in any order.
- For unimodal distributions, the mode is within $\sqrt{3}$ standard deviations of the mean, and the root mean square deviation about the mode is between the standard deviation and twice the standard deviation.

Self-Assessment

1. Fill in the blanks:

(*i*) The consecutive addition of frequencies is called .....

(*ii*) Below 10, more than 40 are the examples of ..... class-intervals.

(*iii*) Sum of the deviations of the items from the ..... is always zero (taking + *ve* and –*ve* signs).

(*iv*) n^{th} root or '*n*' items of a series is termed as .....

(*v*) ..... of a series is the reciprocal of the arithmetic avrage of the reciprocals of the values of its various items.

4.3 Summary

- Measures of central tendency or averages reduce the large number of observations to one figure. Actually the measures of central tendency describe the tendency of items of group around the middle in a frequency distributions of numerical values.
- For the average to be good, it is essential that it is capable of further alzebraic treatment, otherwise its use will become very limited. For example, in the absence of this quality, the combined average of two or more series from their individual averages will not be calculated. This would hinder the possibility to study the average relationship of various parts of a variable, if it is expressed as the sum of two or more variables.
- An average should be capable of being calculated with reasonable ease and within reasonable time. If the time taken is long or the calculations are tedious and complicated, the average shall have only limited use.
- Arithmetic mean is the most widely used method of calculated averages, so much so that when only 'mean' is indicated it is assumed to be arithmetic mean universally. It is obtained by adding up all the observations and dividing it by number of observations.
- Weighted arithmetic mean is the method of calculating a more representative central value and takes into consideration the relative importance of the various figures in the series. Whereas in simple arithmetic mean, equal weight or importance is given to each item. If the central value has to more representative and the data is such that few items are more important than other, the method of weighted arithmetic mean is used.
- An average rate like kilometer per hour, per day items manufactured etc. are required to be found, harmonic mean is calculated. The harmonic mean of a series of values is the reciprocal of the arithmetic mean of the reciprocals of the individual values. Reciprocal tables are used with ease for this. The Harmonic Mean is less than the geometric mean of the same observations.
- The sum of the squared deviations of the items from arithmetic mean is minimum, that is less than the sum of the squared deviations of the items from any other value.
- Median may be defined as 'the middlemost or central value of the series when the values are arranged in ascending or descending order of magnitude'.
- If there is an odd number of an item, then the median is found out by taking the middle most items of the series only after arranging the data in order of magnitude.

Notes

- If there is an even number of items, then the median is half-way between the two middle ones and it is found by taking the average of these two items. For example, if the marks secured by 10 students in an examination are 75, 48, 63, 89, 100, 55, 35, 28, 93 and 79 and we wish to know the median mark, the marks must be arranged either in ascending or descending order of magnitude and then the average of the value of the 5^{th} item and the 6^{th} item will be the median mark.
- The term 'mode' has come from French in which it means 'to be in fashion'. As a statistical language, mode is the value that occurs most frequently in a statistical distribution. Thus 'Mode' is the most representative average and is a position of greatest concentration of values. It has great value conceptually. It is what the doctor means when he describes that a desease of cold and fever usually takes a week to get cured. Similarly, average size of shirt/shoes sold, average family income etc. also cannot be most frequently occurring value.
- In continuous unimodal distributions the median lies, as a rule of thumb, between the mean and the mode, about one third of the way going from mean to mode. In a formula, median $\approx$ (2 × mean + mode)/3. This rule, due to Karl Pearson, often applies to slightly non-symmetric distributions that resemble a normal distribution, but it is not always true and in general the three statistics can appear in any order.

4.4 Key-Words

1. Central tendency : In statistics, the term central tendency relates to the way in which quantitative data tend to cluster around some value.[1] A measure of central tendency is any of a number of ways of specifying this "central value". In practical statistical analysis, the terms are often used before one has chosen even a preliminary form of analysis: thus an initial objective might be to "choose an appropriate measure of central tendency".
2. Harmonic mean : The reciprocal of the arithmetic mean of the reciprocals of a specified set of numbers

4.5 Review Questions

1. What is meant by measures of central tendency? What are the characteristics of good measure of central tendency?
2. Explain the relative importance of arithmetic mean, median and mode as measures of central tendency in statistical analysis.
3. Define mean, median and mode. Mention its merits and demerits.
4. What are the properties of mean, median and mode?
5. Define Harmonic mean and give a situation in which it is used.

Answers: Self-Assessment

1. (*i*) cumulative frequency (*ii*) open-end
   (*iii*) mean (*iv*) geometric mean
   (*v*) harmonic mean

4.6 Further Readings


Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.

Notes

Unit 5: Application of Mean, Median and Mode

CONTENTS

Objectives
Introduction
5.1 Application of Mean
5.2 Application of Median
5.3 Application of Mode
5.4 Summary
5.5 Key-Words
5.6 Review Questions
5.7 Further Readings

Objectives

After reading this unit students will be able to:

- Discuss Application of Mean and Median.
- Know the Application of Mode.

Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

5.1 Application of Mean

The most popular and widely used measure for representing the entire data by one value is what most laymen call an 'average' and what statisticians call the arithmetic mean. Its value is obtained by adding together all the items and by the dividing this total by the number of items. Arithmetic mean may be either (i) simple arithmetic mean, or (ii) weighted arithmetic mean.

Calculation of Arithmetic Mean—Individual Observations

The process of computing mean in case of individual observations (*i.e.* where frequencies are not given) is very simple. Add together the various values of the variable and divide the total by the number of items. Symbolically:

$$\overline{X}^{*} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} = \frac{\Sigma X}{N}$$

$\overline{X}$ = Arithmetic Mean

Notes

$\sum X$ = Sum of all the values of the variable X, *i.e.,* $X_1, X_2, X_3, \ldots\ldots, X_N$.

N = Number of observations.

Steps: The formula suggests two steps in calculating mean:

(i) Add together all the values of the variable X and obtain the total $\sum X$.

(ii) Divide this total by the number of observations.

Example 1: Calculate arithmetic mean from the following data:

Marks obtained by 20 students out of 200			
40	100	144	100
56	106	148	106
68	108	150	108
78	118	156	118
84	128	158	128

Solution:

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

$\sum X$ = Summation of all the items, N = 20.

$\sum X$ = 2202, N = 20

$$\bar{X} = \frac{2202}{20} = 110.1.$$

Answer: Arithmetic mean of the series is = 110.1.

Short-cut Method: The above method is useful only when 'N' is small. Mean of marks cannot be calculated with ease by the above method. Therefore, short-cut method is used. This method is based on the fact that the alzebraic sum of the deviations of a series of individual observations from their mean is always equal to zero.

Arithmetic mean by short-cut is calculated by the following formula:

$$\bar{X} = A + \frac{\sum dx}{N}$$

where, $\bar{X}$ is arithmetic mean, A is assumed mean, N is number of observations, $\sum dx$ is the sum of the deviations from the assumed mean.

Example 2: Using short-cut method, determine the arithmetic mean of the data (given in example 1 taking first 15 students).

Solution:

X	Deviations dx = X – A
40	40 – 100 = – 60
56	56 – 100 = – 44
68	68 – 100 = – 32
78	78 – 100 = – 22
84	84 – 100 = – 16
100	100 – 100 = 0
106	106 – 100 = 6
108	108 – 100 = 8

Notes

118	118 – 100 = 18
128	128 – 100 = 28
144	144 – 100 = 44
148	148 – 100 = 48
150	150 – 100 = 50
156	156 – 100 = 56
158	158 – 100 = 58
N = 15	$\sum dx$ = 142

So let assumed mean A = 100.

A.M. by short-cut method:

$$\bar{X} = A + \frac{\sum dx}{N}$$

$$A = 100 \quad \sum dx = 142, N = 15$$

$$\therefore \quad \bar{X} = 100 + \frac{142}{15}$$

$$= 100 + 9.47$$

$$= 109.47$$

Arithmetic Mean in Discrete Series

A discrete series is obtained from a large number of individual observations. Suppose the marks obtained by 100 students is given. This data can be converted into a discrete series where the marks obtained are accompained by the number of students obtaining it. For example, suppose 10 students obtained 50 marks, 12 students obtained 60, 25 students obtained 78, 3 students obtained 100, 15 students obtained 94, 15 students obtained 82 and 20 students obtained 38. Then instead of writing in form of individual observations, data can be written like this:

Marks Obtained (X)	Number of Students (*f*)
50	10
60	12
78	25
100	3
94	15
82	15
38	20
	Total N = 100

Then this is a discrete series.

Arithmetic Mean by Direct Method

$$\bar{X} = \frac{\sum fx_1 + \sum fx_2 \ldots \sum fx_n}{\sum f} \text{ or } \frac{\sum fx}{\sum f}$$

'*f*' denotes the frequency.

Arithmetic Mean by Short-cut Method

Notes

(i) $$\bar{X} = A + \frac{\Sigma fdx}{\Sigma f}$$

where A is assumed mean.

(ii) $$\bar{X} = A + \frac{\Sigma fdx}{\Sigma f} \times i$$

where i is the common factor of deviations.

Σfdx is the total of the products of the deviations from the assumed average and the respective frequency of the items.

Σf is the summation of all the frequencies.

Example 3: From the following frequency distribution calculate the mean weight of the students.

Weight (in *kgs.*)	64	65	66	67	68	69	70	71	72	73
No. of Students	1	6	10	22	21	17	14	5	3	1

Solution:

Weight (X)	**Number of Students (*f*)**	**Deviation d_2 = (X – A)**	**(*fxdx*) *fdx*.**
64	1	64 – 68 = – 4	– 4 × 1 = – 4
65	6	65 – 68 = – 3	– 3 × 6 = – 18
66	10	66 – 68 = – 2	– 2 × 10 = – 20
67	22	67 – 68 = – 1	– 1 × 22 = – 22
68	21	68 – 68 = 0	0 × 21 = 0
69	17	69 – 68 = 1	1 × 17 = 17
70	14	70 – 68 = 2	2 × 14 = 28
71	5	71 – 68 = 3	3 × 5 = 15
72	3	72 – 68 = 4	4 × 3 = 12
73	1	73 – 68 = 5	5 × 1 = 5
	Σf = 100		Σfdx = 13

Let assumed mean A = 68.

$$\bar{X} = A + \frac{\Sigma fdx}{\Sigma f}$$

A = 68, Σfdx = 13, Σf = 100.

$$\therefore \quad \bar{X} = 68 + \frac{13}{100}$$

$$\bar{X} = 68 + 0.13$$

$$\therefore \quad \bar{X} = 68.13$$

Answer: Mean weight of the students = 68.13 *kg*.

Notes

Example 4: Find the arithmetic mean of the following data using step deviation method:

X	1590	1610	1630	1650	1670	1690	1710	1730
f	1	2	9	48	131	102	40	17

Solution:

X	f	dx **= (X – A)**	**Step deviation** dx'	$-fdx'$
1590	1	1590 – 1670 = – 80	– 8	1 × – 8 = – 8
1610	2	1610 – 1670 = – 60	– 6	2 × – 6 = – 12
1630	9	1630 – 1670 = – 40	– 4	9 × – 4 = – 36
1650	48	1650 – 1670 = – 20	– 2	48 × – 2 = – 96
1670	131	1670 – 1670 = 00	0	0 × 131 = 0
1690	102	1690 – 1670 = 20	2	102 × 2 = 204
1710	40	1710 – 1670 = 40	4	40 × 4 = 160
1730	17	1730 – 1670 = 60	6	17 × 6 = 102
	Σf = 350			$\Sigma fdx'$ = 314

Let assumed mean A = 1670.

$$i = 10\ dx' = \frac{dx}{i}$$

$$\bar{x} = A + \frac{\Sigma fdx'}{\Sigma f} \times i$$

$A = 1670, \Sigma fdx' = 314; \Sigma f = 350; i = 10.$

$$\therefore \quad \bar{x} = 1670 + \frac{314}{350} \times 10$$

$$= 1670 + 8.97$$

$$\therefore \quad \bar{x} = 1678.97$$

Answer: The arithmetic mean of the above series is = 1678.97.

Calculation of the Arithmetic Mean in a Continuous Series

The continuous series expresse the data which is very vast. The calculation of arithmetic mean of this series is similar to that of discrete series after calculating the mid point of each segment of the continuous series which is called the class interval. The continuous series may have three types of class intervals: (1) Exclusive class interval for example, 10—20, 20—30, 30—40 .... etc. (2) Inclusive class interval for example, 0—9, 10—19, 20—29, 30—39 ... etc. If the data is given in the form of inclusive class intervals, it is first converted into exclusive class interval, (3) Cumulative class interval for example, more than 10, more than 20 ... etc. or less than 10, less than 20 ... etc.

Example 5: For the following data calculate the mean marks obtained by the students using: (i) Short-cut method, (ii) Step deviation method.

Marks	10—20	20—30	30—40	40—50	50—60
Number of Students	1	2	3	5	7
Marks	60—70	70—80	80—90	90—100	
Number of Students	12	16	10	4	

Solution:

Notes

X	f	Mid value (M)	dx = M – A fdx	Step *devi.* dx'	fdx'
10—20	1	15	15 – 55 = – 40 = – 40	– 4	– 4 × 1 = – 4
20—30	2	25	25 – 55 = – 30 = – 60	– 3	– 3 × 2 = – 6
30—40	3	35	35 – 55 = – 20 = – 60	– 2	– 2 × 3 = – 6
40—50	5	45	45 – 55 = – 10 = – 50	– 1	– 1 × 5 = – 5
50—60	7	55	55 – 55 = 00 = 0	0	0 × 7 = 0
60—70	12	65	65 – 55 = 10 = 120	1	1 × 12 = 12
70—80	16	75	75 – 55 = 20 = 320	2	2 × 16 = 32
80—90	10	85	85 – 55 = 30 = 300	3	3 × 10 = 30
90—100	4	95	95 – 55 = 40 = 160	4	4 × 4 = 16
	$\Sigma f = 60$		$\Sigma fdx = 690$		$\Sigma fdx' = 69$

Let assumed mean A = 55.

(i) $\bar{X}$ (by short-cut method) $= A + \frac{\Sigma fdx}{\Sigma f}$

$A = 55; \Sigma fdx = 690; \Sigma f = 60.$

$$\bar{X} = 55 + \frac{690}{60}$$

$$= 55 + 11.5$$

$$\bar{X} = 66.5$$

(ii) $\bar{X}$ (by step deviation method) $= A + \frac{\Sigma fdx'}{\Sigma f} \times i$

$A = 55; \Sigma fdx' = 69; \Sigma f = 60, i = 10.$

$$\bar{X} = 55 + \frac{69}{60} \times 10$$

$$= 66.5$$

Mean marks obtained by students = 66.5.

$$= 17.5 + \frac{-62}{80}$$

$$= 17.5 - 0.775$$

$$= 16.725 \text{ rupees.}$$

Answer: 16.275 approx.

Example 6: Find out the missing frequency in the following distribution:

Marks	No. of Students
0—10	4
10—20	7
20—30	?

Notes

30—40	17
40—50	6
50—60	4

The mean of the distribution is **30.2** marks.

Solution: Let x be the missing frequency.

Marks	M.V. (m)	frequency (f)	$m.f.$
0—10	5	4	20
10—20	15	7	105
20—30	25	x	$25x$
30—40	35	17	595
40—50	45	6	270
50—60	55	4	220
		$n = 38 + x$	$\Sigma mf = 1210 + 25x$

$$a = \frac{\Sigma mf}{n} \text{ or } 30.2 = \frac{1210 + 25x}{38 + x}$$

or $30.2\ (38 + x) = 1210 + 25\ x$

or $1147.6 + 30.2\ x = 1210 + 25\ x$

or $30.2\ x - 25\ x = 1210 - 1147.6$

or $5.2\ x = 62.4$

$x = 12$

The missing frequency is, therefore, 12.

Answer: 12.

Example 7: Calculate the Geometric Mean of the following two series:

Series A	Series B
173	0.8974
182	0.0570
75	0.0081
5	0.5677
0.8	0.0002
0.08	0.0984
0.8974	0.0854
	0.5672

Solution:

Series A		Series B	
Values	**logs**	**Values**	**logs**
173	2.2380	0.8974	$\underline{1}.9530$
185	2.2601	0.0570	$\underline{2}.7559$
75	1.8751	0.0081	$\underline{3}.9085$
5	$\underline{0}.6990$	0.5677	1.7541

Notes

0.8	$\bar{1}$.9031	0.0002	$\bar{4}$.3010
0.08	$\bar{2}$.9031	0.0984	$\bar{2}$.9930
0.8974	1.9530	0.0854	$\bar{2}$.9315
		0.05672	1.7538
N = 7	Σlogs = 5.8314	N = 8	Σlogs = 10.3508

Series A

$$\text{G.M.} = \text{Antilog}\left(\frac{\Sigma \text{logs}}{N}\right) = \text{A.L.} = \left(\frac{5.8314}{7}\right)$$

G.M. = Antilog 0.8331 = 6.810

Series B

$$\text{G.M.} = \text{Antilog}\left(\frac{\Sigma \text{logs}}{N}\right) = \text{AL}\left(\frac{10.3508}{8}\right)$$

$$\text{G.M.} = \text{A.L.}\left(\frac{16 + 6.3508}{8}\right) = \text{AL}\bar{2}.7938 = 0.0622.$$

5.2 Application of Median

The median by definition is the middle value of the distribution. Whenever the median is given as a measure, one-half of the items in the distribution have a value the size of the median value or smaller and one-half have a value the size of the median value or larger.

As distinct from the arithmetic mean which is calculated from the *value of every* item in the series, the median is what is called a *positional* average. The term 'position' refers to the place of a value in a series. The place of the median in a series is such that an equal number of items lie on either side of it. For example, if the income of five persons is 2,700, 2,720, 2,750, 2,760, 2,780, then the median income would be Rs. 2,750. Changing any one or both of the first two values with any other numbers with value of 2,750 or less, and on changing of the last two values to any other values of 2,760 and more, would not affect the values of the median which would remain 2,750. In contrast, in case of arithmetic mean the change in the value of a single item would cause the value of the mean the changed. Median is thus the central value of the distribution or the value that divides the distribution into two equal parts. If there are even number of items in a series there is no actual value exactly in the middle of the series and as such the median is indeterminate. In such a case the median is arbitrarily taken to be halfway between the two middle items. For example, if there are 10 items in a series, the median position is 5.5, that is, the median value is halfway between the value of the items that are 5th and 6th in order of magnitude. Thus when N is odd the median is an actual value with the remainder of the series in two equal parts on either side of it. If N is even then the median is a derived figure, *i.e.,* half the sum of the two middle values.

Calculation of Median Individual Observation

Steps:

(i) Arrange the data in ascending or descending order of magnitude. (Both arrangements would give the same answer.)

(ii) Apply the formula: Median = Size of $\frac{N+1}{2}$th item.

Notes

Individual series

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

Example 1: Calculate median from the following data:

80 60 70 55 95 78 43

Solution: Arranging the given data in ascending order, we get

43 55 60 70 78 80 95

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{7+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{8}{2}\text{th item}$$

Median = The size of 4th item, *i.e.* 70

– Median = 70

Example 2: Compute median from the following data:

74 52 63 45 85 69 55 30

Solution: Arranging the given data in ascending order, we get

30 45 52 55 63 69 74 85

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{8+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{9}{2}\text{th item}$$

Median = The size of 4.5th item,

$$\text{Median} = \text{The size of } \frac{4\text{th item} + 5\text{th item}}{2}$$

$$\text{Median} = \text{The size of } \frac{55+63}{2}\text{th item}$$

$$\text{Median} = \frac{118}{2}$$

– Median = 59

Discrete series

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

Example 3: Find the median wage from the data given below:

Notes

Daily wage (Rs.)	100	150	200	250	300	350	400
No. of workers	10	15	25	30	32	28	10

Solution: Median= The size of $\frac{N+1}{2}$th value

Daily wages (Rs.)	**No. of workers**	
x	*f*	*cf*
100	10	10
150	15	25
200	25	50
250	30	80
300	32	112
350	28	140
400	10	150
	N = 150	

$$\text{Median} = \text{The size of } \frac{N+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{150+1}{2}\text{th item}$$

$$\text{Median} = \text{The size of } \frac{151}{2}\text{th item}$$

Median = The size of 75.5th item, *i.e.* 250

– Median wage = Rs. 250

Continuous series

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times c$$

where

L = Lower limit of the median class

$\frac{N}{2}$ = Half of the total frequency

cf = Cumulative frequency value lies just above the median class

f = Actual frequency lies on the median class

C = Class interval of the median class.

Notes

Example 4: Calculate median mark from the following data:

Marks	0—20	20—40	40—60	60—80	80—100
No. of students	8	16	24	12	40

Solution: Median= $L + \frac{\frac{N}{2} - cf}{f} \times c$

Marks	*f*	*cf*
0—20	8	8
20—40	16	24
40—60	24	48
60—80	12	60
80—100	40	100
	N = 100	

$$\text{Median class} = \frac{N}{2}\text{th class}$$

$$= \frac{100}{2}\text{th class}$$

= 50th class, *i.e.* 60 – 80

$L = 60$

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

$cf = 48$

$f = 12$

$c = 20$

Substituting the values in the formula, we get

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times c$$

$$\text{Median} = 60 + \frac{50 - 48}{12} \times 20$$

$$\text{Median} = 60 + \frac{2}{12} \times 20$$

$$\text{Median} = 60 + \frac{40}{12}$$

Median = 60 + 3.33

– Median mark = 63.33.

Calculation of Median in a Series of Individual Observations

Notes

Example 5: Find the median of the following:

Marks obtained by 11 students	40	42	44	48	52	60	68	70	75	80	82

Solution: Median M= Size of $\left(\frac{N+1}{2}\right)$th item.

$N = 11.$

$\therefore$ $M = \text{Size of}\left(\frac{11+1}{2} = \frac{12}{6} = 6\right)$th item.

The 6th item is 60.

$\therefore$ $M = 60.$

Answer: The median marks of the above data is 60.

Example 6: Find the value of median from the following:

Marks — 10, 9, 19, 21, 25, 32, 11.

Solution: The above data is first rearranged in the ascending order.

Marks — 9, 10, 11, 19, 21, 25, 32.

$$M = \left(\frac{N+1}{2}\right)\text{th item.}$$

$N = 7 \quad \therefore M = \left(\frac{7+1}{2} = \frac{8}{2}\right) = 4\text{th item.}$

$\therefore$ $M = 19.$

Answer: The median marks in the above data is 19.

Example 7: Compute median for the following:

X — 9, 19, 21, 6, 12, 18, 17, 20.

Solution: The data is first rearranged in ascending order:

6, 9, 12, 17, 18, 19, 20, 21

$$M = \left(\frac{N+1}{2}\right)\text{th item. } N = 8.$$

$\therefore$ $M = \left(\frac{8+1}{2} = \frac{9}{2}\right) = 4.5\text{th item.}$

$\therefore$ $M = \frac{\text{Size of 4th item} + \text{Size of 5th item}}{2}$

4th item = 17, 5th item = 18.

$$M = \frac{17+18}{2} = \frac{35}{2} = 17.5$$

Answer: The median of the above data is 17.5.

Notes

Calculation of Median in Discrete Series

Example 8: Find out the value of median from the following data:

Weekly Wages (Rs.)	100	50	70	110	80
Number of Workers	15	20	15	18	12

Solution: The data is first rearranged in ascending order (with respect to X).

X (ascending order)	*f*	**Cumulative frequency *c.f.***
50	20	20
70	15	20 + 15 = 35
80	12	35 + 12 = 47
100	15	47 + 15 = 62
110	18	62 + 18 = 80
	Σf = 80	

$$M = \left(\frac{N+1}{2}\right)^{th} \text{ item. Here } N = \Sigma f = 80.$$

$$\therefore \quad M = \left(\frac{80+1}{2} = \frac{81}{2}\right) = 40.5^{th} \text{ item.}$$

40.5th item would lie in the cumulative frequency (*c.f.*) 47. Therefore the Median = 80.

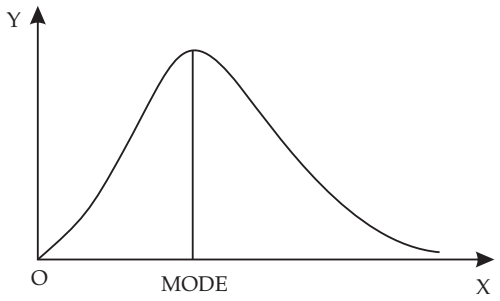
Answer: The median weekly wages = Rs. 80.

5.3 Application of Mode

A third type of "Central value" or "Centre" of the distribution is the value of greatest frequency or, more precisely, of greatest frequency density. Graphically, it is the value on the X-axis below the peak, or highest point of the frequency curve. This average is called the **mode.**

The mode is often said to be the value which occurs most frequently. While this statement is quite helpful in interpreting the mode, it cannot safely be applied to any distribution, because of the vagaries of sampling. Even fairly large samples drawn from a statistical population with a single well-defined mode may exhibit very erratic fluctuations. Hence, mode should be thought as the value which has the greatest density *in its immediate neighbourhood*. For this reason mode is also called the most typical or fashionable value of a distribution.

The following diagram will illustrate the meaning of mode:



The value of the variable at which the curve reaches a maximum is called the mode. It is the value around which the items tend to be most heavily concentrated.

Notes

Although mode is that value which occurs most frequently it does not follow that its frequency represents a majority out of all the total number of frequencies. For example, in the election of college union president the votes obtained by three candidates contesting for presidentship out of a total of 816 votes polled are as follows:

Ramesh	268
Ashok	278
Rakesh	270
Total	816

Mr. Ashok will be elected as president because he has obtained highest votes. But it will be wrong to say that he represents majority because there are more votes against him (268 + 270 = 538) than those for him.

There are many situations in which arithmetic mean and median fail to reveal the true characteristics of data. For example, when we talk of most common wage, most common income, most common height, most common size of shoe or ready-made garments we have in mind mode and the arithmetic mean or median discussed earlier. The mean does not always provide an accurate reflection of the data due to the presence of extreme items. Median may also prove to be quite unrepresentative of the data owing to uneven distribution of the series. For example, the values in the lower half of a distribution range from, say, Rs. 10 to Rs. 100 while the same number of items in the upper half of the series range from Rs. 100 to Rs. 6,000 with most of them near the higher limit. In such a distribution the median value of Rs. 100 will provide little indication of the true nature of the data.

Both these shortcomings may be overcome by the use of mode which refers to the value which occurs most frequently in a distribution. Moreover, mode is simplest to compute since it is the value corresponding to the highest frequency. For example, if the data are:

Size of shoe	5	6	7	8	9	10	11
No. of persons	10	20	25	40	22	15	6

The modal size is '8' since more persons are wearing this size compared to any other size.

Calculation of Mode

Determining the precise value of the mode of a frequency distribution is by no means an elementary calculation. Essentially, it involves fitting mathematically of some appropriate type of frequency curve to the grouped data and the determination of the value on the X-axis below the peak of the curve. However, there are several elementary methods of *estimating* the mode. These methods have been discussed for individual observations, discrete series and continuous series.

Calculation of Mode — Individual Observations

For determining mode count the number of times the various values repeat themselves and the value which occurs the maximum number of times is the modal value. The more often the modal value appears relatively, the more variable the measure is as an average to represent data.

Example 1: Find Mode from the following data:

110, 120, 130, 120, 110, 140, 130, 120, 140, 120

Solution:

Value	Tally Bars	Frequency
110	||	2
120	||||	4
130	||	2
140	||	2
		Total 10

Notes

Since the value 120 occurs the maximum numbers of times, *i.e.*, 4, hence the modal value is 120.



Notes Thus the process of determining mode in case of individual observations essentially involves grouping of data.

When there are two or more values having the same maximum frequency one cannot say which is the modal value and hence mode is said to be ill-defined. Such a series is also known as bimodal or multimodal. For example, observe the following data:

Income (in Rs.) 610, 620, 630, 620, 610, 640, 630, 620, 630, 640.

Size of item	No. of times it occurs
610	2
620	3
630	3
640	2

Mode is ill-defined in this case.

Calculation of Mode — Discrete Series

In discrete series quite often mode can be determined just by inspection, *i.e.,* by looking to that value of the variable around which the items are most heavily concentrated. For example, observe the following data:

Size of garment	No. of persons
28	10
29	20
30	40
31	65
32	50
33	15

From the above data we can clearly say that the modal size is 31 because the value 31 has occurred the maximum number of times, *i.e.,* 65. However, where the mode is determined just by inspection, an error of judgment is possible in those cases where the difference between the maximum frequency and the frequency preceding it or succeeding it is very small and the items are heavily concentrated on either side. In such cases it is desirable to prepare a grouping table and an analysis table. These tables help us ascertaining the modal class.

A grouping table has six columns. In column 1 the maximum frequency is marked or put in a circle; in column 2 frequencies are grouped in two's, in column 3 leave the first frequency and then group remaining in two's; in column 4 group the frequencies in three's; in column 5 leave the first frequency and group the remaining in three's; and in column 6 leave the first two frequencies and then group the remaining in three's. In each of these cases take the maximum total and mark it in a circle or by bold type.

Notes

After preparing the grouping table, prepare an analysis table. While preparing this table put column numbers on the left-hand side and the various probable values of mode on the right-hand side. The values against which frequencies are the highest and marked in the grouping table are then entered by means of a bar in the relevant 'box' corresponding to the values they represent.

The procedure of preparing grouping table and analysis table shall be clear from the following example:

Example 2: From the following data of the weight of 100 persons in a commercial concern determined the modal weight:

Weight (in *kg*)	58	60	61	62	63	64	65	66	68	70
No. of persons	4	6	5	10	20	22	24	6	2	1

Solution: **Grouping Table**

Weight in	**Frequency**					
(*kg*)	**Col. 1**	**Col. 2**	**Col. 3**	**Col. 4**	**Col. 5**	**Col. 6**
58	4	10		15		
60	6		11		21	
61	5	15				35
62	10		30	**52**		
63	20	**42**			**66**	
64	22		**46**			**52**
65	**24**	30		32		
66	6		8		9	
68	2	3				
70	1					

Analysis Table

Col. No.	**Weight in *kg*.**										
	58	60	61	62	63	64	65	66	67	68	70
I							1				
II					1	1					
III						1	1				
IV				1	1	1					
V					1	1	1				
VI						1	1	1			
Total				1	3	5	4	1			

Notes

Since the value 64 has been repeated the maximum number of times, *i.e.*, 5, the modal weight is 64 *kg*. It should be noted that by inspection one is likely to say that the modal value is 65 since it has the highest frequency, *i.e.,* 24. But this is incorrect as revealed by the analysis table and grouping table.

Calculation of Mode — Continuous Series

Steps:

(i) By preparing grouping table and analysis table or by inspection ascertain the modal class.

(ii) Determine the value of mode by applying the following formula:

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i \quad \text{... (i)}$$

where L = Lower limit of the modal class; Δ_1 = the difference between the frequency of the modal class the frequency of the pre-modal class, *i.e.,* preceding class (ignoring signs); Δ_2 = the difference between the frequency of the modal class and the frequency of the post-modal class, *i.e.,* succeeding class (ignoring signs); i = the size of the class-interval of the modal class.

Another form of this formula is:

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \quad \text{... (ii)}$$

where L = Lower limit of the modal class; f_1 = frequency of the modal class, f_0 = frequency of the class preceding the modal class; f_2 = frequency of the class succeeding the modal class.

While applying the above formula for calculating mode it is necessary to see that the class intervals are uniform throughout. It they are unequal they should first be made equal on the assumption that the frequencies are equally distributed throughout the class, otherwise we will get misleading results.



Did u know? In the latter case the value of mode cannot be determined by the above formula and hence mode is *ill-defined.*

A distribution having only one mode is called *unimodal*. If it contains more than one mode, it is called *bimodal* or *multimodal*. If collected data produce a bimodal distribution, the data themselves should be questioned. Quite often such a condition is caused when the size of the sample is small; the difficulty can be remedied by increasing the sample size. Another common cause is the use of non-homogeneous data. Instances where a distribution is bimodal and nothing can be done to change it, the mode should not be used as a measure of central tendency.

Where mode is ill-defined, its value may be ascertained by the following formula based upon the relationship between mean, median and mode.

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean} \quad \text{... (iii)}$$

Notes

Example 3: Calculate mode from the following data:

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

Solution: Since this is a cumulative frequency distribution, we first convert it into a simple frequency distribution.

Marks	No. of Students
0 — 10	3
10 — 20	5
20 — 30	7
30 — 40	10
40 — 50	12
50 — 60	15
60 — 70	12
70 — 80	6
80 — 90	2
90 — 100	8

By inspection the modal class is 50 — 60.

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 50, \Delta_1 = (15 - 12) = 3, \Delta_2 = (15 - 12) = 3, i = 10$$

$$M_0 = 50 + \frac{3}{3+3} \times 10 = 50 + 5 = 55.$$

Example 4: From the following data of the weight of 122 persons determine the modal weight by grouping:

Weight (in *lb.*)	No. of persons	Weight (in *lb.*)	No. of persons
100 — 110	4	140 — 150	33
110 — 120	6	150 — 160	17
120 — 130	20	160 — 170	8
130 — 140	32	170 — 180	2

Solution: By inspection it is difficult to say which is the modal class. Hence, we prepare a grouping table and an analysis table.

Notes

Grouping Table

Weight (in *lb.*)	No. of persons					
	Col. 1	**Col. 2**	**Col. 3**	**Col. 4**	**Col. 5**	**Col. 6**
100 – 110	4	10	26	30	58	**85**
110 – 120	6					
120 – 130	20	**52**				
130 – 140	32		**65**			
140 – 150	**33**	50		**82**	**58**	27
150 – 160	17		25			
160 – 170	8	10				
170 – 180	2					

Analysis Table

Col. No.	Class in which mode is expected to lie		
	120 – 130	**130 – 140**	**140 – 150**
I			1
II	1	1	
III		1	1
IV		1	1
V	1	1	1
VI	1	1	1
Total	3	5	5

This is a bimodal series. Hence mode has to be determined indirectly by applying the formula:

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean.}$$

$$\text{Median} = \text{Size of } \frac{N}{2}\text{th item} = \frac{122}{2} = 61\text{st item.}$$

Hence median lies in the class 130 – 140.

$$\text{Median} = L + \frac{N/2 - c.f.}{f} \times i$$

$L = 130$, $N/2 = 61$; $c.f. = 30$, $f = 32$, $i = 10$.

$$\text{Median} = 130 + \frac{61 - 30}{32} \times 10 = 130 + \frac{310}{32} = 130 + 9.69 = 139.69 \text{ } lb.$$

Notes

Calculation of Mean

Weight in *lb*.	*m*	No. of persons *f*	(*m* – 135)/10 *d*	*fd*
100 — 110	105	4	– 3	– 12
110 — 120	115	6	– 2	– 12
120 — 130	125	20	–1	– 20
130 — 140	135	32	0	0
140 — 150	145	33	+ 1	+ 33
150 — 160	155	17	+ 2	+ 34
160 — 170	165	8	+ 3	+ 24
170 — 180	175	2	+ 4	+ 8
		N = 122		$\sum fd$ = 55

$$\bar{X} = A + \frac{\sum fd}{N} \times i$$

A = 135, $\sum fd$ = 55, N = 122, i = 10

$$\bar{X} = 135 + \frac{55}{122} \times 10 = 135 + 4.51 = 139.51.$$

Mode = 3 Median – 2 Mean.

Mode = (3 × 139.69) – (2 × 139.51) = 419.07 – 279.02 = 140.05

Hence modal weight is 140.05 *lbs*.

Mode when Class Intervals are Unequal

The formula for calculating the value of mode given above is applicable only where there are equal class intervals. If the class intervals are unequal then we must make them equal before we start computing the value of mode. The class interval should be made equal and frequencies adjusted on the assumption that they are equally distributed throughout the class.

Example 5: Calculate the modal income for the following data:

Income (Rs. per month)	No. of Employees
2000 — 2500	8
2500 — 3000	12
3000 — 4000	30
4000 — 4500	3
4500 — 5000	2

Solution: Since class intervals are not equal throughout, we will take 500 as class interval and adjust the frequencies of those classes whose class interval is more than 500. The adjusted frequency distribution is as follows:

Notes

Income (Rs. per month)	No. of Workers
2000 — 2500	8
2500 — 3000	12
3000 — 3500	15
3500 — 4000	15
4000 — 4500	3
4500 — 5000	2

Its is clear from that the mode lies in the class 3000 – 3500.

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 3000, \Delta_1 = (15 - 12) = 3, \Delta_2 = (15 - 15) = 0, i = 20$$

$$M_0 = 3000 + \frac{3}{3+0} \times 500 = 3000 + 500 = 3500$$

Hence modal income = Rs. 3500.

Locating Mode Graphically

In a frequency distribution the value of mode can also be determined graphically. The steps in calculation are:

1. Draw a histogram of the given data.

2. Draw two lines diagonally in the inside of the modal class bar, starting from each upper corner of the bar to the upper corner of the adjacent bar.

3. Draw a perpendicular line from the intersection of the two diagonal lines to the X-axis (horizontal scale) which gives us the modal value.

Example 6: The monthly profits in rupees of 100 shops are distributed as follows:

Profits (Rs.)	No. of shops	Profits (Rs.)	No. of shops
0 — 100	13	300 — 400	20
100 — 200	18	400 — 500	17
200 — 300	27	500 — 600	6

Draw the histogram of the data and hence find the modal value. Check this value by direct calculation.

Solution:

Direct calculation:

Mode lies in the class 200 — 300

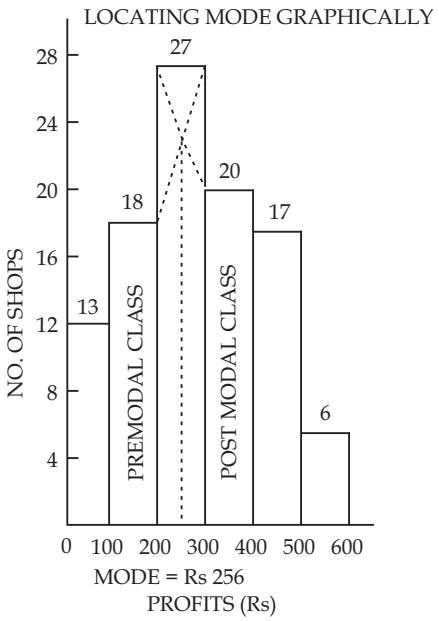
$$\text{Mode} = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

$$L = 200, \Delta_1 = (27 - 18) = 9, \Delta_2 = (27 - 20) = 7, i = 100.$$

$$M_0 = 200 + \frac{9}{9+7} \times 100 = 200 + 56.25 = 256.25.$$

Notes

Mode can also be determined from a frequency polygon in which case a perpendicular is drawn on the base from the apex of the polygon and the point where it meets the base gives the modal value.



However, graphic method of determining mode can be used only where there is one class containing the highest frequency. If two or more classes have the same highest frequency, mode cannot be determined graphically. For example, for the data given below mode cannot be graphically ascertained.

Size of shoe	No. of persons	Size of shoe	No. of persons
2 — 4	10	8 — 10	8
4 — 6	15	10 — 12	2
6 — 8	15		

Self-Assessment

1. Fill in the Blanks

(*i*) Median is better suited for ..... interval series.

(*ii*) In moderately a symmetrical distributions, the distance between the ..... and the ..... is about ..... the distance between the ..... and the ..... .

(*iii*) Given mean 25, mode 24, the median would be ..... .

(*iv*) The mode of distribution is the value that has the greatest ..... of ..... .

(*v*) In a symmetrical distribution mean ..... median ..... mode.

5.4 Summary

- The most popular and widely used measure for representing the entire data by one value is what most laymen call an 'average' and what statisticians call the arithmetic mean. Its value is obtained by adding together all the items and by the dividing this total by the number of items.

Notes

- A discrete series is obtained from a large number of individual observations. Suppose the marks obtained by 100 students is given. This data can be converted into a discrete series where the marks obtained are accompained by the number of students obtaining it.
- The continuous series expresse the data which is very vast. The calculation of arithmetic mean of this series is similar to that of discrete series after calculating the mid point of each segment of the continuous series which is called the class interval.
- As distinct from the arithmetic mean which is calculated from the *value of every* item in the series, the median is what is called a *positional* average. The term 'position' refers to the place of a value in a series. The place of the median in a series is such that an equal number of items lie on either side of it.
- The mode is often said to be the value which occurs most frequently. While this statement is quite helpful in interpreting the mode, it cannot safely be applied to any distribution, because of the vagaries of sampling. Even fairly large samples drawn from a statistical population with a single well-defined mode may exhibit very erratic fluctuations. Hence, mode should be thought as the value which has the greatest density *in its immediate neighbourhood*. For this reason mode is also called the most typical or fashionable value of a distribution.
- Determining the precise value of the mode of a frequency distribution is by no means an elementary calculation. Essentially, it involves fitting mathematically of some appropriate type of frequency curve to the grouped data and the determination of the value on the X-axis below the peak of the curve. However, there are several elementary methods of *estimating* the mode.
- A distribution having only one mode is called *unimodal*. If it contains more than one mode, it is called *bimodal* or *multimodal*. In the latter case the value of mode cannot be determined by the above formula and hence mode is *ill-defined.* If collected data produce a bimodal distribution, the data themselves should be questioned. Quite often such a condition is caused when the size of the sample is small; the difficulty can be remedied by increasing the sample size. Another common cause is the use of non-homogeneous data. Instances where a distribution is bimodal and nothing can be done to change it, the mode should not be used as a measure of central tendency.
- The formula for calculating the value of mode given above is applicable only where there are equal class intervals. If the class intervals are unequal then we must make them equal before we start computing the value of mode. The class interval should be made equal and frequencies adjusted on the assumption that they are equally distributed throughout the class.

5.5 Key-Words

1. Mean : In statistics, mean has three related meanings:
   (*i*) the arithmetic mean of a sample (distinguished from the geometric mean or harmonic mean).
   the expected value of a random variable.
   the mean of a probability distribution.

   There are other statistical measures of central tendency that should not be confused with means - including the 'median' and 'mode'. Statistical analyses also commonly use measures of dispersion, such as the range, interquartile range, or standard deviation. Note that not every probability distribution has a defined mean; see the Cauchy distribution for an example.
2. Median: In statistics and probability theory, median is described as the numerical value separating the higher half of a sample, a population, or a probability distribution, from the lower

Notes

half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, then there is no single middle value; the median is then usually defined to be the mean of the two middle values. A median is only defined on one-dimensional data, and is independent of any distance metric. A geometric median, on the other hand, is defined in any number of dimensions.

In a sample of data, or a finite population, there may be no member of the sample whose value is identical to the median (in the case of an even sample size); if there is such a member, there may be more than one so that the median may not uniquely identify a sample member. Nonetheless, the value of the median is uniquely determined with the usual definition. A related concept, in which the outcome is forced to correspond to a member of the sample, is the medoid…

3. Mode : The mode is the value that appears most often in a set of data.

Like the statistical mean and median, the mode is a way of expressing, in a single number, important information about a random variable or a population. The numerical value of the mode is the same as that of the mean and median in a normal distribution, and it may be very different in highly skewed distributions. The mode is not necessarily unique, since the same maximum frequency may be attained at different values. The most extreme case occurs in uniform distributions, where all values occur equally frequently. The mode of a discrete probability distribution is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value, so, informally speaking, the mode is at the peak.

5.6 Review Questions

1. Give two examples where arithmetic mean and median would be most appropriate average.
2. Can the value of mean, mode and median be the same in a symmetrical distribution ? If yes, state the situation.
3. Discuss the application mean and median.
4. How do you determine median and mode graphically ?
5. 'The arithmetic mean is the best among all the averages.' Give reasons.

Answers: Self-Assessment

1. (*i*) Positional (*ii*) mean, median, 1/3, mean, mode
   (*iii*) 24.67 (*iv*) concentration, frequencies
   (*v*) is equal to, is equal to

5.7 Further Readings



Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 6: Dispersion: Meaning and Characteristics, Absolute and Relative Measures of Dispersion including Range, Quartile Deviation and Percentile

CONTENTS

Objectives
Introduction
6.1 Meaning and Characteristics of Dispersion
6.2 Absolute and Relative Measures of Dispersion
6.3 Range, Quartile Deviation and Percentile
6.4 Summary
6.5 Key-Words
6.6 Review Questions
6.7 Further Readings

Objectives

After reading this unit students will be able to:

- Know the Meaning and Characteristics of Dispersion.
- Explain Absolute and Relative Measures of Dispersion.
- Discuss Range, Quartile Deviation and Percentile.

Introduction

Series of date definitely have a great utility but they fail to reveal many facts about the phenomenon. There may be many different series, whose average/mean may come out to be identicle. But when they are studied in depth, they reveal entirely different stories. For example, the income of 5 people is — Rs. 50, 50, 50, 50, 50. Then the average will come out to be Rs. 50. The incomes of another five (5) people are Rs. 20, 80, 25, 25, 100. The average would again come out to be Rs. 50. Now if we consider incomes of another five people, they are Rs. 150, 20, 10, 10, 60. This would again average to Rs. 50 only. But in all the three cases, the average does not seem to represent the data fully. In the first case, the incomes are equal, in the second case, they have less variations in income but in the third case, there is vast variation of income. Therefore concluding about the data only on the basis of averages, considering it to be representative of the series may be misleading. Therefore, there is a need to measure the variations in the data. These variations are also called dispersion. Measures of central tendency are based on items of a series, therefore, they are called 'averages of the first order'. Measures of dispersion, on the other hand, are based upon average of the deviations of the different values from mean.

They are therefore called 'averages of the second order'.

6.1 Meaning and Characteristics of Dispersion

Meaning and Definition of Dispersion

The term dispersion refers to the variability of the size of items. Dispersion explains the size of various items in a series are not uniform rather, they vary. For example, if in a series the lowest and highest values vary only a little, the dispersion is said to be low. But if this variation is very high, dispersion is said to be considerable. In a series of ten students, the marks obtained are 10, 6, 8, 5, 10, 10, 8, 10, 5, 8.

Notes

(the average = 8). In another class, 10 students obtained the following marks. 10, 10, 5, 2, 10, 10, 3, 10, 10 (the average = 8). The dispersion in the second case is more because the size of items in this series vary considerably, inspite of the fact that the averages of the two have come out to be 8. Some of the important definitions of dispersion are — As per Brooks and Dick, "Dispersion or spread is the degree of the scatter or variations of the variable about a central value." A. L. Bowley defines dispersion as — "Dispersion is the measure of variations of the item." In the words of Prof. L. R. Connor, "Dispersion is a measure of the extent to which the individual items vary." According to Spriegel, "The degree to which numerical data tend to spread about an average value is called the variation or dispersion of data."

All the above definitions suggest that the term dispersion refers to the variability in the size of items. This variability is measured with respect to the average of the series. Therefore measures of dispersion are also termed as averages of the second order.



Did u know? "A measure of variation or dispersion describes the degree of scatter shown by observations and is usually measured by comparing the individual values of the variable with the average of all the values and then calculating the average of all the individual differences.

Characteristics of Dispersion

There are four basic characteristics of dispersion:

(1) **To guage the reliability of the average:** Even after making all the efforts to obtain the most representative average, the efforts prove to be successful when the data is homogeneous. In the absence of homogeneity, a measure of dispersion presents a better description about the structure of the distribution and the place of individual items in it. Therefore, in case of heterogeneous data, dispersion is measured to guage the reliability of the average calculated. When the value of dispersion is small, it is concluded that the average closely represents the data but when value of dispersion comes out to be large, it should be concluded that the average obtained is not very reliable.

(2) **To make a comparative study of the variability of series:** The consistency of uniformity of two series can be compared with the help of dispersion. If the value of dispersion measured comes out to be large, it may be concluded that the series lacks uniformity or consistency. Such studies are very useful in many fields like profit of companies, share values, performance individuals, studies related to demand, supply, prices etc.

(3) **To identify the factors causing variability so that it can be controlled:** Another important purpose of calculating dispersion is to identify the nature and causes of variations in a given data so that measures to control these can be suggested. Thus measures of dispersion are not merely supplementary to the averages, describing their reliability rather, they significantly disclose the quality of data in terms of homogeneity and consistency. They help to evaluate the various causes of heterogeneity and inconsistencies and suggest ways to control these. For example, in industrial production, efficient operation requires control of variation, the causes of which are sought through, inspection and quality control programmes." In social sciences, the measurement of inequality in the distribution of income and wealth requires the measures of variation.

(4) **To serve as a basis for further statistical analysis:** Yet another purpose of measures of dispersion is to help the statistician in carrying out further statistical analysis of the data like studying correlation, regression, testing of hypothesis, analysis of time series etc.

On the basis of the above, it can be concluded that due to inconsistencies and lack of uniformity of the data, averages can not prove to be closely representing the data, in most of the cases. In such a situation, dispersion presents a more better picture about the data, and gives logic to

Notes

find out whether the average calculated is reliable or not. It also helps in comparing the two series and also help in finding out ways to control the variations. In this way dispersion is a very strong tool into the hands of statisticians to know about the structure of data more closely and reliably.

Properties of a Good Measure of Dispersion

Just like the properties of a good measure of central tendency, properties of a good measure of dispersion are:

W. A. Sppur and *C. P. Bonim: Statistical Analysis for Business Decision.*

(1) It should be simple to understand.
(2) It must be easy to calculate.
(3) It must be based on all the items of the series.
(4) It should not be unduly affected by the extreme items.
(5) It should be least affected by the fluctuations in sampling.
(6) It should be capable of further statistical treatment.

6.2 Absolute and Relative Measures of Dispersion

Absolute measures of dispersion: When the dispersion of a series is calculated in terms of the absolute or actual figures in the data and the value of dispersion obtained can be expressed in the same units as the items of data are expressed, such measures are called absolute measures of dispersion. For example, if we calculate dispersion of a series indicating the income of group of persons in rupees, and the value of dispersion is obtained in rupees, it is termed as absolute measure of dispersion.

Relative measures of dispersion: When the value of dispersion is calculated as ratio or percentage of the average it is called relative measure of dispersion.

6.3 Range, Quartile Deviation and Percentile

Range

'Range' is the simplest measure of dispersion which is determined by the two extreme values of the observations and it is the difference between the largest and the smallest value in a distribution.

Uses of Range: (1) Range is very useful in quality control measures taken by the production department. It is checked that the quality should not deteriorate beyond the set value of range. Control charts are prepared for the purpose. (2) Another area where 'range' is very useful is the study of fluctuation of data. Variations in the weather forecasts, movement in the prices of securities etc. can be studied effectively and efficiently with the help of range.

Merits/advantages: (1) Simple to understand and easy to calculate, (2) It presents a broad picture of the data.

Demerits/disadvantages: (1) Gets affected by the extreme items, (2) does not take into consideration towards most of the items and their deviations, (3) Does not give reasonable picture of the data, (4) It is influenced by fluctuations of sampling.

Formula

$$\text{Range} = \text{Largest Value} - \text{Smallest Value}$$

or

$$\text{Maximum} - \text{Minimum}$$

or

$$\text{Range} = L - S$$

or

$$M_1 - M_0$$

Notes

$$\text{Coefficient of Range} = \frac{\text{Largest Value} - \text{Smallest Value}}{\text{Largest Value} + \text{Smallest Value}}$$

or

$$\frac{L - S}{L + S}$$

Individual Series

Example 1: Find out absolute and relative dispersion of range from the following observations:

Marks: 63, 68, 71.5, 83, 50, 27, 64, 38, 40

Solution: Absolute measure of Range = Largest – Smallest

= 83 – 27 = 56 marks

$$\text{Relative Measure} = \frac{\text{Largest} - \text{Smallest}}{\text{Largest} + \text{Smallest}}$$

$$= \frac{83 - 27}{83 + 27} = \frac{56}{110} = 0.509$$

Example 2: Calculate Range and its coefficient of the following series:

S. No.	1	2	3	4	5	6	7	8	9	10
Values	391	384	591	407	672	522	777	733	2488	1490

Solution: Largest = 2488, Smallest = 384,

Range = L – S

2488 – 384

= 2104

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

$$= \frac{2488 - 384}{2488 + 384} = \frac{2104}{2872} = 0.7325$$

Example 3: The yearly income of a person for the last ten years in given below. Find the range and its coefficient.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Income ('000 Rs.)	40	30	80	100	80	90	120	110	130	150

Solution: Range = L – S from the data, L = 150, S = 30.

$\therefore$ Range = 150 – 30 = 120

$$\text{Coefficient of Range} = \frac{L - S}{L + S}$$

or

$$= \frac{150 - 30}{150 + 30} = \frac{120}{180}$$

= 0.66

Notes

Answer: The range for the above data is Rs. 1,20,000 per year and the coefficient of range is 0.66.

Range in Discrete Series

Example 4 : Find out absolute and relative measure of range in the following distribution:

Scores	Frequency
2	6
3	7
4	3
5	11
6	4
7	2
8	5
9	8
10	3

Solution: Absolute Measure of Range = L – S

= 10 – 2 = 8

$$\text{Relative Measure} = \frac{L - S}{L + S}$$

$$= \frac{10 - 2}{10 + 2} = \frac{8}{12} = 0.666$$

Example 5: Calculate Coefficient of Range of the following series:

Size	Frequency
2.5	7
3.5	1
4.5	3
5.5	5
6.5	4
7.5	9
8.5	8
9.5	11
10.5	4

Solution : L = 10.5

S = 2.5

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{10.5 - 2.5}{10.5 + 2.5} = \frac{8}{13} = 0.615$$

Range in Continuous Series

Range can be calculated by the following two methods:

(1) Class mark of the highest class – class mark of the lowest class.

(2) Upper class boundary of the highest class – lower class boundary of the lowest class.

Notes

Example 6: Calculate Coefficient of range from the following distribution:

Marks	No. of Students
25—30	6
30—35	3
35—40	12
40—45	8
45—50	22
50—55	9
55—60	5

Solution: The above question can be calculated by:

(1) Limits Method: L = 60, S = 25

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{60 - 25}{60 + 25} = \frac{35}{85} = 0.411$$

(2) Mid-Value Method: L = 57.5, S = 27.5

$$\text{Coefficient of Range} = \frac{L - S}{L + S} = \frac{57.5 - 27.5}{57.5 + 27.5} = \frac{30}{85} = 0.352$$

Example 7: Find out relative dispersion of range from the following frequency table:

Marks	Frequency
5—10	4
10—15	6
15—20	20
20—25	7
25—30	5
30—35	8
35—40	6
40—45	5
45—50	2

Solution: For the calculation of relative dispersion of range, mean, median and mode will be calculated.

Marks	*f*	*c.f.*	*m.v.*	*dx*	*fdx*
5—10	4	4	7.5	– 5	– 20
10—15	6	10	12.5	– 4	– 24
15—20	20	30	17.5	– 3	– 60
20—25	7	37	22.5	– 2	– 14
25—30	5	42	27.5	–1	– 7
30—35	8	50	32.5	0	0
35—40	6	56	37.5	1	6
40—45	5	61	42.5	2	10
45—50	2	63	47.5	3	6
	N = 63				Σfdx = – 103

Notes

$$\text{Mean} = x + \frac{\sum fdx}{n} \times i$$

$$= 32.5 + \frac{-103}{63} \times 5$$

$$= 32.5 + \frac{-515}{63}$$

$$= 32.5 - 8.17 = 24.33 \text{ approx.}$$

$$\text{Median No.} = \frac{N}{2} = \frac{63}{2} = 31.5$$

$$\text{Median} = l + \frac{i}{f}(m - c)$$

$$= 20 + \frac{5}{7}(31.5 - 30)$$

$$= 20 + \frac{7.5}{7}$$

$$= 20 + 1.07 = 21.07$$

$$\text{Mode} = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

$$= 15 + \frac{20 - 6}{40 - 6 - 7} \times 5$$

$$= 15 + \frac{14 \times 5}{27}$$

$$= 15 + \frac{70}{27} = 15 + 2.59 = 17.59$$

Coefficient of Range Dispersion

(1) By using Mean $\frac{50 - 5}{24.33} = \frac{45}{24.33} = 1.84$

(2) By using Double the Mean $\frac{50 - 5}{24.33 \times 2} = \frac{45}{48.66} = 0.92$

(3) By using Median $\frac{50 - 5}{21.07} = \frac{45}{21.07} = 2.13$

(4) By using Mode $\frac{50 - 5}{17.59} = \frac{45}{17.59} = 2.55$

(5) By using sum of the extremes $\frac{50 - 5}{50 + 5} = \frac{45}{55} = 0.81$

Quartile Deviation or Semi-Interquartile Range

Notes

Quartile is the location-based measure of dispersion. It measures the average amount by which the first and the third quartiles deviate from the second quartile *i.e.*, median.

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}, \text{ Coefficient of variation} = \frac{\text{Q.D.}}{\text{Median}} \times 100,$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

Merits of Q.D.: (1) Easy to compute. (2) It is very useful to know the variability at the centre of the data. (3) It is not much affected by the extreme items. (4) It can be calculated from open end distribution or from a skewed distribution.

Demerits of Q.D.: (1) Based only on the middle part of the data. (2) It is not capable of further mathematical treatment. (3) It is greatly affected by changes in sampling. (4) Gives no indication about variation occurring beyond Q_3 and Q_1.

Third Moment of Dispersion: In this method the deviations of items from mean are cubed, *i.e.,*

$$\text{Third moment of dispersion} = \frac{\sum d^3}{N}$$

$$\text{Coefficient of third moment of dispersion} = \frac{\frac{\sum d^3}{N}}{\sigma}$$

(5) It provides unit of measurement for the normal distribution.

Demerits: (1) If the data is vast, it involves tedious calculations.

Formula

$$\text{Quartile Deviation (Q.D.)} = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

where Q_1 represents First Quartile

Q_3 represents Third Quartile

Q.D. in Individual Series

Example 8: Calculate Q.D. and its coefficient from the following observations relating to marks of 15 students:

48, 52, 56, 62, 66, 47, 51, 58, 60, 66, 68, 70, 64, 73, 63.

Solution: **Array:** 47, 48, 51, 52, 56, 58, 60, 62, 63, 64, 66, 66, 68, 70, 73

$$Q_1 = \text{Value of the } \left[\frac{N+1}{4}\right]^{th} \text{ item}$$

$$= \left[\frac{15+1}{4}\right]^{th} \text{ item } i.e., \ 4^{th} \text{ item} = 52$$

Notes

$$Q_3 = \text{Value of the } \left[3\left(\frac{N+1}{4}\right)\right]^{th} \text{ item}$$

$$= \left[\frac{3(15+1)}{4}\right]^{th} \text{ item } i.e., 12^{th} \text{ item} = 66$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{66 - 52}{2} = \frac{14}{2} = 7$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{66 - 52}{66 + 52} = \frac{14}{118} = 0.118$$

Q.D. in Discrete Series

Example 9: Calculate quartile deviation and its coefficient from the following data:

Height of students (in *cms*.)	120	122	124	126	130	140	150	160
No. of students	1	3	5	7	10	3	1	1

Solution:

Calculation of Q_3 and Q_1

X	***f***	***c.f.***
120	1	1
122	3	4
124	5	9
126	7	16
130	10	26
140	3	29
150	1	30
160	1	31

$$Q_1 = \text{Size of } \left(\frac{N+1}{4}\right)^{th} \text{ item} = 8^{th} \text{ item}$$

$$\therefore \quad Q_1 = 224.$$

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right)^{th} \text{ item} = 24^{th} \text{ item}$$

$$\therefore \quad Q_3 = 230$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{230 - 224}{2} = 3 \text{ } cms.$$

Notes

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{230 - 224}{230 + 224} = 0.01321.$$

Answer: Q.D. for the above data is found to be 3 *cms.* and coefficient of Q.D. = 0.01321.

Q.D. in Continuous Series

Example 10: Find quartile deviation and its relative measure:

Variable	Frequency	Variable	Frequency
20—29	306	50—59	96
30—39	182	60—69	42
40—49	144	70—79	32

Solution:

Calculation OF Q.D.

Variable	*f*	*c.f.*
20—29	306	306
30—39	182	488
40—49	144	632
50—59	96	728
60—69	42	770
70—79	34	804

$$Q_1 = \left(\frac{805}{4}\right)^{th} \text{ item} = 201.25^{th} \text{ item}$$

which lies in class 20—29 or class 19.5 – 29.5

$$Q_1 = 19.5 + \left(\frac{10}{306} \times 201\right) = 26.07$$

$$Q_3 = \frac{3(805)^{th}}{4} \text{ item} = 603.75^{th} \text{ item.}$$

which lies in 40—49 class or 39.5 — 49.5.

$$Q_3 = 39.5 + \left(\frac{10}{144} + 155\right) = 47.49$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{47.49 - 26.07}{2} = 10.71$$

Notes

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

$$= \frac{47.49 - 26.07}{47.49 + 26.07} = 0.2912$$

Answer: For the given data Q.D. = 10.71 and coeff. of Q.D. = 0.2912.

Example 11: Compute Quartile Deviation and its coefficient from the following data:

Mid-Value	3	4	5	6	7	8	9
Frequency	11	14	20	24	20	16	5

Solution:

Class	*f*	*C.f.*
2.5—3.5	11	11
3.5—4.5	14	25
4.5—5.5	20	45
5.5—6.5	24	69
6.5—7.5	20	89
7.5—8.5	16	105
8.5—9.5	5	110
Total	**N = 110**	

$$Q_1 = \text{Size of } \frac{N}{4} = \frac{110}{4} = 27.5^{th} \text{ item}$$

27.5^{th} item which lies in 4.5—5.5 group

$$Q_1 = l_1 + \frac{i}{f}(q_1 - c)$$

$$= 4.5 + \frac{1}{20}(27.5 - 25)$$

$$= 4.5 + \frac{2.5}{20}$$

$$= 4.5 + 0.125 = 4.625$$

$$Q_3 = \text{Size of } \frac{3N}{4} = \frac{3 \times 110}{4} = \frac{330}{4} = 82.5^{th} \text{ item}$$

82.5^{th} item which lies in 6.5 — 7.5 group.

$$Q_3 = l_1 + \frac{i}{f}(q_3 - c)$$

$$= 6.5 + \frac{1}{20}(82.5 - 69)$$

Notes

$$= 6.5 + \frac{13.5}{20}$$

$$= 6.5 + 0.675 = 7.175$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{7.175 - 4.625}{2} = \frac{2.55}{2} = 1.275$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{7.175 - 4.625}{7.175 + 4.625}$$

$$= \frac{2.55}{11.8} = 0.216$$

Example 12: For a distribution, the coefficient of quartile deviation = 0.4, and the difference of two quartiles = 40. Find the values of quartiles.

Solution: Given: C of Q.D. = 0.4, $Q_3 - Q_1 = 40$, $Q_1 = ?$, $Q_3 = ?$

$$\text{C of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

or $$= 0.4 = \frac{40}{Q_3 + Q_1}$$

or $$Q_3 + Q_1 = \frac{40}{0.4} = 100$$

$$Q_3 - Q_1 = 40$$

$$\frac{Q_3 + Q_1 = 100}{2Q_3 = 140}$$

$\therefore$ $$Q_3 = 140/2 = 70, \; Q_1 = 100 - 70 = 30$$

Example 13: From the following data, calculate Quartile Coefficient of Dispersion.

Wages Less than	10	20	30	40	50	60	70
No. of Workers	5	8	15	20	30	33	35

Solution:

Wages in Rs.	**No. of workers *(f)***	***c.f.***
0—10	5	5
10—20	3	8
20—30	7	15
30—40	5	20
40—50	10	30
50—60	3	33
60—70	2	35
Total	**N = 35**	

Notes

$$Q_1 = \text{Size of } \frac{N}{4} = \frac{35}{4} = 8.75^{th} \text{ item}$$

8.75^{th} item which lies in 20—30 group.

$$Q_1 = l_1 + \frac{i}{f}(q_1 - c)$$

$$= 20 + \frac{10}{7}(8.75 - 8)$$

$$= 20 + \frac{10 \times .75}{7}$$

$$= 20 + 1.07 = 21.07$$

$$Q_3 = \text{Size of } \frac{3N}{4} = \frac{3 \times 35}{4} = \frac{105}{4} = 26.25^{th} \text{ item}$$

26.25^{th} item which lies in 40—50 group

$$Q_3 = l_1 + \frac{i}{f}(q_3 - c)$$

$$= 40 + \frac{10}{10}(26.25 - 20)$$

$$= 40 + \frac{10 \times 6.25}{10}$$

$$= 40 + 6.25 = 46.25$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{46.25 - 21.07}{2} = \frac{25.18}{2} = 12.59$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{46.25 - 21.07}{46.25 + 21.07} = \frac{25.18}{67.32}$$

$$= 0.374$$

Percentile

In statistics, **a percentile** (or centile) is the value of a variable below which a certain percent of observations fall. For example, the 20^{th} percentile is the value (or score) below which 20 percent of the observations may be found. The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is in the 86^{th} percentile, it is higher than 85% of the other scores.

The 25^{th} percentile is also known as the first quartile (Q_1), the 50^{th} percentile as the median or second quartile (Q_2), the 75^{th} percentile as the third quartile (Q_3).

Notes

Definition

There is no standard definition of percentile, however all definitions yield similar results when the number of observations is very large.

Nearest rank

One definition of percentile, often given in texts, is that the P-th percentile $(0 \le P < 100)$ of N ordered values (arranged from least to greatest) is obtained by first calculating the (ordinal) rank

$$n = \frac{P}{100} \times N + \frac{1}{2}$$

rounding the result to the nearest integer, and then taking the value that corresponds to that rank.

(Note that the rounded value of n is just the least integer which exceeds $\frac{P}{100} \times N$.)

For example, by this definition, given the numbers

15, 20, 35, 40, 50

the rank of the 30th percentile would be

$$n = \frac{30}{100} \times 5 + \frac{1}{2} = 2.$$

Thus the 30th percentile is the second number in the sorted list, 20.

The 35th percentile would have rank

$$n = \frac{35}{100} \times 5 + \frac{1}{2} = 2.25,$$

so the 35th percentile would be the second number again (since 2.25 rounds down to 2) or 20

The 40th percentile would have rank

$$n = \frac{40}{100} \times 5 + \frac{1}{2} = 2.5,$$

so the 40th percentile would be the third number (since 2.5 rounds up to 3), or 35.

The 100th percentile is defined to be the largest value. (In this case we do not use the above definition with P = 100, because the rank n would be greater than the number N of values in the original list.)

In lists with fewer than 100 values the same number can occupy more than one percentile group.

Linear interpolation between closest ranks

An alternative to rounding used in many applications is to use **linear interpolation** between the two nearest ranks.

In particular, given the N sorted values $v_1 \le v_2 \le v_3 \le \ldots \le v_N$, we define the *percent rank* corresponding to the n^{th} value as:

$$p_n = \frac{100}{N}\left(n - \frac{1}{2}\right).$$

In this way, for example, if N = 5th percent rank corresponding to the third value is

$$p_3 = \frac{100}{5}\left(3 - \frac{1}{2}\right) = 50.$$

Notes

The value v of the P-th percentile may now be calculated as follows:

If $P < p_1$ or $P > p_N$, then we take $v = v_1$ or $v = v_N$, respectively.

If there is some integer k for which $P = p_k$, then we take $v = v_k$.

Otherwise, we find the integer k for which $p_k < P < p_{k+1}$, and take $v = v_k + \frac{P - p_k}{p_{k+1} - p_k}(v_{k+1} + v_k) =$

$v_k + N \times \frac{P - p_k}{100}(v_{k+1} - v_k)$.

Using the list of numbers above, the 40th percentile would be found by linearly interpolating between the 30th percentile, 20, and the 50th, 35. Specifically:

$$v = 20 + 5 \times \frac{40 - 30}{100}(35 - 20) = 27.5$$

This is halfway between 20 and 35, which one would expect since the rank was calculated above as 2.5.

It is readily confirmed that the 50th percentile of any list of values according to this definition of the P-th percentile is just the sample median.

Moreover, when N is even the 25th percentile according to this definition of the P-th percentile is the median of the first $\frac{N}{2}$ values (*i.e.*, the median of the lower half of the data).

Weighted percentile

In addition to the percentile function, there is also a *weighted percentile*, where the percentage in the total weight is counted instead of the total number. There is no standard function for a weighted percentile. One method extends the above approach in a natural way.

Suppose we have positive weights $w_1, w_2, w_3, ..., w_N$ associated, respectively, with our N sorted sample values. Let

$$S_n = \sum_{k=1}^{n} w_k ,$$

the n-th **partial sum** of the weights. Then the formulas above are generalized by taking

$$p_n = \frac{100}{S_N}\left(S_n - \frac{w_n}{2}\right)$$

and

$$v = v_k + \frac{p - p_k}{p_{k+1} - p_k}(v_{k+1} + v_k) .$$

Alternative methods

Some software packages, including **Microsoft Excel** (up to the version 2007) use the following method, noted as an alternative by NIST to estimate the value, v_p, of the P-th percentile of an ascending ordered dataset containing N elements with values $v_1 \le v_2 \le \ldots \le v_N$.

The rank is calculated:

$$n = \frac{P}{100}(N - 1) + 1$$

Notes

and then split into its integer component k and decimal component d, such that $n = k + d$.

Then v_p is calculated as:

$$v_p = \begin{cases} v_1, & \text{for } n = 1 \\ v_N, & \text{for } n = N \\ v_k + d(v_{k+1} - v_k), & \text{for } 1 =< n < N \end{cases}$$

The primary method recommended by **NIST** is similar to that given above, but with the rank calculated as

$$n = \frac{P}{100}(N + 1)$$

These two approaches give the rank of the 40th percentile in the above example as, respectively:

$$n = \frac{40}{100}(5 - 1) + 1 = 2.6$$

and

$$n = \frac{40}{100}(5 + 1) = 2.4$$

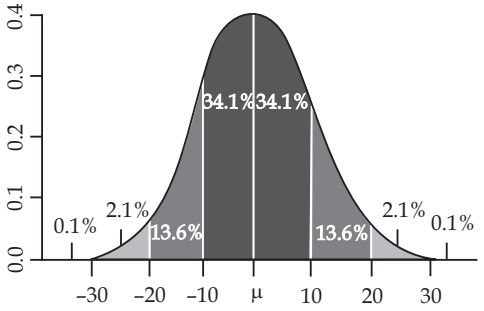
The values are then interpolated as usual based on these ranks, yielding 29 and 26, respectively, for the 40th percentile.

Applications

When **ISPs** bill **"burstable" internet bandwidth,** the 95th or 98th percentile usually cuts off the top 5% or 2% of bandwidth peaks in each month, and then bills at the nearest rate. In this way infrequent peaks are ignored, and the customer is charged in a fairer way. The reason this statistic is so useful in measuring data through put is that it gives a very accurate picture of the cost of the bandwidth. The 95th percentile says that 95% of the time, the usage is below this amount. Just the same, the remaining 5% of the time, the usage is above that amount.

Physicians will often use infant and children's **weight and height percentile** to assess their growth in comparison to national averages.

The normal curve and percentiles



The dark blue zone represents observations within one **standard deviation** (σ) to either side of the **mean** (μ), which accounts for about 68.2% of the population. Two standard deviations from the mean (dark and medium blue) account for about 95.4%, and three standard deviations (dark, medium, and light blue) for about 99.7%.

Notes

The methods given above are approximations for use in small-sample statistics. In general terms, for very large populations percentiles may often be represented by reference to a **normal curve** plot. The normal curve is plotted along an axis scaled to **standard deviation,** or sigma, units. Mathematically, the normal curve extends to negative **infinity** on the left and positive infinity on the right. Note, however, that a very small portion of individuals in a population will fall outside the – 3 to + 3 range.

In humans, for example, a small portion of all people can be expected to fall above the + 3 sigma height level.

Percentiles represent the area under the normal curve, increasing from left to right. Each standard deviation represents a fixed percentile. Thus, rounding to two decimal places, – 3 is the 0.13^{th} percentile, – 2 the 2.28^{th} percentile, – 1 the 15.87^{th} percentile, 0 the 50^{th} percentile (both the mean and median of the distribution), + 1 the 84.13^{th} percentile, + 2 the 97.72^{nd} percentile, and + 3 the 99.87^{th} percentile. Note that the 0^{th} percentile falls at negative infinity and the 100^{th} percentile at positive infinity.

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) A good measure of dispersion is the one which is not defined rigidly.

(*ii*) Range is the best measure of dispersion.

(*iii*) Quartile Deviation is more suitable in case of openend distributions.

(*iv*) Absolute measure of dispersion can be used for purposes of comparison.

6.4 Summary

- The term dispersion refers to the variability of the size of items. Dispersion explains the size of various items in a series are not uniform rather, they vary. For example, if in a series the lowest and highest values vary only a little, the dispersion is said to be low.
- "A measure of variation or dispersion describes the degree of scatter shown by observations and is usually measured by comparing the individual values of the variable with the average of all the values and then calculating the average of all the individual differences.
- Therefore, in case of heterogeneous data, dispersion is measured to guage the reliability of the average calculated. When the value of dispersion is small, it is concluded that the average closely represents the data but when value of dispersion comes out to be large, it should be concluded that the average obtained is not very reliable.
- The consistency of uniformity of two series can be compared with the help of dispersion. If the value of dispersion measured comes out to be large, it may be concluded that the series lacks uniformity or consistency. Such studies are very useful in many fields like profit of companies, share values, performance individuals, studies related to demand, supply, prices etc.
- It can be concluded that due to inconsistencies and lack of uniformity of the data, averages can not prove to be closely representing the data, in most of the cases. In such a situation, dispersion presents a more better picture about the data, and gives logic to find out whether the average calculated is reliable or not. It also helps in comparing the two series and also help in finding out ways to control the variations. In this way dispersion is a very strong tool into the hands of statisticians to know about the structure of data more closely and reliably.
- When the dispersion of a series is calculated in terms of the absolute or actual figures in the data and the value of dispersion obtained can be expressed in the same units as the items of data are expressed, such measures are called absolute measures of dispersion. For example, if we calculate dispersion of a series indicating the income of group of persons in rupees, and the value of dispersion is obtained in rupees, it is termed as absolute measure of dispersion.

Notes

- Quartile is the location-based measure of dispersion. It measures the average amount by which the first and the third quartiles deviate from the second quartile *i.e.*, median.
- In statistics, **a percentile** (or centile) is the value of a variable below which a certain percent of observations fall. For example, the 20th percentile is the value (or score) below which 20 percent of the observations may be found. The term percentile and the related term percentile rank are often used in the reporting of scores from norm-referenced tests. For example, if a score is in the 86th percentile, it is higher than 85% of the other scores.
- In addition to the percentile function, there is also a *weighted percentile*, where the percentage in the total weight is counted instead of the total number. There is no standard function for a weighted percentile. One method extends the above approach in a natural way.
- When **ISPs** bill **"burstable" internet bandwidth,** the 95th or 98th percentile usually cuts off the top 5% or 2% of bandwidth peaks in each month, and then bills at the nearest rate. In this way infrequent peaks are ignored, and the customer is charged in a fairer way. The reason this statistic is so useful in measuring data through put is that it gives a very accurate picture of the cost of the bandwidth. The 95th percentile says that 95% of the time, the usage is below this amount. Just the same, the remaining 5% of the time, the usage is above that amount.
- In general terms, for very large populations percentiles may often be represented by reference to a **normal curve** plot. The normal curve is plotted along an axis scaled to **standard deviation,** or sigma, units. Mathematically, the normal curve extends to negative **infinity** on the left and positive infinity on the right. Note, however, that a very small portion of individuals in a population will fall outside the – 3 to + 3 range.
- Percentiles represent the area under the normal curve, increasing from left to right. Each standard deviation represents a fixed percentile. Thus, rounding to two decimal places, – 3 is the 0.13th percentile, – 2 the 2.28th percentile, – 1 the 15.87th percentile, 0 the 50th percentile (both the mean and median of the distribution), + 1 the 84.13th percentile, + 2 the 97.72nd percentile, and + 3 the 99.87th percentile. Note that the 0th percentile falls at negative infinity and the 100th percentile at positive infinity.

6.5 Key-Words

1. Absolute measures : Absolute measures of Dispersion are expressed in same units in which original data is presented but these measures cannot be used to compare the variations between the two series. Relative measures are not expressed in units but it is a pure number. It is the ratios of absolute dispersion to an appropriate average such as co-efficient of Standard Deviation or Co-efficient of Mean Deviation.

   Relative measures : These measures are calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measured. If the original data is in dollar or kilometers, we do not use these units with relative measure of dispersion. These measures are a sort of ratio and are called coefficients. Each absolute measure of dispersion can be converted into its relative measure.

2. Quartile deviation : The quartile deviation is a slightly better measure of absolute dispersion than the range. But it ignores the observation on the tails. If we take difference samples from a population and calculate their quartile deviations, their values are quite likely to be sufficiently different. This is called sampling fluctuation. It is not a popular measure of dispersion. The quartile deviation calculated from the sample data does not help us to draw any conclusion (inference) about the quartile deviation in the population.

Notes

6.6 Review Questions

1. Explain with example the term 'Dispersion'. Give its definition and discuss the characteristics of dispersion.
2. What are the uses of 'range' as a method of measuring dispersion ? Give its advantages and disadvantages.
3. Give the merits and demerits of quartile deviation. What is the third movement of dispersion?
4. Define Percentile. Discuss the application of Percentile.

Answers: Self-Assessment

1. (*i*) F (*ii*) F (*iii*) T (*iv*) F

6.7 Further Readings



Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Unit 7: Mean Deviation and Standard Deviation

Notes

CONTENTS

Objectives
Introduction
7.1 The Mean Deviation
7.2 The Standard Deviation
7.3 Summary
7.4 Key-Words
7.5 Review Questions
7.6 Further Readings

Objectives

After reading this unit students will be able to:

- Describe the Mean Deviation.
- Explain the Standard Deviation.

Introduction

The two methods of dispersion discussed earlier in this book, namely, range and quartile deviation, are not measures of dispersion in the strict sense of the term because they do not show the scatterness around an average. However, to study the formation of a distribution we should take the deviations from an average. The two other measures, namely, the average deviation and the standard deviation, help us in achieving this goal.

The average deviation is sometimes called the mean deviation. It is the average difference between the items in a distribution and the median or mean of that series. Theoretically, there is an advantage in taking the deviations from median because *the sum of the deviations of items from median is minimum when signs are ignored.* However, in practice the arithmetic mean is more frequently used in calculating the value of average deviation and this is the reason why it is more commonly called mean deviation. In any case, the average used must be clearly stated in a given problem so that any possible confusion in meaning is avoided.

The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.



Did u know? In statistics **standard deviation** (represented by the symbol sigma, σ) shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.

In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be

Notes

conducted multiple times. The reported margin of error is typically about twice the standard deviation—the radius of a 95 percent confidence interval. In science, researchers commonly report the standard deviation of experimental data, and only effects that fall far outside the range of standard deviation are considered statistically significant—normal random error or variation in the measurements is in this way distinguished from causal variation. Standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the volatility of the investment.

7.1 The Mean Deviation

The average deviation or mean deviation is a measure of dispersion that is based upon all the items in a distribution. It is the arithmetic mean of the deviations of the data from its central value, may it be arithmetic mean, median or mode. While, considering the deviations from its central value, only absolute values are taken into consideration, *(i.e.,* without considering the positive or negative signs). Mean deviation is denoted by δ (delta).

Mean/Average Deviation (denoted by δ)

$$\delta_{\bar{X}} = \frac{\Sigma|d\bar{X}|}{N} \text{ (Deviation taken from Arithmetic Mean)}$$

$$\delta_{M} = \frac{\Sigma|dM|}{N} \text{ (Deviation taken from Median)}$$

$$\delta_{Mo} = \frac{\Sigma|dMo|}{N} \text{ (Deviation taken from Mode)}$$

Coefficient of Dispersion:

$$\text{Coeff. of } \delta_{\bar{X}} = \frac{\delta_{\bar{X}}}{\bar{X}} \ (\bar{X} \text{ is arithmetic mean})$$

$$\text{Coeff. of } \delta_{M} = \frac{\delta_{M}}{M} \text{ (M is Median)}$$

$$\text{Coeff. of } \delta_{Mo} = \frac{\delta_{Mo}}{Mo} \text{ (Mo is Mode)}$$

Example 1: A batch of 10 students obtained the following marks out of 100. Calculate the mean deviation and its coefficient.

Marks: 58, 39, 22, 11, 44, 28, 49, 55, 41 and 42.

Solution: The median value for the series is:

$$\left(\frac{N+1}{2}\right)^{th} \text{ item} = \frac{11}{2} = 5.5^{th} \text{ item.}$$

The series in ascending order:

11, 22, 28, 39, 41, 42, 44, 49, 55, 58

$$\therefore \text{ Median} = \frac{41+42}{2} = \frac{83}{2} = 41.5.$$

Calculation of Deviation from Median

Marks	Deviation (d_M)
11	11 – 41.5 = |30.5|
22	22 – 41.5 = |19.5|
28	28 – 41.5 = |13.5|

Notes

39	39 – 41.5 = |2.5|
41	41 – 41.5 = |.5|
42	42 – 41.5 = |.5|
44	44 – 41.5 = |2.5|
49	49 – 41.5 = |7.5|
55	55 – 41.5 = |13.5|
58	58 – 41.5 = |16.5|
	$\sum d_M = 107$

$$\text{Dispersion} = \frac{\sum d_M}{N},\ N = 10,\ \sum d_M = 107.$$

$$\therefore \quad \text{Dispersion} = \frac{107}{10} = 10.7 \text{ marks.}$$

$$\text{Coefficient} = \frac{\delta M}{M} = \frac{10.7}{41.5} = 0.26 \text{ marks.}$$

Answer: Mean deviation (from median) δ_M for the given data is 10.7 marks and the coefficient is 0.26.

Example 2: Calculate mean deviation from arithmetic mean from the following data:
10.500, 10.250, 10.375, 10.625, 10.750, 10.125, 10.375, 10.625, 10.500, 10.125.

Solution: When the data is in fractions and the mean value comes in fractions the following method may be used to avoid tedious calculations.

$$\delta_{\bar{X}} = \frac{1}{N}\left(\bar{X}_y - \bar{X}_x\right)$$

or

$$\delta_M = \frac{1}{N}\left(M_y - M_x\right)$$

where $\bar{X}_y / M_y$ is sum of items above Mean/Median $\bar{X}_x / M_y$ is sum below Mean/ Median.

Mean for the data given = $\frac{104.25}{10} = 10.425$.

The value of items above mean $\left(\bar{X}_y\right)$

10.500 + 10.500 + 10.625 + 10.625 + 10.750 = 53.

The value of items below mean $\left(\bar{X}_x\right)$

10.125 + 10.125 + 10.250 + 10.375 + 10.375 = 51.25.

$$\text{Mean deviation} = \frac{1}{10}(53 - 51.25)$$

$$= \frac{1.75}{10} = 0.175.$$

(This method is also called short-cut method of calculation Mean deviation).

Notes

Answer: The mean deviation from arithmetic mean of the given data is 0.175.

Example 3: Find the mean deviation for the following data (from median and mean)

Items	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	15	16	21	10	17	8	4	2	1	2	2	0	2

Solution: **To find Median**

Items in ascending order	**Frequency**	**Cumulative Frequency**
0	15	15
1	16	31
2	21	52
3	10	62
4	17	79
5	8	87
6	4	91
7	2	93
8	1	94
9	2	96
10	2	98
11	0	98
12	2	100
	$\Sigma f = 100$	

$$\text{Median} = \frac{(N+1)}{2}\text{th item} = 55.5\text{th item}$$

$\therefore$ Median = 2.

Deviation from Median

Items	f	**Deviation** (d_M)	fd_M
0	15	0 – 2 = |2|	2 × 15 = 30
1	16	1 – 2 = |1|	1 × 16 = 16
2	21	2 – 2 = |0|	00
3	10	3 – 2 = |1|	1 × 10 = 10
4	17	4 – 2 = |2|	2 × 17 = 34
5	8	5 – 2 = |3|	3 × 8 = 24
6	4	6 – 2 = |4|	4 × 4 = 16
7	2	7 – 2 = |5|	5 × 2 = 10
8	1	8 – 2 = |6|	6 × 1 = 6
9	2	9 – 2 = |7|	7 × 2 = 14
10	2	10 – 2 = |8|	8 × 2 = 16
11	0	11 – 2 = |9|	9 × 0 = 00
12	2	12 – 2 = |10|	10 × 2 = 20
			$\Sigma\lvert fd_M\rvert = 196$

Notes

Mean deviation from Median $\delta_M = \frac{\Sigma|fd_M|}{N}$

Here, $\Sigma|fd_M| = 196$, N = 100.

$\therefore \quad \delta_M = \frac{196}{100} = 1.96.$

Mean deviation from mean $\delta_{\bar{X}} = \frac{\Sigma|fd_{\bar{X}}|}{N}$

$$\bar{X} = \frac{\Sigma fX}{\Sigma f}$$

fx = 00, 16, 42, 30, 68, 40, 24, 14, 8, 18, 20, 00, 24.

$\Sigma fx = 304$, $\Sigma f = 100$

$\therefore \quad \bar{X} = \frac{304}{100} = 3.04$

Deviation from Mean

Item	f	deviation $d_{\bar{X}}$	$fd_{\bar{X}}$
0	15	|3.04|	45.6
1	16	|2.04|	32.64
2	21	|1.04|	21.84
3	10	|0.04|	0.4
4	17	|0.96|	16.32
5	8	|1.96|	15.68
6	4	|2.96|	11.84
7	2	|3.96|	7.92
8	1	|4.96|	4.96
9	2	|5.96|	11.96
10	2	|6.96|	13.96
11	0	|7.96|	00
12	2	|8.96|	17.92
			$\Sigma fd_{\bar{X}} = 201.04$

Mean deviation $\delta_{\bar{X}} = \frac{\Sigma|fd_{\bar{X}}|}{N}$

$\Sigma fd_{\bar{X}} = 201.04$, N = 100

$\therefore \quad \delta_{\bar{X}} = \frac{201.04}{100} = 2.01$

Answer: Mean deviation from Median δ_M is equal to 1.96 whereas that from arithmetic mean $\delta_{\bar{X}} = 2.01$.

Notes

Example 4: Calculate mean deviation from the following data:

Class-interval	Frequency	Class-interval	Frequency
10–30	6	90–110	21
30–50	53	110–130	26
50–70	85	130–150	4
70–90	56	150–170	4

Given that the median of the above data is 60.5.

Solution:

Class-interval	f	Mid-points	Deviation d_M	fd_M
10–30	6	20	|40.5|	243.0
30–50	53	40	|20.5|	1086.5
50–70	85	60	|0.5|	42.5
70–90	56	80	|17.5|	1092.0
90–110	21	100	|39.5|	2139.5
110–130	26	120	|59.5|	1547.0
130–150	4	140	|79.5|	318.0
150–170	4	160	|99.5|	398.0
				Σfd_M = 6,866.5

$$\delta_M = \frac{\Sigma|fd_M|}{N}, \quad \Sigma|fd_M| = 6{,}866.5,\ N = 255$$

$$\therefore \quad \delta_M = \frac{6{,}866.5}{255} = 26.93.$$

Example 5: Calculate mean deviation from mean from the following data:

X	0–100	100–200	200–300	300–400	400–500	500–600	600–700
f	6	5	8	15	7	6	3

Solution:

X	f	Mid (M) values	Deviation from $\bar{X}$ $(d_{\bar{X}})$	$fd_{\bar{X}}$
0–100	6	50	|284|	1,704
100–200	5	150	|184|	920
200–300	8	250	|84|	672
300–400	15	350	|16|	240
400–500	7	450	|116|	812
500–600	6	550	|216|	1,296
600–700	3	650	|316|	948
	Σf = 50			$\Sigma fx_{\bar{X}}$ = 6,592

Notes

$$\bar{X} = \frac{\Sigma fM}{\Sigma f} = \frac{16,700}{50} = 334$$

$$\delta_{\bar{X}} = \frac{\Sigma \left| fd_{\bar{X}} \right|}{N}$$

$$\Sigma \left| fd_{\bar{X}} \right| = 6,592, \; N = 50$$

$$\delta_{\bar{X}} = \frac{6,592}{50} = 131.84$$

Merits and Limitations of Mean Deviation

Merits

(i) The outstanding advantage of the average deviation is its relative simplicity. It is simple to understand and easy to compute. Anyone familiar with the concept of the average can readily appreciate the meaning of the average deviation. If a situation requires a measure of dispersion that will be presented to the general public or any group not thoroughly grounded in statistics, the average deviation is very useful.

(ii) It is based on each and every item of the data. Consequently change in the value of any item would change the value of mean deviation.

(iii) Mean deviation is less affected by the values of extreme items than the standard deviation.

(iv) Since deviations are taken from a central value, comparison about, formation of different distributions can easily be made.

Limitations

(i) The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items. For example if from twenty, fifty is deducted we write 30 and not – 30. This is mathematically wrong and makes the method **non-algebraic.** If the signs of the deviations are not ignored the net sum of the deviations will be zero if the reference point is the mean or approximately zero if the reference point is median.

(ii) This method may not give us very accurate results. The reason is that mean deviation gives us best results when deviations are taken from median. But median is not a satisfactory measure when the degree of variability in a series is very high. And if we compute mean deviation from mean that is also not desirable because the sum of the deviations from mean (ignoring signs) is greater than the sum of the deviations from median (ignoring signs). If mean deviation is computed from mode that is also not scientific because the value of mode cannot always be determined.

(iii) It is not capable of further algebraic treatment.

(iv) It is rarely used in sociological studies.

Because of these limitations its use is limited and it is overshadowed as a measure of variation by the superior standard deviation.

Usefulness of the Mean Deviation: The serious drawbacks of the average deviation should not blind us to its practical utility. Because of its simplicity in meaning and computation, it is especially effective in reports presented to the general public or to groups not familiar with statistical methods. This measure is useful for small samples with no elaborate analysis required. Incidentally it may be mentioned that the National Bureau of Economic Research has found in its work on forecasting business cycles, that the average deviation is the most practical measure of dispersion to use for this purpose.

Notes

7.2 The Standard Deviation

The standard deviation concept was introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as *root-mean square deviation* for the reason that it is the square root of the means of the squared deviations from the arithmetic mean. Standard deviation is denoted by the small Greek letter σ (read as sigma).

The standard deviation measures the absolute dispersion or variability of a distribution; the greater the amount of dispersion or variability, the greater the standard deviation, the greater will be the magnitude of the deviations of the values from their mean. A small standard deviation means a high degree of uniformity of the observations as well as homogeneity of a series; a large standard deviation means just the opposite. Thus if we have two or more comparable series with identical or nearly identical means, it is the distribution with the smallest standard deviation that has the most representative mean. Hence standard deviation is extremely useful in judging the representativeness of the mean.

Difference between Average Deviation and Standard Deviation

Both these measures of dispersion are based on each and every item of the distribution. But they differ in the following respects:

(i) Algebraic signs are ignored while calculating mean deviation whereas in the calculation of standard deviations, signs are taken into account.

(ii) Mean deviation can be computed either from median or mean. The standard deviation, on the other hand, is always computed from the arithmetic mean because the sum of the squares of the deviations of items from arithmetic mean is the least.

Calculation of Standard Deviation—Individual Observations

In case of individual observations, standard deviation may be computed by applying any of the following two methods:

1. By taking deviations of the items from the actual mean.
2. By taking deviations of the items from an assumed mean.

1. Deviations taken from Actual Mean: When deviations are taken from actual mean the following formula is applied:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

where $x = \left(X - \bar{X}\right)$ and N = number of observations.

Steps : (i) Calculate the actual mean of the series, *i.e.,* $\bar{X}$.

(ii) Take the deviations of the items from the mean, *i.e.,* find $\left(X - \bar{X}\right)$. Denote these deviation by x.

(iii) Square these deviations and obtain the total Σx^2.

(iv) Divide Σx^2 by the total number of observations, *i.e.,* N, and extract the square-root. This gives us the value of standard deviation.

Example 6: Calculate standard deviation from the following observations of marks of 5 students of a tutorial group:

Marks out of 25 **Notes**

8 12 13 15 22

Solution:

CALCULATION OF STANDARD DEVIATION

X	$(X - \bar{X})$	x^2
8	– 6	36
12	– 2	4
13	– 1	1
15	+ 1	1
22	+ 8	64
$\Sigma X = 70$	$\Sigma x = 0$	$\Sigma x^2 = 106$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \text{ where } x = (X - \bar{X})$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{70}{5} = 14$$

$$\Sigma x^2 = 106, N = 5$$

$$\sigma = \sqrt{\frac{106}{5}} = \sqrt{21.2} = 4.604.$$

2. Deviations taken from Assumed Mean: When the actual mean is in fractions, say, in the above case 123.674, it would be too cumbersome to take deviations from it and then obtaining squares of these deviations. In such a case, either the mean may be approximated or else the deviations be taken from an assumed mean and the necessary adjustment be made in the value of standard deviation. The former method of approximation is less accurate and, therefore, invariably in such a case deviations are taken from assumed mean.

When deviations are taken from assumed mean the following formula is applied:

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

Steps : (i) Take the deviations of the items from an assumed mean *i.e.*, obtain (X –A). Denote these deviations by d. Take the total of these deviations, *i.e.*, obtain Σd.

(ii) Square these deviations and obtain the total Σd^2.

(iii) Substitute the value of Σd^2, Σd and N in the formula.

Example 7: Following figures give the income of 10 persons in rupees. Find the standard deviation.

227, 235, 255, 269, 292, 299, 312, 321, 333, 348

Notes

Solution:

CALCULATION OF STANDARD DEVIATION

X	(X–280)d	d^2
227	– 53	2809
235	– 45	2025
255	– 25	625
269	– 11	121
292	+ 12	144
299	+ 19	361
312	+ 32	1024
321	+ 41	1681
333	+ 53	2809
348	+ 68	4624
N = 10	$\sum d$ = 91	$\sum d^2$ = 16223

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$$

$\sum d^2$ = 16223, N = 10, $\sum d$ = 91

$$\sigma = \sqrt{\frac{16223}{10} - \left(\frac{91}{10}\right)^2} = \sqrt{1622.3 - 82.81} = 39.24.$$

Calculation of Standard Deviation—Discrete Series

For calculating standard deviation in discrete series any of the following methods may be applied:

1. Actual mean method.
2. Assumed mean method.
3. Step deviation method.

1. Actual Mean Method: When this method is applied deviations are taken from the actual mean, *i.e.*, we find $\left(X - \bar{X}\right)$ and denote these deviations by x. These deviations are then squared and multiplied by the respective frequencies. The following formula is applied:

$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

where $x = \left(X - \bar{X}\right)$.

However, in practice this method is rarely used because if the actual mean is in fractions the calculations take a lot of time.

2. Assumed Mean Method: When this method is used, the following formula is applied.

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Notes

where $d = (X - A)$.

Steps : (i) Take the deviations of the items from an assumed mean and denote these deviations by d.

(ii) Multiply the deviations by the respective frequencies and obtain the total, Σfd.

(iii) Obtain the squares of the deviations, *i.e.*, calculate d^2.

(iv) Multiply the squared deviations by respective frequencies and obtain the total, Σfd^2.

Substitute the values in the above formula.

Example 8: Calculate the standard deviation from the data given below:

Size of item	Frequency	Size of item	Frequency
3.5	3	7.5	85
4.5	7	8.5	32
5.5	22	9.5	8
6.5	60		

Solution:

CALCULATION OF STANDARD DEVIATION

X Size of item	f	$(X-6.5)d$	fd	fd^2
3.5	3	– 3	– 9	27
4.5	7	– 2	– 14	28
5.5	22	– 1	– 22	22
6.5	60	0	0	0
7.5	85	+ 1	+ 85	85
8.5	32	+ 2	+ 64	128
9.5	8	+ 3	+ 24	72
	N = 217		Σfd = 128	Σfd^2 = 362

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

$\Sigma fd^2 = 362$, $\Sigma fd = 128$, N = 217

$$\sigma = \sqrt{\frac{362}{217} - \left(\frac{128}{217}\right)^2} = \sqrt{1.67 - (.59)^2}$$

$$= \sqrt{1.67 - 0.35} = 1.149.$$

3. Step Deviation Method: When this method is used we take a common factor from the given data. The formula for computing standard deviation is:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

Notes

where $d = \frac{(X - A)}{i}$ and i = class interval

The use of the above formula simplifies calculations.

Example 9: Find the standard deviation for the following distribution:

X	4.5	14.5	24.5	34.5	44.5	54.5	64.5
f	1	5	12	22	17	9	4

Solution:

Calculation of Standard Deviation

X	f	(X–34.5)/10 d	fd	fd^2
4.5	1	– 3	– 3	9
14.5	5	– 2	– 10	20
24.5	12	– 1	– 12	12
34.5	22	0	0	0
44.5	17	+ 1	+ 17	17
54.5	9	+ 2	+ 18	36
64.5	4	+ 3	+ 12	36
	N = 70		Σfd = 22	Σfd^2 = 130

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

Here Σfd^2 = 130, Σfd = 22, N = 70, i = 10.

$$\sigma = \sqrt{\frac{130}{70} - \left(\frac{22}{70}\right)^2} \times 10 = \sqrt{1.857 - .1} \times 10 = 1.326 \times 10$$

= 13.26.

Calculation of Standard Deviation—Continuous Series

In continuous series any of the methods discussed above for discrete frequency distribution can be used. However, in practice it is the step deviation method that is mostly used. The formula is:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$d = \frac{(m - A)}{i}$; i = Class interval.

Steps : (i) Find the mid-points of various classes.

(ii) Take the deviations of these mid-points from an assumed mean and divide by the class interval. Denote these deviations by d.

Notes

(iii) Multiply the frequency of each class with these deviations and obtain Σfd.

(iv) Square the deviations and multiply them with the respective frequencies of each class and obtain Σfd^2.

Thus, the only difference in procedure in case of continuous series is to find mid-points of the various classes.

Example 10: The following table gives the distribution of income of 100 families in a village. Calculate standard deviation:

Income (Rs.)	No. of families
0–1000	18
1000–2000	26
2000–3000	30
3000–4000	12
4000–5000	10
5000–6000	4

(B.Com., Kerala Univ., 1996)

Solution:

Calcultion of Standard of Diviation

Income (Rs.)	*m.p.* *m*	*f*	(*m*–2500)/1000 *d*	*fd*	fd^2
0–1000	500	18	– 2	– 36	72
1000–2000	1500	26	– 1	– 26	26
2000–3000	2500	30	0	0	0
3000–4000	3500	12	+ 1	+ 12	12
4000–5000	4500	10	+ 2	+ 20	40
5000–6000	5500	4	+ 3	+ 12	36
		N = 100		Σfd = – 18	Σfd^2 = 186

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{186}{100} - \left(\frac{-18}{100}\right)^2} \times 1000 = \sqrt{1.86 - .0324} \times 1000$$

$$= 1.3519 \times 1000 = 1351.9.$$

Mathematical Properties of Standard Deviation

Standard deviation has some very important mathematical properties which considerably enhance its utility in statistical work.

1. **Combined standard deviation:** Just as it is possible to compute combined mean of two or more than two groups, similarly we can also compute combined standard deviation of two or more groups. Combined standard deviation is denoted by σ_{12} and is computed as follows:

Notes

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2\sigma_2^2}{N_1 + N_2}}$$

where σ_{12} = combined standard deviation; σ_1 = standard deviation of first group; σ_2 = standard deviation of second group; $d_1 = \left(\overline{X}_1 - \overline{X}_{12}\right)$; $d_2 = \left(\overline{X}_2 - \overline{X}_{12}\right)$.

The above formula can be extended to find out the standard deviation of three or more groups. For example, combined standard deviation of three groups would be:

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$$d_1 = \left|\overline{X}_1 - \overline{X}_{123}\right|,\ d_2 = \left|\overline{X}_2 - \overline{X}_{123}\right|,\ d_3 = \left|\overline{X}_3 - \overline{X}_{123}\right|.$$

Example 11: The number examined, the mean weight and the standard deviation in each group of examination by two medical examiners is given below. Find the mean weight and standard deviation of both the groups taken together.

A	50	113	6.5
B	60	120	8.2

Solution:

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

$$N_1 = 50,\ N_2 = 60,\ \overline{X}_1 = 113,\ \overline{X}_2 = 120$$

$$\overline{X}_{12} = \frac{(50 \times 113) + (60 \times 120)}{50 + 60} = \frac{5650 + 7200}{110} = \frac{12850}{110} = 116.82$$

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$$N_1 = 50,\ \sigma_1 = 6.5,\ N_3 = 60,\ \sigma_2 = 8.2$$

$$d_1 = \left|\overline{X}_1 - \overline{X}_{12}\right| = (113 - 116.82) = -3.82$$

$$d_2 = \left|\overline{X}_2 - \overline{X}_{12}\right| = (120 - 116.82) = 3.18.$$

Substituting the values

$$\sigma_{12} = \sqrt{\frac{50(6.5)^2 + 60(8.2)^2 + 50(-3.82)^2 + 60(3.18)^2}{50 + 60}}$$

$$= \sqrt{\frac{2112.5 + 4034.4 + 729.62 + 606.744}{110}} = \sqrt{\frac{7483.264}{100}}$$

$$= \sqrt{68.03} = 8.25.$$

Example 12: The number of workers employed, the mean wage (in Rs.) per month and the standard deviation (in Rs.) in each section of a factory are given below. Calculate the mean

wage and standard deviation of all the workers taken together.

Notes

Section	No. of workers employed	Mean wage in Rs.	Standard deviation in Rs.
A	50	113	6
B	60	120	7
C	90	115	8

Solution:

$$\bar{X}_{123} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}$$

$$= \frac{(50\times 113) + (60\times 120) + (90\times 115)}{60 + 60 + 90}$$

$$= \frac{5650 + 7200 + 10350}{200} = \frac{23200}{200} \text{ Rs. } 116.$$

Combined standard deviation of three series.

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1d_1^2 + N_2d_2^2 + N_3d_3^2}{N_1 + N_2 + N_3}}$$

$$d_1 = \left|\bar{X}_1 - \bar{X}_{123}\right| = \left|113 - 116\right| = 3$$

$$d_2 = \left|\bar{X}_2 - \bar{X}_{123}\right| = \left|120 - 116\right| = 4$$

$$d_3 = \left|\bar{X}_3 - \bar{X}_{123}\right| = \left|115 - 116\right| = 1$$

$$\sigma_{123} = \sqrt{\frac{50(6)^2 + 60(7)^2 + 90(8)^2 + 50(3)^2 + 60(4)^2 + 90(-1)^2}{50 + 60 + 90}}$$

$$= \sqrt{\frac{1800 + 2940 + 5760 + 450 + 960 + 90}{200}} = \sqrt{\frac{12000}{200}} = \sqrt{60} = 7.75.$$

2. **Standard deviation of *n* natural numbers:** The standard deviation of the first *n* natural numbers can be obtained by the following formula:

$$\sigma = \sqrt{\frac{1}{12}\left(N^2 - 1\right)}$$

Thus, the standard deviation of natural numbers 1 to 10 will be

$$\sigma = \sqrt{\frac{1}{12}\left(10^2 - 1\right)} = \sqrt{\frac{1}{12}\times 99} = \sqrt{8.25} = 2.872.$$

Note: The answer would be the same when direct method of calculating standard deviation is used. But this holds good only for natural numbers from 1 to *n* in continuation without gaps.

3. The sum of the squares of the deviations of items in the series from their arithmetic mean is minimum. In other words, the sum of the squares of the deviations of items of any series from a value other than the arithmetic mean would always be greater. This is the reason why standard deviation is always computed from the arithmetic mean.

Notes

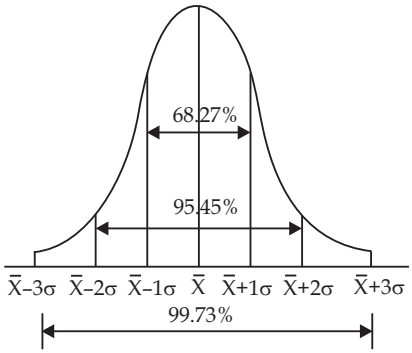
4. For a symmetrical distribution, the following area relationship holds good:

Mean $\pm 1\sigma$ covers 68.27% items.

Mean $\pm 2\sigma$ covers 95.45% items.

Mean $\pm 3\sigma$ covers 99.73% items.

This can be illustrated by the following diagram:



Relation between Measures of Dispersion

In a normal distribution there is a fixed relationship between the three most commonly used measures of dispersion. The quartile deviation is smallest, the mean deviation next and the standard deviation is largest, in the following proportion:

$$\text{Q.D.} = \frac{2}{3}\sigma \text{ ; M.D.} = \frac{4}{5}\sigma$$

These relationships can be easily memorised because of the sequence 2, 3, 4, 5. The same proportions tend to hold true for many distributions that are quite normal. They are useful in estimating one measure of dispersion when another is known, or in checking roughly the accuracy of a calculated value. If the computed σ differs very widely from its value estimated from Q.D. or M.D. either an error has been made or the distribution differs considerably from normal.

Another comparison may be made of the proportion of items that are typically included within the range of one Q.D., M.D. or S.D. measured both above and below the mean. In a normal distribution:

$\bar{X} \pm \text{Q.D.}$ includes 50 per cent of the items.

$\bar{X} \pm \text{M.D.}$ includes 57.51 per cent of the items.

$\bar{X} \pm \sigma$ includes 68.27 per cent of the items.

Coefficient of Variation

The Standard deviation discussed above is an absolute measure of dispersion. The corresponding relative measure is known as the *coefficient of variation.* This measure developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problems where we want to compare the variability of two or more than two series. That series (or group) for which the coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous. On the other hand, the series for which coefficient of variation is less is

Notes

said to be less variable or more consistent, more uniform more stable or more homogeneous. Coefficient of variation is denoted by the symbol C.V. and is obtained as follows:

Coefficient of variation or C.V. = $\frac{\sigma}{\bar{X}} \times 100$.

It may be pointed out that although any measure of dispersion can be used in conjunction with any average in computing relative dispersion, statisticians, in fact, almost always use the standard deviation as the measure of dispersion and the arithmetic mean as the average. When the relative dispersion is stated in terms of the arithmetic mean and the standard deviation, the resulting percentage is known as the coefficient of variation of coefficient of variability.

A distinction is sometimes made between coefficient of variation and coefficient of standard deviation.

The former is always a percentage, the latter is just the ratio of standard deviation to mean, *i.e.*, $\left(\frac{\sigma}{\bar{X}}\right)$.

Example 13: The scores of two batsmen A and B in ten innings during a certain season are:

A	32	28	47	63	71	39	10	60	96	14
B	19	31	48	53	67	90	10	62	40	80

Find (using coefficient of variation) which of the batsmen A, B is more consistent in scoring. **(B.Com., Calcutta Univ., 1996)**

Solution:

Calculation of Coefficient of Variation

X	**(X – 46)** x	x^2	**Y**	$(Y - \bar{Y})$ y	y^2
32	– 14	196	19	– 31	961
28	– 18	324	31	– 19	361
47	+ 1	1	48	– 2	4
63	+ 17	289	53	+ 3	9
71	+ 25	625	67	+ 17	289
39	– 7	49	90	+ 40	1600
10	– 36	1296	10	– 40	1600
60	+ 14	196	62	+ 12	144
96	+ 50	2500	40	– 10	100
14	– 32	1024	80	+ 30	900
$\Sigma X = 460$	$\Sigma x = 0$	$\Sigma x^2 = 6500$	$\Sigma Y = 500$	$\Sigma y = 0$	$\Sigma y^2 = 5968$

Batsman A

C.V. = $\frac{\sigma}{\bar{X}} \times 100$

$\bar{X} = \frac{460}{10} = 46$

$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{6500}{10}} = 25.5$

Batsman B

C.V. = $\frac{\sigma}{\bar{Y}} \times 100$

$\bar{Y} = \frac{500}{10} = 50$

$\sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{5968}{10}} = 24.43$

Notes

$$\text{C.V.} = \frac{25.5}{46} \times 100 = 55.43 \qquad \text{C.V.} = \frac{24.43}{50} \times 100 = 48.86$$

Since coefficient of variation is less for batsman B hencebatsman B is more consistent.

Example 14: A panel of two judges P and O graded seven dramatic performances by independently awarding marks as follows:

Performance:	1	2	3	4	5	6	7
Marks by P:	46	42	44	40	43	41	45
Marks by O:	40	38	36	35	39	37	41

Find out coefficient of variation in the marks awarded by two judges and interpret the result.

Solution:

Coefficient of Variation of the Marks Obtained by P and O

Marks by P **X**	$(\mathbf{X} - \bar{\mathbf{X}})$ x	x^2	**Marks by O** **Y**	$(\mathbf{Y} - \bar{Y})$ y	y^2
46	+ 3	9	40	+ 2	4
42	–1	1	38	0	0
44	+ 1	1	36	– 2	4
40	– 3	9	35	– 3	9
43	0	0	39	+ 1	1
41	– 2	4	37	– 1	1
45	+ 2	4	41	+ 3	9
$\Sigma X = 301$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma Y = 266$	$\Sigma y = 0$	$\Sigma y^2 = 28$

Marks by P

$$\bar{X} = \frac{\Sigma X}{N} = \frac{301}{7} = 43$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{28}{7}} = 2$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{2}{43} \times 100 = 4.65$$

Marks by O

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{266}{7} = 38$$

$$\sigma = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{28}{7}} = 2$$

$$\text{C.V.} = \frac{\sigma}{\bar{Y}} \times 100 = \frac{2}{38} \times 100 = 5.26.$$

Notes

The average marks obtained by P are higher. Hence his performance is better. The coefficient of variation is lower in case of P hence he is a more consistent student.

Example 15: Suppose that samples of polythene bags two manufactures, A and B are tested by prospective buyer for bursting pressure, with the following results:

Bursting Pressure (*lb.*)	Number of bags	
	A	B
5.0–9.9	2	9
10.0–14.9	9	11
15.0–19.9	29	18
20.9–24.9	54	32
25.0–29.9	11	27
30.0–34.9	5	13
	110	110

Which set of the bags has the highest average bursting pressure ? Which has more uniform pressure ? If prices are the same, which manufacture's bags would be preferred by the buyer ? Why ?

Solution: For determining the set of bags having average bursting pressure, calculate arithmetic mean and for finding out set of bags having more uniform pressure compute coefficient of variation.

Manufacturer A

Calculation of Mean and Standard Deviation

Bursting pressure (*lb.*)	m	f	$\left(\frac{m-17.45}{5}\right)$ d	fd	fd^2
4.95–9.95	7.45	2	– 2	– 4	8
9.95–14.95	12.45	9	– 1	– 9	9
14.95–19.95	**17.45**	29	0	0	0
19.95–24.95	22.45	54	+ 1	+ 54	54
24.95–29.95	27.45	11	+ 2	+ 22	44
29.95–34.95	32.45	5	+ 3	+ 15	45
	N = 110			Σfd = 78	Σfd^2 = 160

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i$$

Here $A = 17.45$, $\Sigma fd = 78$, $N = 110$, $i = 5$

$$\bar{X} = 17.45 + \frac{78}{110} \times 5 = 17.45 + 3.55 = 21.$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{160}{110} - \left(\frac{78}{110}\right)^2} \times 5$$

Notes

$$= \sqrt{1.455 - 0.503} \times 5 = \sqrt{0.952} \times 5 = 0.976 \times 5 = 4.88$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{4.88}{21} \times 100 = 23.24\%.$$

Manufacturer B

Calculation of Mean and Standard Deviation

Bursting pressure (*lb.*)	*m*	*f*	$\left(\frac{m - 17.45}{5}\right)$ *d*	*fd*	fd^2
4.95–9.95	7.45	9	– 2	–18	36
9.95–14.95	12.45	11	– 1	– 11	11
14.95–19.95	**17.45**	18	0	0	0
19.95–24.95	22.45	54	+ 1	+ 54	54
24.95-29.95	27.45	27	+ 2	+ 54	108
29.95–34.95	32.45	13	+ 3	+ 39	117
		N = 110		Σfd = 96	Σfd^2 = 304

$$\bar{X} = A + \frac{\Sigma fd}{N} \times i = 17.45 + \frac{96}{110} \times 5 = 17.45 + 4.36 = 21.18$$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \times i = \sqrt{\frac{304}{110} - \left(\frac{96}{110}\right)^2} \times 5$$

$$= \sqrt{2.764 - 0.762} \times 5 = 1.4149 \times 5 = 7.075$$

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100 = \frac{7.075}{21.81} \times 100 = 32.44\%$$

Since the average bursting pressure is higher for manufacturer B, the bags of manufacturer B have a higher bursting pressure. The bags of manufacturer A have more uniform pressure since the coefficient of variation is less for manufacturer A. If princes are the same, the bags of manufacturer A should be preferred by the buyer because they have more uniform pressure.

Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1918. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in the original series. Variance is defined as follows:

$$\text{Variance} = \frac{\Sigma\left(X - \bar{X}\right)^2}{N}.$$

Thus, variance is nothing but the square of the standard deviation, *i.e.,*

$$\text{Variance} = \sigma^2$$

or $$\sigma = \sqrt{\text{Variance.}}$$

Notes

In a frequency distribution where deviations are taken from assumed mean, variance may directly be computed as follows:

$$\text{Variance} = \left\{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2\right\} \times i$$

where $= \frac{(X - A)}{i}$ and i = class interval.

Example 16: The weights of a number of packages are given as follows:

16.1, 15.9, 15.8, 16.3, 16.2, 16.0, 15.9, 16.0, 16.1, 16.0, 15.9, 16.1, 16.0, 16.0.

From a frequency table. Find the standard deviation and the variance.

Solution:

Weight X	Tally Bar	Frequency f	$(X - A)$ d	fd	fd^2
15.8	|	1	– .3	– .3	.09
15.9	|||	3	– .2	– .6	.12
16.0	||||	5	– .1	– .5	.05
16.1	|||	3	0	0	0
16.2	|	1	+ .1	+ .1	01
16.3	|	1	+ .2	+ .2	04
		N = 14		Σfd = – 1.1	Σfd^2 = 0.31

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

$\Sigma fd^2 = 0.31$, $\Sigma fd = -1.1$, $N = 14$

$$= \sqrt{\frac{.31}{14} - \frac{(-1.1)^2}{14}} = \sqrt{0.022 - .0062} = 0.126$$

Variance $= \sigma^2 = (.126)^2 = 0.0159$.

Merits and Limitations of Standard Deviation

Merits

(i) The standard deviation is the best measure of variation because of its mathematical characteristics. It is based on every item of the distribution. Also it is amenable to algebraic treatment and is less affected by fluctuations of samlping than most other measures of dispersion.

(ii) It is possible to calculate the combined standard deviation of two or more groups. This is not possible with any other measure.

(iii) For comparing the variability of two or more distributions coefficient of variation is considered to be most appropriate and this is based on mean and standard deviation.

(iv) Standard deviation is most prominently used in further statistical work. For example, in computing skewness, correlation, etc., use is made of standard deviation. It is a key note in sampling and provides a unit of measurement for the normal distribution.

Notes

Limitations

(i) As compared to other measures it is difficult to compute. However, it does not reduce the importance of this measure because of high degree of accuracy of results it gives.

(ii) It gives more weight to extreme items and less to those which are near the mean. It is because of the fact that the squares of the deviations which are big in size would be proportionately greater than the squares of those deviations which are comparatively small. The deviations 2 and 8 are in the ratio of 1: 4 but their sqaures, *i.e.*, 4 and 64, would be in the ratio of 1: 16.

Correcting Incorrect Values of Standard Deviation

Mistakes in calculations are always possible. Sometimes it so happens that while calculating mean and standard deviation we unconsciously copy out wrong items. For example, an item 21 may be copied as 12. Similarly one item 127 may be taken as only 27. In such cases if the entire calculations are done again, it would become too tedious a task. By adopting a very simple procedure we can correct the incorrect values of mean and standard deviation. For obtaining correct mean we find out correct ΣX by deducting from the original ΣX the wrong items and adding to it the correct items.

Similarly for calculating correct standard deviation we obtain the value of correct ΣX^2. The following illustration shall clarify the calculations.

Example 17: A student obtained the mean and standard deviation of 100 observations as 40 and 5.1 respectively. It was later found that one observation was wrongly copied as 50, the correct figure being 40. Find the correct mean and standard deviation.

Solution:

Correct Mean:

We are given $\overline{X} = 40$, $\sigma = 5.1$, N = 100

$$\overline{X} = \frac{\Sigma X}{N}$$

$$40 = \frac{\Sigma X}{100} \text{ or } \Sigma X = 4000$$

But correct ΣX = ΣX – Wrong items + Correct items = 4000 – 50 + 40 = 3990.

$$\text{Correct } \overline{X} = \frac{\text{Correct } \Sigma X}{N} = \frac{3990}{100} = 39.9.$$

Correct Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma X^2}{N} - \left(\overline{X}\right)^2}$$

$$5.1 = \sqrt{\frac{\Sigma X^2}{100} - (40)^2}$$

Squaring, we get

$$26.01 = \frac{\Sigma X^2}{100} - 1600$$

$$2601 = \Sigma X^2 - 1{,}60{,}000 \text{ or } \Sigma X^2 = 2601 + 1, 60, 000 = 1, 62, 601.$$

Correct ΣX^2 = Incorrect ΣX^2 – wrong item square + Correct item square

Notes

$$\text{Correct } \Sigma X^2 = 162601 - (50)^2 + (40)^2 = 162601 - 2500 + 1600 = 161701$$

$$\text{Correct } \sigma = \sqrt{\frac{\text{Correct } \Sigma X^2}{N} - \left(\text{Correct } \bar{X}\right)^2}$$

$$= \sqrt{\frac{161701}{100} - (39.9)^2} = \sqrt{1617.01 - 1592.01} = \sqrt{25} = 5.$$

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) Mean deviation can be calculated from arithmetic mean, median or mode.

(*ii*) Mean deviation ignores the sings of deviations.

(*iii*) Standard deviation is an absolute measure of dispersion.

(*iv*) Standard deviations of more than two component parts cannot be combined in one.

(*v*) Mean deviation is least when deviations are taken from median.

7.3 Summary

- The average deviation is sometimes called the mean deviation. It is the average difference between the items in a distribution and the median or mean of that series. Theoretically, there is an advantage in taking the deviations from median because *the sum of the deviations of items from median is minimum when signs are ignored*. However, in practice the arithmetic mean is more frequently used in calculating the value of average deviation and this is the reason why it is more commonly called mean deviation. In any case, the average used must be clearly stated in a given problem so that any possible confusion in meaning is avoided.
- In statistics **standard deviation** (represented by the symbol sigma, σ) shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values.
- The standard deviation of a random variable, statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data.
- The average deviation or mean deviation is a measure of dispersion that is based upon all the items in a distribution. It is the arithmetic mean of the deviations of the data from its central value, may it be arithmetic mean, median or mode. While, considering the deviations from its central value, only absolute values are taken into consideration, *(i.e.,* without considering the positive or negative signs). Mean deviation is denoted by δ (delta).
- The outstanding advantage of the average deviation is its relative simplicity. It is simple to understand and easy to compute. Anyone familiar with the concept of the average can readily appreciate the meaning of the average deviation. If a situation requires a measure of dispersion that will be presented to the general public or any group not thoroughly grounded in statistics, the average deviation is very useful.
- The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items. For example if from twenty, fifty is deducted we write 30 and not – 30. This is mathematically wrong and makes the method **non-algebraic.** If the signs of the deviations are not ignored the net sum of the deviations will be zero if the reference point is the mean or approximately zero if the reference point is median.
- The serious drawbacks of the average deviation should not blind us to its practical utility. Because of its simplicity in meaning and computation, it is especially effective in reports

Notes

presented to the general public or to groups not familiar with statistical methods. This measure is useful for small samples with no elaborate analysis required. Incidentally it may be mentioned that the National Bureau of Economic Research has found in its work on forecasting business cycles, that the average deviation is the most practical measure of dispersion to use for this purpose.

- The standard deviation concept was introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of studying dispersion. Its significance lies in the fact that it is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion. Standard deviation is also known as *root-mean square deviation* for the reason that it is the square root of the means of the squared deviations from the arithmetic mean. Standard deviation is denoted by the small Greek letter σ (read as sigma).
- The standard deviation measures the absolute dispersion or variability of a distribution; the greater the amount of dispersion or variability, the greater the standard deviation, the greater will be the magnitude of the deviations of the values from their mean. A small standard deviation means a high degree of uniformity of the observations as well as homogeneity of a series; a large standard deviation means just the opposite. Thus if we have two or more comparable series with identical or nearly identical means, it is the distribution with the smallest standard deviation that has the most representative mean. Hence standard deviation is extremely useful in judging the representativeness of the mean.
- Mean deviation can be computed either from median or mean. The standard deviation, on the other hand, is always computed from the arithmetic mean because the sum of the squares of the deviations of items from arithmetic mean is the least.
- When the actual mean is in fractions, say, in the above case 123.674, it would be too cumbersome to take deviations from it and then obtaining squares of these deviations. In such a case, either the mean may be approximated or else the deviations be taken from an assumed mean and the necessary adjustment be made in the value of standard deviation.
- The sum of the squares of the deviations of items in the series from their arithmetic mean is minimum. In other words, the sum of the squares of the deviations of items of any series from a value other than the arithmetic mean would always be greater. This is the reason why standard deviation is always computed from the arithmetic mean.
- In a normal distribution there is a fixed relationship between the three most commonly used measures of dispersion. The quartile deviation is smallest, the mean deviation next and the standard deviation is largest, in the following proportion:

$$\text{Q.D.} = \frac{2}{3}\sigma \text{ ; M.D.} = \frac{4}{5}\sigma$$

- These relationships can be easily memorised because of the sequence 2, 3, 4, 5. The same proportions tend to hold true for many distributions that are quite normal. They are useful in estimating one measure of dispersion when another is known, or in checking roughly the accuracy of a calculated value. If the computed σ differs very widely from its value estimated from Q.D. or M.D. either an error has been made or the distribution differs considerably from normal.
- The Standard deviation discussed above is an absolute measure of dispersion. The corresponding relative measure is known as the *coefficient of variation.* This measure developed by Karl Pearson is the most commonly used measure of relative variation. It is used in such problems where we want to compare the variability of two or more than two series. That series (or group) for which the coefficient of variation is greater is said to be more variable or conversely less consistent, less uniform, less stable or less homogeneous.
- The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1918. The concept of variance is highly important in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variation in the original series.

Notes

- The standard deviation is the best measure of variation because of its mathematical characteristics. It is based on every item of the distribution. Also it is amenable to algebraic treatment and is less affected by fluctuations of samlping than most other measures of dispersion.
- Mistakes in calculations are always possible. Sometimes it so happens that while calculating mean and standard deviation we unconsciously copy out wrong items. For example, an item 21 may be copied as 12. Similarly one item 127 may be taken as only 27. In such cases if the entire calculations are done again, it would become too tedious a task. By adopting a very simple procedure we can correct the incorrect values of mean and standard deviation. For obtaining correct mean we find out correct ΣX by deducting from the original ΣX the wrong items and adding to it the correct items.

7.4 Key-Words

1. Mean Deviation : The mean deviation or the average deviation is defined as the mean of the absolute deviations of observations from some suitable average which may be the arithmetic mean, the median or the mode. The difference ( ) is called deviation and when we ignore the negative sign, this deviation is written as and is read as mod deviations.
2. Standard Deviation : In statistics and probability theory, standard deviation (represented by the symbol sigma, ?) shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean; high standard deviation indicates that the data points are spread out over a large range of values

7.5 Review Questions

1. Discuss 'mean deviation' method of measuring dispersion giving its merits and demerits.
2. Explain the concept of Standard deviation. What are its merits and demerits ?
3. Discuss the mathematical Properties of Standard deviation.
4. Distinguish between mean deviation and Standard deviation.

Answers: Self-Assessment

1. (*i*) T (*ii*) T (*iii*) T (*iv*) F (*v*) T

7.6 Further Readings



1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 8: Skewness and Kurtosis: Karl Pearson, Bowley, Kelly's Methods

CONTENTS

Objectives
Introduction
8.1 Meaning, Definition and Types of Skewness
8.2 Karl Pearson, Bowley and Kelly's Methods
8.3 Kurtosis
8.4 Summary
8.5 Key-Words
8.6 Review Questions
8.7 Further Readings

Objectives

After reading this unit students will be able to:

- Describe the Meaning, Definition and Types of Skewness.
- Know the Measures of Skewness.
- Explain Karl Pearson, Bowley and Kelly's Methods.
- Understand Kurtosis.

Introduction

Measuring of central tendencies reveal the concentration of frequencies towards the central value of the series and methods of dispersion reveal the dispersal of values in relation to the central value. But the nature of dispersal of values on either sides of an average is not known by measuring dispersion. Similarly, Kurtosis is yet another measure which tells us about the form of a distribution. Thus, it can be said that the central tendencies and dispersion measures should be supplemented by measures of skewness and kurtosis so that a more elaborate picture about the distribution given can be obtained. The study becomes more important in subjects of economics, sociology and other social sciences where normal distribution in a series usually does not occur. However, studies hold importance in biological sciences and other physical sciences as well.

8.1 Meaning, Definition and Types of Skewness

Skewness—Meaning and Definition

The word 'skewness' is the opposite of symmetry and its presence tells us that a particular distribution is not symmetrical or in other words it is skewed. The word 'skewness' can be understood by the following definitions given by eminent statisticians, economists and mathematicians.

(1) As per *Croxten* and *Cowden*, "When a series is not symmetrical it is skewed."

(2) In the words of *Simpson* and *Kafka*, "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the arithmetic mean, median and mode are identicle. The more the mean moves away from mode, the larger the asymmetry or skewness."

Notes

(3) According to *Garrett*, "A distribution is said to be skewed when the mean and median fall at different points in the distribution and the balance or centre of gravity is shifted to one side or the other.

(4) *Riggleman* and *Frisbee* have defined skewness as, "Skewness is the lack of symmetry. When a frequency distribution is plotted on a chart, skewness present in items tends to the disperse chart more on one side of the mean than on other."

Did u know? The measures of 'Skewness' tell about the pattern of dispersal of items from an average, whether it is symmetrical or not. The nature of distribution is further studied deeply by calculating 'Moments' which reveals whether the symmetrical curve is normal, more flat than normal or more peaked than normal.

From the above discussion it is clear that skewness is the lack of symmetry. Measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with symmetrical or normal distribution. In a symmetrical distribution, frequencies go on increasing upto a point and then begin to decrease in the same fashion. There are various possible patterns of symmetrical distribution and normal distribution which is bell-shaped is one of these. Some of the possible patterns of the symmetrical distribution are:

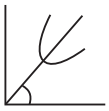


Figure 1

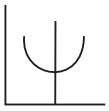


Figure 2

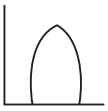


Figure 3



Figure 4

Symmetrical but not bell shaped



Figure 5

Normal distribution

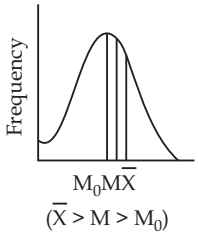
Figure 5: (Symmetrical bell-shaped distribution)

In a symmetrical distribution, mean = median = mode and they lie at the centre of the distribution. When symmetry is disturbed, these values are pulled apart.

Types of Skewness

The skewness may be broadly of two types:

(a) Positive skewness: A distribution in which more than half of the area under the curve is to the right side of the mode, it is said to be a positively skewed distribution. In this type of skewness the right tail is longer than the left tail. In this case, mean is greater than median and the median is greater than the mode and $Q_3 - M > M - Q_1$. Diagrammatically,



($\bar{X} > M > M_0$)

Figure: 6

Notes

(b) Negative skewness: A distribution in which more than half of the area under the distribution curve is to the left side of the mode, it is said to be a negatively skewed distribution. In this case, the elongated tail is to the left and mean is less than the median which is less than mode and $Q_3 - M < M - Q_1$.

Diagrammatically, the negative skewness can be explained as below:

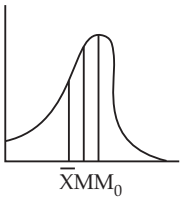


Figure: 7

8.2 Karl Pearson's, Bowley and Kelly's Methods

The following are the main methods of measuring *Skewness* of data:

(1) Karl Pearson's Method

(2) Bowley's Method

(3) Kelly's Method

(1) Karl Pearson's Method

The method of skewness given by Karl Pearson is also called as First Measure of Skewness. This method is based on the difference between the 'mean' and 'mode'. Thus,

$S_k = \bar{X} - Z$, [where S_k = skewness; $\bar{X}$ = arithmetic mean; Z = mode]

In a symmetrical distribution, mean and mode coincide, so skewness will be zero. If $\bar{X} > Z$, the skewness will be positive and will have positive sign. If $\bar{X} < Z$, the skewness will be negative and will have negative sign.

- **Karl Pearson's Co-efficient of Skewness**

  Karl Pearson has given a formula for relative measure of skewness. It is also known as Karl Pearson's of coefficient Skewness of Pearsonian Coefficient of Skewness. The formula is that the difference between the mean and mode is divided by the standard deviation.

  $$\text{Coefficient of } S_k = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}} = \frac{\bar{X} - Z}{\sigma} \quad \text{... (1)}$$

  Mode = 3 Median – 2 Mean

  Or $Z = 3M - 2\bar{X}$

  Substituting the value of modes in equation (1)

  $$\text{Coefficient of } S_k = \frac{\bar{X} - (3M - 2\bar{X})}{\sigma} = \frac{3(\bar{X} - M)}{\sigma}$$

  In a distribution, we have more than one mode, *i.e.*, mode is ill-defined, we cannot apply the above state formula. Then we have the following alternative formula:

  $$\text{Coefficient of } S_k = \frac{3(\bar{X} - M)}{\sigma}$$

Notes

- **Coefficient of Skewness in Individual Observations**

Example 1: Calculate the co-efficient of skewness of the following data by using Karl Pearson's method.

Marks:	2	4	4	6	7

Solution :

Marks (X)	$x = (X - A)$ **A = 4**	x^2
2	– 2	4
4	0	0
4	0	0
6	2	4
7	3	9
$\Sigma X = 23$		$\Sigma x^2 = 17$

$$\text{Mean } (\bar{X}) = \frac{\Sigma X}{N} = \frac{23}{5} = 4.6$$

As 4 is repeated twice therefore mode = 4.

Now $$\text{S.D. } (\sigma) = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{17}{5}} = \sqrt{3.40} = 1.844$$

Thus, co-efficient of skewness $$S_k = \frac{\bar{X} - \text{Mode}}{\sigma} = \frac{4.6 - 4}{1.844} = \frac{0.6}{1.844} = 0.325$$

- **Co-efficient of Skewness in Continuous Series**

Example 2: Find out the coefficient of skewness for the following distribution:

Class:	0–10	10–20	20–30	30–40	40–50
Frequency:	14	23	27	21	15

Solution:

Class (x)	**Frequency** (f)	**Mid-value** (m)	**Deviation** $d = (m - A)$ **A = 25**	fd	fd^2
0–10	14	5	– 20	– 280	5,600
10–20	23	15	– 10	– 230	2,300
20–30	27	25	0	0	0
30–40	21	35	+ 10	+ 210	2,100
40–50	15	45	+ 20	+ 300	6,000
	$\Sigma f = 100$			$\Sigma fd = 0$	$\Sigma fd^2 = 16000$

Notes

$$\bar{X} = A + \frac{\sum fd}{\sum f} = 25 + \frac{0}{100} = 25$$

Mode (Z) is located in the class interval 20–30.

$$Z = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 20 + \frac{27 - 23}{2 \times 27 - 23 - 21} \times 10$$

$$= 20 + \frac{4}{10} \times 10 = 20 + 4 = 24$$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{16000}{100} - \left(\frac{0}{100}\right)^2}$$

$$= \sqrt{160} = 12.65 \text{ Approx.}$$

$$\text{Coefficient of Skewness} = \frac{\bar{X} - Z}{\sigma} = \frac{25 - 24}{12.65} = \frac{1}{12.65} = 0.08$$

Example 3: The Karl Pearsons coefficient of skewness of a distribution is 0.32. The Standard Deviation is 6.5 and Mean is 29.6. Find mode.

Solution: Given $S_k = 0.32$, $\bar{X} = 29.6$, $\sigma = 6.5$, $Z = ?$

As we know $S_k = \frac{\bar{X} - Z}{\sigma}$

Substituting the values, we get

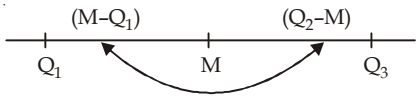
$$0.32 = \frac{29.6 - Z}{6.5}$$

$$\Rightarrow \quad 29.6 - Z = 6.5 \times 0.32$$

$$\Rightarrow \quad Z = 29.6 - 2.08 = 27.52$$

(2) Bowley's Method

This method is called Second Measure of Skewness, which is propounded by Dr. A.L. Bowley. This method is based on relative positions of the median and the two quartiles. In a symmetrical distribution, the upper and lower quartiles are equidistant from median *i.e.*,



Mathematically,

$$Q_3 - M = M - Q_1 \text{ or } Q_3 + Q_1 = M + M \text{ or } Q_3 + Q_1 - 2M = 0$$

Thus, here skewness is absent.

But in an asymmetrical distribution first and third quartiles are not equidistant from median, *i.e.*,

$$Q_3 + Q_1 - 2M \neq 0$$

Notes

In this case skewness is present. It is important to note that if Q_1 is farther away from median than the Q_3, then skewness will be negative and if the case is opposite the skewness will be positive. Dr. Bowley has given the following method of skewness.

$$S_k = (Q_3 - M) - (M - Q_1) = Q_3 + Q_1 - 2M$$

Coefficient of Skewness is

$$\text{Coefficient of } S_k = \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)}$$

$$\text{Coefficient of } S_k = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$$

The calculation of median and quartiles in the case of individual, discrete and continuous series is already explained in Unit - 4.

Example 4: If sum and difference of two quartiles are 22 and 8 respectively. Find the co-efficient of skewness when the median is 10.5.

Solution: Given $Q_3 - Q_1 = 8$; $Q_3 + Q_1 = 22$ and M = 10.5

$$\text{Now, Coefficient of } S_k = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} = \frac{22 - 2(10.5)}{8} = \frac{22 - 21}{8} = \frac{1}{8} = 0.125$$

Example 5: If Bowley's co-efficient of skewness is – 0.36, Q_1 = 8.6 and median = 12.3. What is the quartile co-efficient of dispersion ?

Solution: Given, Bowley's Coefficient of S_k = – 0.36, Q_1 = 8.6, M = 12.3 Coefficient of Q.D = ?

$$\text{Coefficient of } S_k = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1} \text{ or } -0.36 = \frac{Q_3 + 8.6 - 2 \times 12.3}{Q_3 - 8.6}$$

or $-0.36(Q_3 - 8.6) = Q_3 + 8.6 - 24.6$

or $-0.36Q_3 + 3.096 = Q_3 - 16$

or $-0.36Q_3 - Q_3 = -16 - 3.096$

or $-1.36Q_3 = -19.096$

or $Q_3 = \frac{19.095}{1.36} = 14.04$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{14.04 - 8.6}{14.04 + 8.6} = \frac{5.44}{22.64} = 0.24$$

(3) Kelly's Method

Prof. Kelly has given a formula which is based on deciles or percentiles. It is defined as

or $S_k = P_{90} + P_{10} - 2P_{50}$

or $S_k = D_9 + D_1 - 2D_5$

Coefficient of skewness is defined as

Notes

$$\text{Coefficient of } S_k = \frac{(P_{90}-P_{50})-(P_{50}-P_{10})}{(P_{90}-P_{50})+(P_{50}-P_{10})} = \frac{P_{90}+P_{10}-2P_{50}}{P_{90}-P_{10}}$$

$$\text{Coefficient of } S_k = \frac{D_9+D_1-2D_5}{D_9-D_1}$$

Example 6: Find out the Kelly's co-efficient of skewness of the data given below:

Class:	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency:	3	10	17	7	6	4	2	1

Solution:

Class	**Frequency (*f*)**	**Cumulative Frequency (*c.f.*)**
0–10	3	3
10–20	10	13
20–30	17	30
30–40	7	37
40–50	6	43
50–60	4	47
60–70	2	49
70–80	1	50
	$\Sigma f = 50$	

P_{10} = Size of $\frac{10N}{100}$ th item = $\frac{10(50)}{100}$ th item = 5th item

P_{10} = Size of 5th item which lies in 10–20 group

$$P_{10} = l_1 + \frac{\left(\frac{10N}{100} - c.f.\right)}{f} \times i$$

$$P_{10} = 10 + \frac{(5-3)}{10} \times 10 = 10 + 2 = 12$$

P_{50} = Size of $\frac{50 \times N}{100}$ th item = $\frac{50 \times 50}{100}$ th item = 25th item.

P_{50} = Size of 25th item which lies in 20-30 group.

$$P_{50} = l_1 + \frac{\left(\frac{50N}{100} - c.f.\right)}{f} \times i$$

$$P_{50} = 20 + \frac{(25-13)}{17} \times 10 = 20 + \frac{120}{17} = 27.06$$

Notes

$$P_{90} = \text{Size of } \frac{90 \times N}{100} \text{ th item} = \frac{90 \times 50}{100} \text{ th item} = 45\text{th item.}$$

Now P_{90} = Size of 45th item which lies in 50–60 group.

$$P_{90} = l_1 + \frac{\frac{90N}{100} - c.f.}{f} \times i$$

$$P_{90} = 50 + \frac{(45-43)}{4} \times 10 = 50 + \frac{20}{4} = 55$$

Measures of Skewness at a Glance

Methods	Formula
1. Karl Pearson's Method	
(a) Absolute Skewness	$S_k = \bar{X} - Z$
When mode is ill-defined	$S_k = 3(\bar{X} - M)$
(b) Coefficient of Skewness	Co-efficient of $S_k = \frac{\bar{X} - Z}{\sigma} = \frac{3(\bar{X} - M)}{\sigma}$
When mode is ill-defined	
2. Bowley's Method	
(a) Absolute Skewness	$S_k = (Q_3 - M) - (M - Q_1) = Q_3 + Q_1 - 2M$
(b) Coefficient of Skewness	$= \frac{(Q_3 - M) - (M - Q_1)}{(Q_3 - M) + (M - Q_1)} = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}$
3. Kelly's Method	
(a) Absolute Skewness	$= P_{90} + P_{10} - 2P_{50}$ or $D_9 + D_1 - 2D_5$
(b) Coefficient of Skewness	$= \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}$ or $= \frac{D_9 + D_1 - 2D_5}{D_9 - D_1}$

8.3 Kurtosis

Besides averages, variation and skewness, a fourth characteristic used for description and comparison of frequency distributions is the peakedness of the distribution. Measures of peakedness are known as measures of **kurtosis.**

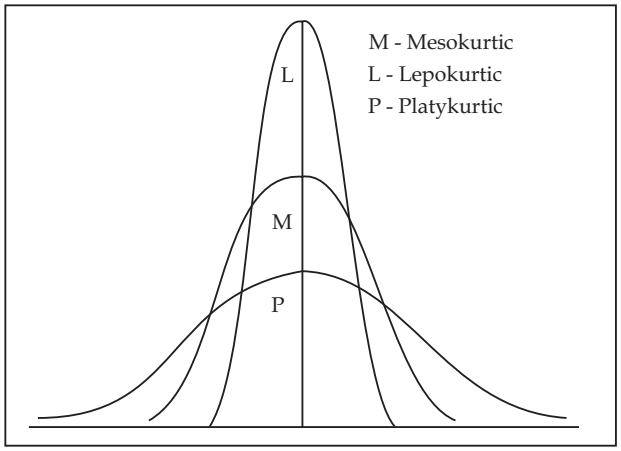
Notes Kurtosis in Greek means "*bulginess*". In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve.

In other words, measures of kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. If a curve is more peaked than the normal curve, it is called '*leptokurtic*'.

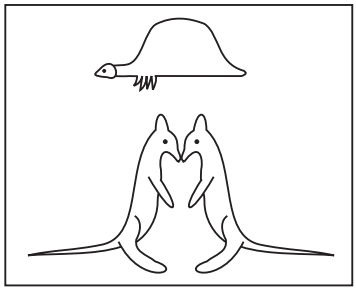
Notes

In such a case the items are more closely bunched around the mode. On the other hand, if a curve is more flat-topped than the normal curve, it is called '*platykurtic*'. The normal curve itself is known as '*mesokurtic*'. The condition of peakedness or flat-toppedness itself is known as kurtosis or excess. The concept of kurtosis is rarely used in elementary statistics.

The following diagram illustrates the shapes of three different curves mentioned above:



The above diagram clearly shows that these curves differ widely with regard to convexity, an attribute which Karl Pearson referred to as 'kurtosis'. Curve M is a normal one and is called 'mesokurtic'. Curve L is more peaked than M and is called 'leptokurtic'. Curve P is less peaked (or more flat-topped) than curve M and is called 'platykurtic'.



A famous British statistician Willian S. Gosset ("Student") has very humorously pointed out the nature of these curves in the sentence, "Platykurtic curves, like the platypus, are squat with short tails; lepto-kurtic curves are high with long tails like the kangaroos noted for lapping." Gosset's little sketch is reproduced above.

Measures of Kurtosis

The most important measure of kurtosis is the value of the coefficient β_2. It is defined as:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ where } \mu_4 = \text{4th moment and } \mu_2 = \text{2nd moment.}$$

For a normal curve the value of $\beta_2 = 3$. When the value of β_2 is greater than 3 the curve is more peaked than the normal curve, *i.e.*, leptokurtic. When the value of β_2 is less than 3 the curve is less peaked than the normal curve, *i.e.*, platykurtic. The normal curve and other curves with $\beta_2 = 3$ are called mesokurtic.

Sometimes γ_2, the derivative of β_2, is used as a measure of kurtosis, γ_2 is defined as

$$\gamma_2 = \beta_2 - 3.$$

Notes

For a normal distribution $\gamma_2 = 0$. If γ_2 is positive, the curve is leptokurtic and if γ_2 is negative, the curve is platykurtic.

Example 7: The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the skewness and kurtosis of the distribution.

Solution:

Testing Skewness

We are given $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Skewness is measured by the coefficient β_1

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

Here $\mu_2 = 2.5$, $\mu_3 = 0.7$

Substituting the values, $\beta_1 = \frac{(0.7)^2}{(2.5)^3} = +0.031$

Since $\beta_1 = +0.031$, the distribution is slightly skewed.

Testing Kurtosis:

For testing kurtosis we compute the value of β_2. When a distribution is normal or symmetrical, $\beta_2 = 3$. When a distribution is more peaked than the normal, β_2 is more than 3 and when it is less peaked than the normal, β_2 is less than 3.

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

$\mu_4 = 18.75$, $\mu_2 = 2.5$

$$\therefore \quad \beta_2 = \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3$$

Since β_2 is exactly three, the distribution is mesokurtic.

Self-Assessment

1. Fill in the blanks:

(*i*) If $Q_3 = 30$, $Q_1 = 20$, Med = 25, Coeff. of Sk. shall be .....

(*ii*) If $\overline{X} = 50$, Mode = 48, $\sigma = 20$, the Coefficient of Skewness shall be .....

(*iii*) If Coeff. of Sk. = 0.8, Median = 35, $\sigma = 12$, the mean shall be .....

(*iv*) In a symmetrical distribution the coefficient of skewness is .....

(*v*) The limits for Bowley's coefficient of skewness are .....

8.4 Summary

- The nature of distribution is further studied deeply by calculating 'Moments' which reveals whether the symmetrical curve is normal, more flat than normal or more peaked than normal.

Notes

Similarly, Kurtosis is yet another measure which tells us about the form of a distribution. Thus, it can be said that the central tendencies and dispersion measures should be supplemented by measures of skewness and kurtosis so that a more elaborate picture about the distribution given can be obtained. The study becomes more important in subjects of economics, sociology and other social sciences where normal distribution in a series usually does not occur. However, studies hold importance in biological sciences and other physical sciences as well.

- In the words of *Simpson* and *Kafka*, "Measures of skewness tell us the direction and the extent of skewness. In symmetrical distribution the arithmetic mean, median and mode are identicle. The more the mean moves away from mode, the larger the asymmetry or skewness."
- Measure of skewness indicates the difference between the manner in which items are distributed in a particular distribution compared with symmetrical or normal distribution. In a symmetrical distribution, frequencies go on increasing upto a point and then begin to decrease in the same fashion. There are various possible patterns of symmetrical distribution and normal distribution which is bell-shaped is one of these.
- A distribution in which more than half of the area under the curve is to the right side of the mode, it is said to be a positively skewed distribution. In this type of skewness the right tail is longer than the left tail. In this case, mean is greater than median and the median is greater than the mode and $Q_3 - M > M - Q_1$.
- A distribution in which more than half of the area under the distribution curve is to the left side of the mode, it is said to be a negatively skewed distribution. In this case, the elongated tail is to the left and mean is less than the median which is less than mode and $Q_3 - M < M - Q_1$.
- Kurtosis in Greek means "*bulginess*". In statistics kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve. In other words, measures of kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. If a curve is more peaked than the normal curve, it is called '*leptokurtic*'. In such a case the items are more closely bunched around the mode. On the other hand, if a curve is more flat-topped than the normal curve, it is called '*platykurtic*'. The normal curve itself is known as '*mesokurtic*'. The condition of peakedness or flat-toppedness itself is known as kurtosis or excess. The concept of kurtosis is rarely used in elementary statistics.
- A famous British statistician Willian S. Gosset ("Student") has very humorously pointed out the nature of these curves in the sentence, "Platykurtic curves, like the platypus, are squat with short tails; lepto-kurtic curves are high with long tails like the kangaroos noted for lapping."

8.5 Key-Words

1. Skewness and Kurtosis : Skewness and kurtosis are terms that describe the shape and symmetry of a distribution of scores. Unless you plan to do inferential statistics on your data set skewness and kurtosis only serve as descriptions of the distribution of your data. Be aware that neither of these measures should be trusted unless you have a large sample size.

   Skewness refers to whether the distribution is symmetrical with respect to its dispersion from the mean. If on one side of the mean has extreme scores but the other does not, the distribution is said to be skewed. If the dispersion of scores on either side of the mean are roughly symmetrical (i.e. one is a mirror reflection of the other, the distribution is said to be not skewed.

2. Kelly's Methods : In probability theory, the Kelly criterion, or Kelly strategy or Kelly formula, or Kelly bet, is a formula used to determine the optimal

Notes

size of a series of bets. In most gambling scenarios, and some investing scenarios under some simplifying assumptions, the Kelly strategy will do better than any essentially different strategy in the long run. It was described by J. L. Kelly, Jr in 1956.[1] The practical use of the formula has been demonstrated.

8.6 Review Questions

1. What do you understand by skewness ? Give various definitions. What are the various methods of measuring it.
2. Give the concept of kurtosis.
3. Distinguish between Pearson's and Bowley's measure of skewness.
4. State the formula for calculating Karl Pearson's coefficient of skewness.

Answers: Self-Assessment

1. (*i*) 0 (*ii*) 0.1 (*iii*) 108.2 (*iv*) zero (*v*) ± 1

8.7 Further Readings



Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 9: Correlation: Definition, Types and its Application for Economists

CONTENTS

Objectives
Introduction
9.1 Definition and Types of Correlation
9.2 Application of Correlation for Economists
9.3 Summary
9.4 Key-Words
9.5 Review Questions
9.6 Further Readings

Objectives

After reading this unit students will be able to:

- Know Correlation and Types of Correlation.
- Discuss the Application of Correlation for Economists.

Introduction

Correlation means a relation between two groups. In statistics, it is the measure to indicate the relationship between two variables in which, with changes in the values of one variable, the values of other variable also change. These variables may be related to one item or may not be related to one item but have dependence on the other due to some reason. For example, the data on height and weights of a group of people would relate to each member of the group but prices of sugar and sugarcane are two different series altogether but there would be some relation between the values of the two, prices of sugar depending upon the prices of sugarcane. This technique provides a tool into the hands of decision-makers because it provides better understanding of the trends and their dependance on other factors so that the range of uncertainities associated with decision-making is reduced.

9.1 Definition and Types of Correlation

Definition of Correlation

The term correlation indicates the relationship between two variables in which with changes in the value of one variable, the values of the other variable also change. Correlation has been defined by various eminent statisticians, mathematicians and economists. Some of the important definitions of correlation are given below:

(1) According to *La Yun Chow,* "Correlation analysis attempts to determine the degree of relationship between variables."

(2) As per *W. I. King,* "Correlation means that between two series or groups of data there exists some casual connections. ..... If it is proved true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions, we consider that the fact is established and that a relationship exists. This relationship is called correlation."

(3) In the words of *L. R. Conner,* "If two more quantities vary in sympathy so that movements in the one tend to be accompanied by corresponding movements in the other/others then they are said to be correlated.

Notes

(4) *Croxton* and *Lowden* define correlation as, "When the relationship of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

(5) As per *Prof. Boddington,* "Whenever some definite connection exists between two or more groups, classes or series of data, there is said to be correlation."

From the above definitions it can be said that correlation is a statistical tool which requires about the relationship between two or more variables.

Utility: Correlation has immense utility in various fields of knowledge. Some of the important areas where correlation has been used successfully are:

(1) In the field of genetics: Galton and Pearson developed a method of assessing correlation which was used in studying many problems of biology and genetics.

(2) In the field of management: Basically, management is all about making decisions. Correlation technique presents a strong tool into the hands of the manager which reduces the range of uncertainity associated with decision-making. Moreover, it also helps in identifying the stabilishing factors for a disturbed economic situation.

(3) Other field of social sciences: Correlation helps in determining the interrelationships between different variables and in this way it is very helpful in promoting research and opening new frontiers of knowledge.

In this way it can be said that correlation has immense utility in various fields in promoting research and opening new frontiers of knowledge.



Did u know? Correlation is very useful in understanding the economic behaviour. It helps in locating those variables on which other variables depend. In this way various economic events can be analysed.

"Correlation" and "cause and effect relationship"

Correlation measures a degree of the relationship between two or more variables but it does not indicate any kind of cause and effect relationship between the variables. If, high degree of correlation is found exist between two variables, it implies that there must be a reason for such close relationship, but the cause and effect relation can be revealed specifically when other knowledge of the factor involved being brought to bear on the situation. This means, to establish a 'functional relationship' between two or more variables, one has to go beyond the confines of statistical analysis to other factors. (Functional relationship means that two or more factors are interdependent. In fact, although, high degree of correlation may mean that two or more variables are mutually dependent, but at the same time, this high degree of correlation may be due to many other reasons like:

(1) The two variables are being affected by a third variable or by more than one variable.

(2) The two variables might be mutually affecting each other and neither of them is the cause or the effect.

(3) The high degree of correlation between two variables comes out just by chance or by sheer coincidence.

Therefore, although high degree of correlation does not necessarily indicate the cause and effect relationship. The quantitative tool requires the support of proper knowledge and logic about the variables on the basis of which the results should be interpreted. In this way, although 'correlation' in a strong tool it needs to be used carefully by those who have knowledge otherwise its misuse is quite likely.

Types of Correlation

Correlation can be classified as given ahead:

(1) Positive and negative correlation: When the values of the two variables move in the same direction, *i.e.,* an increase in one is associated with an increase in other, or *vice versa,* the correlation

Notes

is said to be positive. If the values of two variables move in the opposite directions *i.e.,* an increase in the value of one variable is associated with fall in other, or *vice versa,* the correlation is said to be negative. For example, the price and supply are positively correlated but price and demand are negatively correlated.

(2) **Linear and non-linear correlation:** If, in response to a unit change in the value of one variable, there is a constant change in the value of the other variable, the correlation between them is said to be linear. This means, the relation between variables fits in $Y = a + bX$. But when no constant change in variable is registered for a given unit change in other variable, non-linear or curvilinear correlation is said to exist.

(3) **Simple, multiple and partial correlation:** When relation between two variables is studied, it is simple correlation. When three or more factors are studied together to find relationships, it is called multiple correlation. In partial correlation, two or more factors are agreed to be involved but correlation is studied between only two factors, considering other factors to be constant.

9.2 Application of Correlation for Economists

The cause and effect relation existing between economic events is especially difficult to ascertain because of the presence of innumerable variable elements. In solving his problems the economist can not, like the physicist or chemist, eliminate all causes except one and then by experiment determine the effect of that one. Causes must be dealt with *en masse.* Since any effect is the result of many combined causes the economist is never sure that a given effect will follow a given cause. In stating an economic law he always has to postulate "other things remaining the same," with, perhaps, little appreciation of what the other things may be. It is rarely, if ever, possible for the economist to state more than "such and such a cause *tends* to produce such and such an effect." Events can only be stated to be more or less probable. He is dealing mainly, therefore, with correlation and not with simple causation.

The problems of economics are similar to certain problems of biology, such as the effect of environment and heredity upon the individual. In dealing with the question of heredity Karl Pearson says: "Taking our stand then on the observed fact that a knowledge neither of parents nor of the whole ancestry will enable us to predict with certainty in a variety of important cases the character of the individual offspring we ask: What is the correct method of dealing with the problem of heredity in such cases ? The causes A, B, C, D, E , . . . which we have as yet succeeded in isolating and defining are not always followed by the effect X, but by any one of the effects U, V, W, X, Y, Z. We are therefore not dealing with causation but correlation, and there is therefore only one method of procedure possible; we must collect statistics of the frequency with which U, V, W, X, Y, Z, respectively, follow on A, B, C, D, E . . . From these statistics we know the most *probable* result of the causes A, B, C, D, E and the frequency of each deviation from this most probable result. The recognition that in the existing state of our knowledge the true method of approaching the problem of heredity is from the statistical side, and that the most that we can hope at present to do is to give the *probable* character of the offspring of a given ancestry, is one of the great services of Francis Galton to biometry."

Just as the biologists cannot predict a man's height or color of eyes or temper or combativeness by knowing those qualities in his ancestors, so economists cannot predict that a definite call rate in Wall Street will go with a given percentage of reserves to deposits in New York banks or that a given supply of wheat will result in a definite price per bushel. But, on the other hand, just as it has been observed that there *is* a relation existing between a man's stature and the stature of his ancestors, so it has been observed that a relation *does* exist between bank reserves and call rates and between supply of wheat and its price per bushel.

In order to deal in a satisfactory way with such questions as those given above it is necessary to accumulate statistics of the supposedly related phenomena. In order to have those statistics indicate anything it is necessary to obtain a method of measuring the extent of correlation between the phenomena.

The commonly used method of measuring the amount of correlation between any two series of economic statistics is to represent the two series graphically upon the same sheet of cross-section paper and then compare the fluctuations of one series with those of the other. The quantity theory of

prices has been tested in this way by Dr. E. W. Kemmerer. Dr. Kemmerer builds up the following price equation:

Notes

$$P_s = \frac{MR+CR_c}{NE+N_cE_c}$$

in which:

P_s = the average price (weighted by the total flows) of all commodities sold for money and deposit currency during a unit of time.

M = the total currency in circulation during the unit of time
R = the average number of times each unit of currency changes hands during the unit of time.
} = the flow of currency

NE = the flow of goods exchanged for currency.

C = the volume of deposit currency exchanged for goods.
R_c = the average rate of turnover of such deposit currency.
} = flow of deposit currency.

N_cE_c = the flow of goods exchanged for deposit currency.

Dr. Kemmerer then attempts to find the answer that facts give to the following questions:

1. Do the bank reserves vary directly with the money supply ?
2. Does the proportion of bank reserves to check circulation vary directly with the degree of business distrust existing in the country ?
3. Is "a relative increase in the circulating media accompanied by a corresponding and proportionate increase in general prices and a relative decrease in the circulating media, by a corresponding and proportionate decrease in general prices," or, in the language of the formula, is

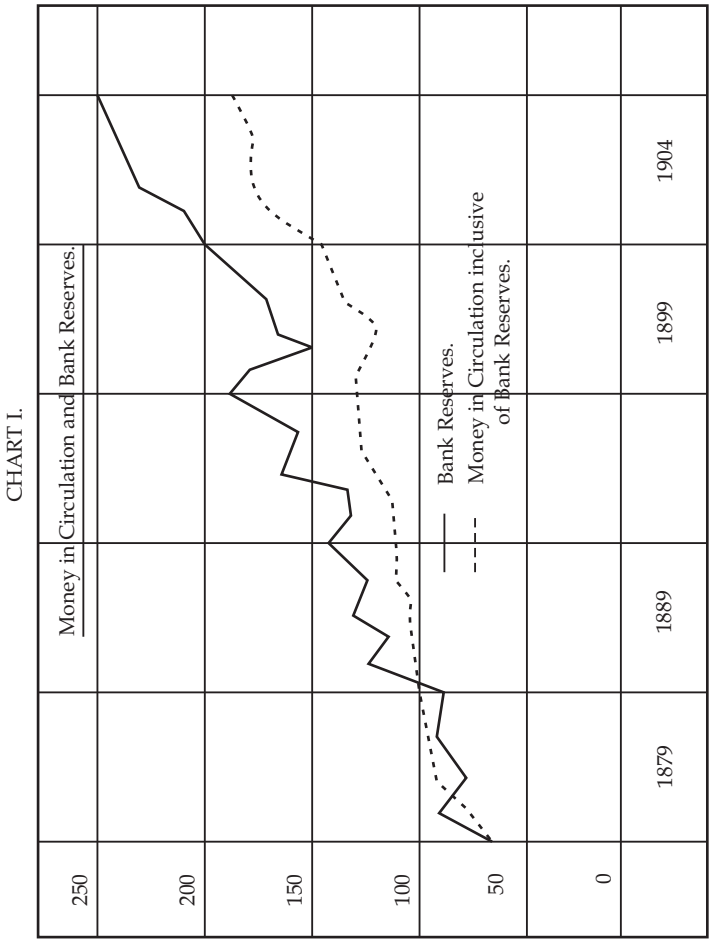
$$P_s = \frac{MR+CR_c}{NE+N_cE_c}$$

borne out by the facts ?

All of the questions to be tested by the statistics collected are questions of correlation. Dr. Kemmerer makes the tests graphically, as has been stated, by comparing the fluctuations of the two curves based upon the pair of series of statistics being considered. The charts presented by Dr. Kemmerer from which his conclusions are drawn are given below.

In the case of the correlation of bank reserves and money in circulation, inclusive of bank reserves, Dr. Kemmerer concludes, "There can be no question but that when due allowance is made for fluctuations in business confidence, the evidence of Chart I strongly supports the contention that there exists a close relationship between the amount of money in circulation and the amount of the country's bank reserves." In the case of the correlation of business distrust and the ratio of bank reserves to check circulation the conclusion is, "the chart substantiates the contention . . . that the ratio of check circulation to bank reserves is a function of business confidence . . . "The final test of the quantity theory is the amount of correlation between the figures for the right and left-hand sides of the equation $P_s = \frac{MR+CR_c}{NE+N_cE_c}$.

Notes



Upon examination of the curves plotted from the two series of statistics representing general prices and relative circulation (the left and right-hand sides, respectively, of the price equation) Dr. Kemmerer concludes, "The general movement of the two curves taken as a whole is the same, while the individual variations from year to year exhibit a striking similarity."

The graphic method of comparing fluctuations is well enough as a preliminary, *but does it enable anyone to tell anything of the extent of the correlation between the series of figures being considered* ? Is Dr. Kemmerer warranted in deducing his conclusions from observation of the charts ? It seems to the writer that one opposing the quantity theory might draw opposite conclusions with as much (or as little) reason. *The charts do not answer the questions proposed.* The painstaking collection of statistics to test correlation is useless if there be no more reliable method to measure correlation. A numerical measure of the correlation must be found if we wish to determine the *extent* to which the fluctuations of one series synchronize with the fluctuations of another series.

A second illustration of a conclusion based upon graphic representation is that of Ira Cross in his study of strike statistics. He says, upon consideration of data taken from the Twenty-first Annual Report of the United States Bureau of Labor, "the percentage of successful strikes decreases during periods of business prosperity and increases during 'hard times.' " In the accompanying charts the per cent. of establishments in which strikes were successful is plotted, first, with the per capita exports and imports and second, with index numbers of wholesale prices. The foreign trade and the price statistics are taken as indicative of the activity of business, as indices of prosperity.

A third illustration of a conclusion relating to correlation is taken from the *London Statist* of April 4, 1908, where the proposition is made that, "When commodities advance prices of Stock Exchange securities recede; when commodities recede Stock Exchange securities advance." The proposition is

Notes

supported by reference to the following chart showing the yearly average price of consols and Sauerbeck's index numbers of prices.

The foregoing illustrations show the need by economists of a quantitative measure of correlation. Such a measure has been widely used in biological statistics and used to a limited extent in economic statistics. G.U. Yule has used the measure in his study of "Pauperism ;" R.H. Hooker has used it in his "Correlation of the Weather and Crops;" J. P. Norton applied it in his study of the "New York Money Market." This measure, the coefficient of correlation, will be computed for the data upon which the conclusions quoted above are based. The formula for the coefficient of correlation is

$$r = \frac{\Sigma xy}{n\sigma_1\sigma_2};$$

where:

x = deviation from arithmetic mean = $X - M_1$

y = deviation from arithmetic mean= $Y - M_2$

σ_1 = standard deviation of X series

σ_2 = standard deviation of Y series

n = number of items.

The coefficient of correlation "serves as a measure of any statement involving two qualifying adjectives, which can be measured numerically, such as tall men have tall sons,' 'wet springs bring dry summers,' 'short hours go with high wages.' " It is not the purpose in what follows to go through the mathematical derivation of the coefficient of correlation, but to test the formula empirically in order to ascertain how it actually varies for given series of statistics and to point out some of its features.

However, it should be noted at this point that the coefficient of correlation is not empirical but was derived by *a priori* reasoning. It was found by assuming that a large number of independent causes operate upon each of the two series X and Y, producing normal distributions in both cases. Upon the assumption that the set of causes operating upon the series X is *not independent* of the set of causes operating upon the series Y the value $r = \frac{\Sigma xy}{n\sigma_1\sigma_2}$ is obtained. This value becomes zero when the operating causes are absolutely independent. Hence the value of r was taken as a measure of correlation. In what follows *no assumption concerning the type of distribution of the X and Y series will be made.*

Some appreciation of the meaning of the coefficient of correlation can be obtained by the consideration of a few simple applications. Suppose that we consider the two series of measurements:

X = l, 2, 3, 4, 5 $\quad M_1 = 3$

Y = 6, 8, 10, 12, 14 $\quad M_2 = 10$

Deviations.		Square of Deviation.		Product of Deviations.	
x	y	x^2	y^2	xy	
– 2	– 4	4	16	8	$\sigma_1 = \sqrt{2}$
– 1	– 2	1	4	2	
0	0	0	0	0	$\sigma_2 = 2\sqrt{2}$
+ 1	+ 2	1	4	2	$r = \frac{20}{5\sqrt{2}.2\sqrt{2}} = 1$
+ 2	+ 4	4	16	8	

Notes

In the above illustration the numbers were chosen so that for an increase of 1 unit in the X series there is an increase of 2 units in the Y series. Thus the correlation is perfect and r equals +1. If the Y series had been 14, 12, 10, 8, 6 (the X series remaining the same) the value of r would have been −1. Thus −1 stands for perfect *negative* correlation, an increase in one series corresponding to a decrease in the other. It should also be noted in this connection that the coefficient of correlation (r) cannot be less than −1 nor more than + l.

The above illustration suggests the question, "Will a linear relationship between X and Y *always* give perfect correlation ?"

Assume the linear relationship

$$Y = aX + b$$

Since $y = Y - M_2$ and $x = X - M_2$

$$M_2 + y = a(x + M_1) + b \text{ or } y = ax$$

(since $b - aM_1 - M_2 = 0$)

and $$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}} = \frac{\sum ax^2}{\sqrt{\sum x^2 . \sum a^2 x^2}} = \frac{a\sum x^2}{\sqrt{a^2\left(\sum x^2\right)^2}} = \pm 1$$

(The sign of r depends upon the sign of a.)

Therefore a linear relationship between two variables will give a correlation coefficient of +1 or −1 depending upon whether large values of one occur with large values of the other or large values of one occur with small values of the other.

The converse of the above proposition is likewise true, *i.e.,* if the coefficient of correlation (r) equals 1 then the relationship between the X and Y series is linear.

Assume $r = 1$

then $\left(\sum xy\right)^2 - \sum x^2 . \sum y^2 = 0$

Letting $x_1 = \lambda_1 y_1, x_2 = \lambda_2 y_2 \ldots x_n = \lambda_n y_n$ the above expression becomes

$$y_1^2 y_2^2 (\lambda_1 - \lambda_2)^2 + y_1^2 y_3^2 (\lambda_1 - \lambda_3)^2 + \ldots + y_r^2 y_s^2 (\lambda_r - \lambda_s)^2 + \ldots = 0$$

The only way in which this expression can equal zero is by having

$$\lambda_1 = \lambda_2 = \lambda_3 = \ldots = \lambda_n$$

and it follows that

$$x_1 = \lambda_1 y_1, x_2 = \lambda_1 y_2 \ldots x_n = \lambda_1 y_n$$

or

$$x = \lambda_1 y$$

which denotes a linear relationship between X and Y.

That any relation other than a linear one will not lead to r = l is illustrated by the following:

Let $Y = X^2$

$X = 1, 2, 3, 4, 5,$ $\quad M_1 = 3$

$Y = 1, 4, 9, 16, 25,$ $\quad M_2 = 11$

Notes

x	y	x^2	y^2	xy	
– 2	– 10	4	100	20	
– 1	– 7	1	49	7	σ_1 = 1.41
0	– 2	0	4	0	σ_2 = 8.65
+ 1	+ 5	1	25	5	r = 0.981
+ 2	+ 14	4	196	28	
Total		10	374	60	

Although the two series increase regularly, so that deviations of like signs always correspond, yet the correlation is not perfect *because a linear relation does not exist between X and Y.*

If the number of items in each series be increased to 11 and the Y items remain squares of the X's the value of *r* will be 0.974.

If there be no law connecting the X and Y series the products of the deviations (*xy*) are as apt to be negative as positive. The expression Σxy will therefore tend to approach zero. With a very large number of measurements the correlation coefficient will approximate zero.

From the condition of no relationship to the condition of a linear relationship existing between the pair of series of measurements the correlation coefficient varies from 0 to ± 1.

Suppose that we are investigating the relation existing between two series of measurements X and Y. Let points be plotted on cross-section paper whose coordinates are corresponding measurements X_1 and Y_1. If there be a relationship existing between the two series, the points thus located will not lie chaotically all over the plane, but they will range themselves about some curve or locus. This curve, which has been called the *curve of regression,* is illustrated in the accompanying diagram. The straight line best fitting the points is called the line of regression.

For example suppose we consider the two series of index numbers for the period 1879-1904 inclusive, representing (1) money in circulation in the United States inclusive of bank reserves, and (2) bank reserves. Let points be located with abscissas proportionate to the money in circulation and with ordinates proportionate to the bank reserves of the same year. The chart on the next page shows that these points lie near a straight line, the line of regression.

The coefficient of correlation (*r*) *is a measure of the closeness of the grouping of the points about this line of regression.* If the points should all range themselves on a line then *r* would equal + 1 or — 1 depending upon whether, looking left to right, the line sloped upward or downward.

We will now derive the equation of the line of regression. Let X and Y be associated measurements and *x* and *y* be associated deviations from the respective arithmetic means. A *linear* relation between the measurements is of the form

$$Y = a_1X + b_1$$

The relation between the deviations will be of form

$$y = a_1x \text{ or } y - a_1x = 0$$

Since all of the points are not located exactly upon a straight line the substitution of the values x_1, y_1, x_2, y_2, etc. in the equations will give residues v_1, v_2, etc. as follows:

$$y_1 - a_1x_1 = v_1$$

$$y_2 - a_1x_2 = v_2$$

$$y_n - a_1x_n = v_n$$

Notes

The values $\frac{v_1}{\sqrt{1+a_1^2}}, \frac{v_2}{\sqrt{1+a_1^2}}, \ldots . \frac{v_n}{\sqrt{1+a_1^2}}$ equal the distances of the various points to the straight line $y = a_1x$.

The equation of a line such that the sum of the squares of the distances from the given points is a minimum will now be found. In other words that value of a_1 will be taken which makes $v_1^2 + v_2^2 + \ldots . v_n^2$ = a minimum. To find the value of a_1, for which $(y_1 - a_1x_1)^2 + (y_2 - a_1x_2)^2 + \ldots . + (y_n - a_1x_n)^2$ will be a minimum, differentiate with respect to a_1 and obtain $-2x_1(y_1 - a_1x_1) - 2x_2(y_2 - a_1x_2) \ldots . -2x_n(y_n - a_1x_n)$. In order that the original function be a minimum, this derivative must equal zero. We will then have

$$(x_1y_1 - a_1x_1^2) + (x_2y_2 - a_1x_2^2) \ldots . + (x_ny_n - a_1x_n^2) = 0, \text{ or}$$

$$\Sigma xy - a_1 \Sigma x^2 = 0$$

$$a_1 = \frac{\Sigma xy}{\Sigma x^2}$$

Similarly if $x = a_2y$, then $\Sigma xy - a_2 \Sigma y^2 = 0$ will give the value of a_2 for which the sum of the squares of the distances of the given points to the straight line $X = a_2Y + b_2$ is a minimum, or

$$a_2 = \frac{\Sigma xy}{\Sigma y^2}$$

Let $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{\Sigma xy}{n\sigma_1\sigma_2} \text{ and } \Sigma x^2 = n\sigma_1^2 \Sigma y^2 = n\sigma_2^2.$$

The equations between the deviations are:

$$y = r \frac{\sigma_2}{\sigma_1} x$$

$$x = r \frac{\sigma_1}{\sigma_2} y$$

It may seem that the two equations just given are inconsistent. But it must be remembered that these equations do not give the relationship existing between *any* corresponding pair of deviations unless all of the points lie exactly on a straight line and there be perfect correlation. For all cases of imperfect correlation a *given* deviation x will occur with several different deviations y (if we have a large number of measurements). If these deviations y are distributed according to the normal law of distribution then the given value x substituted in the first equation will give the mean of the deviations occuring with the deviation x and if a given value y be substituted in the second equation the value of x resulting will be the mean of the deviations of the associated characteristics.

Since $$y = Y - M_2 \text{ and } x = X - M_1$$

$$Y = M_2 + r\frac{\sigma_2}{\sigma_1}(X - M_1)$$

Notes

and $$X = M_1 + r\frac{\sigma_1}{\sigma_2}(Y - M_2)$$

The coefficients $r\frac{\sigma_2}{\sigma_1}$ and $r\frac{\sigma_1}{\sigma_2}$ are called the coefficients of regression of Y upon X and of X upon Y respectively. The first coefficient $\left(r\frac{\sigma_2}{\sigma_1}\right)$ and the reciprocal of the second $\left(\frac{\sigma_2}{r\sigma_1}\right)$ are the *slopes* of the lines of regression. If X and Y be measured in terms of their respective standard deviations as units the slopes of the lines of regression will be *r* and $\frac{1}{r}$. *In other words, the slope of the line of regression of Y upon X, each series being measured in terms of its standard deviation, is equal to the coefficient of correlation for the two series. For perfect positive correlation the line would make an angle of 45° with the X axis for perfect negative correlation the line would make an angle of 135° with the x axis, and for no correlation the line would be parallel to the x axis.*

The correlation coefficients show that there is a very great difference in the degree of correlation of different pairs of series of statistics. The full significance of the "probable error," which is used as a measure of unreliability of any determination, cannot be developed at this point. It is sufficient to note that, "When *r* is not greater than its probable error we have no evidence that there is any correlation, for the observed phenomena might easily arise from totally unconnected causes; but, when *r* is greater than, say, six times its probable error, we may be practically certain that the phenomena are not independent of each other, for the chance that the observed results would be obtained from unconnected causes is practically zero."

The high degree of correlation (+0.98) between money in circulation inclusive of bank reserves and bank reserves is due to the tendency of the two items to vary together during the long time period and not due to correspondence of minor fluctuations. The reasons for the great increase of money in circulation in the United States during the period 1879-1904 are the great increase of population and the industrial expansion. Likewise the number of banks increased in order to serve the increased population and this meant an increase of total reserves. It is self-evident that the long time tendency of the two series of statistics must be upward in a growing country. It seemed to me that the bank reserves during the 26 years, 1879-1904, would be as closely correlated with the *population* as with total circulation. The computation of the correlation coefficient between bank reserves and population gave +0.98. It is the variation upwards of both series during the entire period that causes the high coefficient.

The correlation coefficient between the index numbers of business distrust and the rates of bank reserves to check circulation for the same years is 0.53. When the index numbers of business distrust for one year are correlated with the ratio of bank reserves to check circulation the following year the coefficient is 0.72. As Dr. Kemmerer has suggested (but not verified), there is a closer correlation "when proper allowance is made for the time required for alterations in business confidence to exert their influence on bank reserves."* The lowest correlation (+ 0.23), that between relative circulation and general prices, is not high enough to warrant a conclusion that the items vary together. The smallness of the correlation indicated may have resulted either because the quantity theory is in error or because the statistics are not adequate to test the theory. Whatever may be the fact, the statistics and the method of measuring correlation presented by Dr. Kemmerer do not demonstrate that general prices move in sympathy with relative circulation.

The amount of correlation indicated in each case is small—considering the number of years taken, so small that no conclusion as to the connection between the two series can be drawn. The correlation coefficient in the last instance, *i. e.,* between per cent. of successful strikes and business distrust, suggests an opposite conclusion to that indicated by the other coefficients and that of Mr. Cross. The

Notes

analysis shows that the conclusion that there is negative correlation between *general* prosperity and per cent. of successful strikes is not warranted.

Finally, what is the degree of correlation between the prices of British Consols and Sauerbeck's index numbers of the prices of commodities ? The chart on indicates a greater degree of correlation (negative) between the *minor* fluctuations of the two series than shown by any of the pairs of series that we have considered. The coefficient of correlation based upon statistics for the 57 years from 1851 to 1907, inclusive, is -0.58 ± 0.06. A correlation coefficient of -0.58 based upon 57 pairs of items warrants the conclusion that the two series have inverse movements.

The relations between the *average* deviations, x and y, of the two series of statistics being considered are:

$$y = -1.465x \text{ and } x = -0.2295x$$

The equations of regression are:

$$Y = 225.6 - 1.465\ X \text{ and } X = 19.439 - 0.2295\ Y$$

For certain pairs of time-series (corresponding items occur at same time) of measurements a correlation coefficient approximating zero may be obtained even though graphs of the statistics show that the up-and-down fluctuations occur together. This result will come about if the *long-time* variations show opposite tendencies, as, for instance, in the statistics of marriages and bank clearings in the United Kingdom. On the other hand, a *high* correlation coefficient may be obtained for two series having the same long-time tendency regardless of the non-correspondence of the short-time fluctuations. For example, the coefficient for the two series, population and bank reserves, came out to be 0.98. This high coefficient comes from the fact that the long-time variation of both series is the same. Consequently, before it is legitimate to draw any conclusions as to the meaning of a lack of correlation, or amount of correlation between two series of measurements it is necessary to ascertain the periodic and the secular variations in the two series. This correlation coefficient may be large through the correspondence of either secular or periodic variation, or both. It may be null because one variation covers up the other.

Three methods have been used for isolating the short-time variations of time-series of measurements. They will now be considered.

1. If upon plotting the two series being compared with time as abscissa and the measurements as ordinates, *periodic* variations appear at approximately equal intervals of time the curve may be "smoothed" and the secular variations may be eliminated as follows:
   (a) Ascertain the length of the wave by finding the number of time units between corresponding parts of the waves, *i. e.,* crest to crest, or hollow to hollow. Let 1 represent the number of time units found.
   (b) Average groups of 1 consecutive measurements, placing the points, determined by these averages at the middle of each group of measurements. Take enough groups so that the points obtained will indicate the general tendency of the series.
   (c) Draw a smooth curve through the points located by the process described in (b). This curve shows the secular tendency.
   (d) Subtract (this can be done graphically on cross-section paper) the ordinates of the "smoothed" curve from those of the original curve in order to obtain the series of measurements of the periodic fluctuation. Let d stand for any one of these differences.
   (e) The coefficients computed for corresponding ordinates of two smoothed curves, and for corresponding differences, d and d', give measures of the secular and periodic correlation, respectively.

   The method described above has been applied by Mr. R. H. Hooker in his paper "On the Correlation of the Marriage-Rate with Trade," and by Mr. G. U. Yule in his study of " Changes in Marriage and Birth-Rates in England and Wales during the Past Half Century. " The following table gives the correlation coefficients computed in the articles named for the *periodic* variation:

Notes

Series	Period	Deviations from	Coefficient of Correlation
{Marriage rate ..... / Imports plus exports per capita .....}	1861–1895	9 yr. means	+ 0.86
{Marriage rate ..... / Amount of bank clearings per capita}	1876–1895	9 yr. means	+ 0.47
{Marriage rate ..... / Sauerbeck's index numbers of prices.}	1865–1896	11yr. means	+ 0.795
{Marriage rate ..... / Hartley's index numbers of unemployment}	1870-1895	11 yr. means	– 0.873

The effect of using the deviations rather than the original series in computing the coefficient is shown by the comparison of the first correlation coefficient of + 0.86, given above, with the correlation coefficient of + 0.18, obtained for the same two series of *original* measurements for the same period, 1861–1895.

Using the deviation-method, Mr. Yule computed the correlation coefficients between *first,* the marriage rate of one year (m), and *second,* exports (e), imports (i), total trade (t), the price of wheat (w), and bank clearings (c) for the same year, and for each of several preceding years in order to answer the question, "does the maximum amount of correlation occur when corresponding items are of same year or when the marriage rate of one year is paired with the business item for a preceding year ?"

Mr. Yule says, "Fitting a parabola to the three values thus determined, a maximum correlation of about 0.482 must subsist between the birth-rate and the marriage-rate of 2.17 (two years and two months) previously."

Further analysis leads Mr. Yule to the conclusion that birth-rate is independently (not through marriage-rate only) sensitive to short-time economic changes and that the birth-rate is lowered after a depression, not only because of a decrease in the number of marriages during such depression, but also to a decrease in fertility.

2. In case the statistics show a long-time tendency with no *regular periodic* fluctuation Mr. R. H. Hooker has suggested that the "differences between successive values of the two variables, instead of the differences from the arithmetic means"* be correlated. Put into mathematical symbols we have:

Letting $\begin{Bmatrix} X_0, X_1 .... X_n \\ X'_0, X'_1 .... X'_n \end{Bmatrix}$ represent two series of measurements, and $\begin{Bmatrix} d_1, d_2 ... d_n \\ d'_1, d'_2 ... d'_n \end{Bmatrix}$ represent differences between any two consecutive measurements, and $\begin{Bmatrix} d_m \\ d'_m \end{Bmatrix}$ represent the respective means of these differences,

then $$d_m = \frac{X_n - X_o}{n} = \frac{\Sigma d}{n}, \text{ and}$$

$$d'_m = \frac{X'_n - X'_o}{n} = \frac{\Sigma d'}{n};$$

and the standard deviations of the differences are

$$\delta = \sqrt{\frac{\Sigma (d - d_m)^2}{n}}$$

Notes

$$\delta' = \sqrt{\frac{\Sigma(d' - d'_m)^2}{n}};$$

and the coefficient of correlation is

$$\rho = \frac{\Sigma(d - d_m)(d' - d'_m)}{n\delta'\delta} = \frac{\Sigma dd' - nd_m d'_m}{\sqrt{\left(\Sigma d^2 - nd_m^2\right)\left(\Sigma d'^2 - nd'^2_m\right)}}$$

Comparing this method of differences with the method described in (1) Mr. Hooker says, "Correlation of the deviations from an instantaneous average (or trend) may be adopted to test the similarity of more or less marked periodic influences, correlation of the difference between successive values will probably prove most useful where the similarity of the shorter rapid changes (with no apparent periodicity) are the subject of investigation, or where the normal level of one or both series of observations does not remain constant." He finds that the ordinary correlation coefficient (r) for the price of corn in Iowa and total production in the United States for the period 1870 – 1899 is – 0.28, while ρ = – 0.84.



Notes

The coefficient of correlation (r) *is a measure of the closeness of the grouping of the points about this line of regression.* If the points should all range themselves on a line then r would equal + 1 or — 1 depending upon whether, looking left to right, the line sloped upward or downward.

I have computed ρ for the statistics of corn production in the United States and the average farm price on December 1* for the period 1866 – 1906 and finds ρ = – 0.833 ± 0.034. Letting x represent the production *difference* in millions of bushels, and y represent the price *difference* in cents per bushel, the equations of regression are

$$y = -0.0256\,x + 1.132$$

$$x = -27.05\,y + 46.42$$

A graphic representation of the points whose abscissas and ordinates are the corresponding production and price differences, respectively, and the line of regression is given. The lack of correlation between the original pair of series is shown by the chart.

From the equations of regression such statements as the following can be made:

(i) For no change in corn production there is an increase in price of 1.132 cents per bushel.

(ii) For an increase in production of 100 million bushels the price decreases 1.43 cents per bushel.

(iii) For a decrease in production of 100 million bushels the price increases 3.69 cents per bushel.

(iv) For a stationary price the production must increase 46 million bushels per year.

It seemed to me that if *percentage* changes in price and production were used instead of absolute changes a still closer correlation might result. The computation of ρ from such percentages, however, gave – 0.794.

In the preceding illustrations the amount of correlation between the differences was greater than that between the original series. The method of differences has also been used by the writer for Kemmerer's statistics (considered on page 15 of this article) of (1) money in circulation, and (2) bank reserves for the period 1879 – 1904 with the result ρ = + 0.392, whereas the value of r is 0.98. This shows that there is a lack of correspondence of the short-time variations in these two series.

Notes

3. A third method of eliminating the long-time tendency and thus isolating the short-time fluctuations is to assume some curve, represented by an algebraic equation, which "fits" the statistics in question. The first step in the process is to select some curve, which, for *a priori* or other reasons is considered the best representation of the "growth element."* The second step is to fit the curve to the statistics; stated algebraically, to determine the constants in the equation of the curve by use of the actual data. Finally the deviations of the original measurements from the smooth curve (called by Norton "the growth axis") are computed. The accuracy with which one law, the geometric, $y = bc^x$, describes the population of the United States, and consequently many things that depend upon population is shown by the following diagram. The full points are fixed by the actual population according to each of the censuses from 1850. The smooth line is the graph of the equation

$$y = 24{,}086{,}000\ (1.0238)^x,$$

which equation was determined from the actual population.

Prof. J. P. Norton has applied the method here described to determine the correlation existing between percentage of reserves to deposits of New York Associated Banks and call rates.* Weekly statistics were taken for the period 1885 – 1900. The growth axes assumed were the geometric curve, $y = bc^x$, and the straight line $y = a$, respectively. (y = the measurement, x = time measured in weeks, while a, b and c are constants to be determined from the data.) The typical periodic fluctuations of *percentage* deviations of reserves and loans were also correlated by this method, using $y = bc^x$ as the growth function in both cases. The following table gives the correlation coefficients, ρ.

Series	ρ
Reserve deviations and discount rate	-0.37 ± 0.02
(a) Reserve and (b) Loan Periods Immediate.....	$+0.49 \pm 0.07$
(a) precedes (b) by one week	$+0.62 \pm 0.06$
(a) precedes (b) by two weeks	$+0.87 \pm 0.02$
(a) precedes (b) by three weeks	$+0.96 \pm 0.01$
(a) precedes (b) by four weeks	$+0.91 \pm 0.05$

The conclusion from this study is that "the loan period is really the shadow of the reserve period" ... and apparently follows the latter by "an interval of approximately three weeks."

Up to this point the problem before us has been the measurement of the amount of correlation between two variables. This is the simplest case of the general problem of the measurement of the amount of correlation between one series of measurements, and a group of any number of series of measurements. The solution of the general problem leads to very complex relations, § and it will not be taken up here. The case of three variables will be considered briefly.

Messrs. R. H. Hooker and G. U. Yule have considered the problem, To find the relation between the production of wheat in India during the period 1890 – 1904 (years ending March 31), the price of wheat (calendar years), and the exports of the subsequent twelve months, 1891 – 1905 (years ending March 31). The correlation of the annual differences according to the method described in (2) of gives the following results:

Series Correlated	Coefficient Correlation
1. Exports and Production	+ 0.77
2. Exports and Price	+ 0.86
3. Production and Price combined in the ratio 1: 1, and Exports ...	+ 0.90
4. Production and Price combined in the ratio 3: 1, and Exports ...	+ 0.81
5. Production and Price in the ratio 1: 3, and Exports ...	+ 0.58

Notes

The table indicates that exports depend upon production and price, and depend equally upon them.

Messrs. Hooker and Yule give the following general solution of the special problem just considered:

To find the maximum correlation coefficient between x_1 and $x_2 + bx_3$ that results from considering b a variable, where x_1, x_2, and x_3 are the deviations of the series X_1, X_2, and X_3 from their respective arithmetic averages.

Let $x_2 + bx_3 = z$

then $\Sigma(x_1 z) = \Sigma x_1 x_2 + \Sigma b x_1 x_3 = n\left(r_{12}\sigma_1\sigma_2 + br_{13}\sigma_1\sigma_3\right)$

and $\Sigma z^2 = n\left(\sigma_2{}^2 + b^2\sigma_3{}^2 + 2br_{23}\sigma_2\sigma_3\right)$

Hence $\sqrt{x_1 z} = \dfrac{r_{12}\sigma_2 + br_{13}\sigma_2}{\sqrt{\sigma_2{}^2 + b^2\sigma_3{}^2 + 2br_{23}\sigma_2\sigma_3}}$

To find the value of b for which this is a maximum, differentiate with respect to b and equate to zero; then

$$b = \frac{\left(r_{13} - r_{12}r_{23}\right)\sigma_2}{\left(r_{12} - r_{13}r_{23}\right)\sigma_3}$$

which gives the maximum value

$$\sqrt{x_1 z} = \sqrt{\frac{r_{12}{}^2 + r_{13}{}^2 - 2r_{12}r_{23}r_{31}}{1 - r_{23}{}^2}}$$

Computing $\sqrt{x_1 z}$ from the data of Indian production, price, and exports of wheat the value 0.905 is obtained.

Mr. G. U. Yule, in the paper already referred to,* has worked out the general solution of the problem of the correlation between three variables. In the course of the solution the problem just considered is solved incidentally. The argument is similar to that used in the case of two variables and so it will not be repeated here. A concrete notion of the results secured by Mr. Yule can be obtained from the following explanation taken from Mr. Hooker's article on the "Correlation of the Weather and the Crops."

"I have in the first place formed the ordinary coefficient $r = \dfrac{\Sigma(xy)}{\sqrt{n\sigma_1\sigma_2}}$ between the crop and (a) rainfall, (b) accumulated temperature above 42°. But rainfall and temperature are themselves correlated; hence an apparent influence of, say, rainfall upon a crop may really be due to rainfall conditions being dependent upon temperature, or *vice versa.* Hence it seemed desirable to calculate the *partial* or *net* correlation coefficients, *i.e.* (following the notation given in Mr. Yule's paper of 1897).

$$\rho_{12} = \frac{r_{12} - r_{12}r_{23}}{\sqrt{\left(1 - r_{13}{}^2\right)\left(1 - r_{23}{}^2\right)}}, \quad \rho_{13} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1 - r_{23}{}^2\right)\left(1 - r_{12}{}^2\right)}}$$

"This partial coefficient (ρ) may be regarded as a truer indication of the connection between the crop and each factor alone, inasmuch as, speaking approximately, we may say that the effect of the other factor is eliminated. It may be observed, moreover, that the relative influence of rainfall and temperature upon the crop is given by $\dfrac{\rho_{12}}{\rho_{13}}$; or, more accurately, this fraction measures the relative

Notes

effect of changes equal in amount to their respective standard deviations in the rainfall and temperature. In discussing the figures in the tables I shall accordingly utilize the partial correlation coefficients rather than the others. Finally, I have worked out what Mr. Yule calls the coefficient of double correlation between the crop and rainfall and accumulated temperature above 42°,

$$R = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}\cdot r_{23}\cdot r_{13}}{\left(1 - r_{23}^2\right)}},$$

or as it may also be written,

$$R = \sqrt{1 - \left(1 - r_{12}^2\right)\left(1 - \rho_{13}^2\right)},$$

a form which is quicker to calculate. This may be regarded as a measure of the joint influence of the rainfall and the temperature upon the crop. For the sake of brevity, I shall speak of R as measuring the effect of the 'weather,' using this term in the strictly limited sense of consisting only of these two factors. . . .

"I propose to regard a coefficient between 0.3 and 0.5 as *suggestive* of dependence. Values below 0.3 I shall, as a rule, ignore, in the absence of any corroborative evidence. Perhaps I may remark that I believe that some statisticians would consider themselves justified in drawing deductions from lower coefficients than those I have adopted as my limits."*

Mr. Yule notes that the partial or net correlation coefficient retains three of the chief properties of the ordinary coefficients: " (1) it can only be zero if both net regressions are zero; (2) it is a symmetrical function of the variables (*i.e.*, $\rho_{12} = \rho_{21}$); (3) it cannot be greater than unity."

The various illustrations which have been cited show the importance of questions of correlation in economics. The ordinary graphic method of measuring correlation is inadequate. The coefficient of correlation is simple and yet is sensitive to small changes. It has been used in many fields of statistics by Galton, Pearson, Yule, Hooker, Elderton and others. The experience of these writers warrants the adoption of the coefficient of correlation by economists as one of their standard averages.

Self-Assessment

1. Fill in the blanks:

(*i*) If more than one items is assigned the same rank ..... adjustment is made.

(*ii*) Probable error for coefficient of correlation can be found by the formula ..... .

(*iii*) Coefficient of determination is ..... .

(*iv*) Limits of correlation is ..... to ..... .

(*v*) The relationship between three or more variables is studied with the help of ..... correlation.

9.3 Summary

- Correlation means a relation between two groups. In statistics, it is the measure to indicate the relationship between two variables in which, with changes in the values of one variable, the values of other variable also change. These variables may be related to one item or may not be related to one item but have dependence on the other due to some reason.
- The term correlation indicates the relationship between two variables in which with changes in the value of one variable, the values of the other variable also change. Correlation has been defined by various eminent statisticians, mathematicians and economists.
- Correlation is very useful in understanding the economic behaviour. It helps in locating those variables on which other variables depend. In this way various economic events can be analysed. Moreover, it also helps in identifying the stabilishing factors for a disturbed economic situation.
- Correlation measures a degree of the relationship between two or more variables but it does not indicate any kind of cause and effect relationship between the variables. If, high degree of

Notes

correlation is found exist between two variables, it implies that there must be a reason for such close relationship, but the cause and effect relation can be revealed specifically when other knowledge of the factor involved being brought to bear on the situation. This means, to establish a 'functional relationship' between two or more variables, one has to go beyond the confines of statistical analysis to other factors. (Functional relationship means that two or more factors are interdependent.

- When the values of the two variables move in the same direction, *i.e.,* an increase in one is associated with an increase in other, or *vice versa,* the correlation is said to be positive. If the values of two variables move in the opposite directions *i.e.,* an increase in the value of one variable is associated with fall in other, or *vice versa,* the correlation is said to be negative. For example, the price and supply are positively correlated but price and demand are negatively correlated.
- When relation between two variables is studied, it is simple correlation. When three or more factors are studied together to find relationships, it is called multiple correlation. In partial correlation, two or more factors are agreed to be involved but correlation is studied between only two factors, considering other factors to be constant.
- The cause and effect relation existing between economic events is especially difficult to ascertain because of the presence of innumerable variable elements. In solving his problems the economist can not, like the physicist or chemist, eliminate all causes except one and then by experiment determine the effect of that one. Causes must be dealt with *en masse.* Since any effect is the result of many combined causes the economist is never sure that a given effect will follow a given cause. In stating an economic law he always has to postulate "other things remaining the same," with, perhaps, little appreciation of what the other things may be. It is rarely, if ever, possible for the economist to state more than "such and such a cause *tends* to produce such and such an effect." Events can only be stated to be more or less probable. He is dealing mainly, therefore, with correlation and not with simple causation.
- Just as the biologists cannot predict a man's height or color of eyes or temper or combativeness by knowing those qualities in his ancestors, so economists cannot predict that a definite call rate in Wall Street will go with a given percentage of reserves to deposits in New York banks or that a given supply of wheat will result in a definite price per bushel. But, on the other hand, just as it has been observed that there *is* a relation existing between a man's stature and the stature of his ancestors, so it has been observed that a relation *does* exist between bank reserves and call rates and between supply of wheat and its price per bushel.
- The commonly used method of measuring the amount of correlation between any two series of economic statistics is to represent the two series graphically upon the same sheet of cross-section paper and then compare the fluctuations of one series with those of the other. The quantity theory of prices has been tested in this way by Dr. E. W. Kemmerer. Dr. Kemmerer builds up the following price equation:
- In the case of the correlation of bank reserves and money in circulation, inclusive of bank reserves, Dr. Kemmerer concludes, "There can be no question but that when due allowance is made for fluctuations in business confidence, the evidence of Chart I strongly supports the contention that there exists a close relationship between the amount of money in circulation and the amount of the country's bank reserves."
- The graphic method of comparing fluctuations is well enough as a preliminary, *but does it enable anyone to tell anything of the extent of the correlation between the series of figures being considered*? Is Dr. Kemmerer warranted in deducing his conclusions from observation of the charts ? It seems to the writer that one opposing the quantity theory might draw opposite conclusions with as much (or as little) reason. *The charts do not answer the questions proposed.* The painstaking collection of statistics to test correlation is useless if there be no more reliable method to measure correlation. A numerical measure of the correlation must be found if we wish to determine the *extent* to which the fluctuations of one series synchronize with the fluctuations of another series.

Notes

- Report of the United States Bureau of Labor, "the percentage of successful strikes decreases during periods of business prosperity and increases during 'hard times.' " In the accompanying charts the per cent. of establishments in which strikes were successful is plotted, first, with the per capita exports and imports and second, with index numbers of wholesale prices. The foreign trade and the price statistics are taken as indicative of the activity of business, as indices of prosperity.
- The coefficient of correlation "serves as a measure of any statement involving two qualifying adjectives, which can be measured numerically, such as tall men have tall sons,' 'wet springs bring dry summers,' 'short hours go with high wages.' " It is not the purpose in what follows to go through the mathematical derivation of the coefficient of correlation, but to test the formula empirically in order to ascertain how it actually varies for given series of statistics and to point out some of its features.
- The correlation coefficients show that there is a very great difference in the degree of correlation of different pairs of series of statistics. The full significance of the "probable error," which is used as a measure of unreliability of any determination, cannot be developed at this point. It is sufficient to note that, "When *r* is not greater than its probable error we have no evidence that there is any correlation, for the observed phenomena might easily arise from totally unconnected causes; but, when *r* is greater than, say, six times its probable error, we may be practically certain that the phenomena are not independent of each other, for the chance that the observed results would be obtained from unconnected causes is practically zero."
- The amount of correlation indicated in each case is small—considering the number of years taken, so small that no conclusion as to the connection between the two series can be drawn. The correlation coefficient in the last instance, *i. e.,* between per cent. of successful strikes and business distrust, suggests an opposite conclusion to that indicated by the other coefficients and that of Mr. Cross. The analysis shows that the conclusion that there is negative correlation between *general* prosperity and per cent. of successful strikes is not warranted.
- The coefficient for the two series, population and bank reserves, came out to be 0.98. This high coefficient comes from the fact that the long-time variation of both series is the same. Consequently, before it is legitimate to draw any conclusions as to the meaning of a lack of correlation, or amount of correlation between two series of measurements it is necessary to ascertain the periodic and the secular variations in the two series. This correlation coefficient may be large through the correspondence of either secular or periodic variation, or both. It may be null because one variation covers up the other.
- For a stationary price the production must increase 46 million bushels per year.

  It seemed to me that if *percentage* changes in price and production were used instead of absolute changes a still closer correlation might result. The computation of ρ from such percentages, however, gave – 0.794.
- In the preceding illustrations the amount of correlation between the differences was greater than that between the original series. The method of differences has also been used by the writer for Kemmerer's statistics (considered on page 15 of this article) of (1) money in circulation, and (2) bank reserves for the period 1879 – 1904 with the result ρ = + 0.392, whereas the value of *r* is 0.98. This shows that there is a lack of correspondence of the short-time variations in these two series.
- Mr. G. U. Yule, in the paper already referred to,* has worked out the general solution of the problem of the correlation between three variables. In the course of the solution the problem just considered is solved incidentally. The argument is similar to that used in the case of two variables and so it will not be repeated here. A concrete notion of the results secured by Mr. Yule can be obtained from the following explanation taken from Mr. Hooker's article on the "Correlation of the Weather and the Crops."
- The ordinary graphic method of measuring correlation is inadequate. The coefficient of correlation is simple and yet is sensitive to small changes. It has been used in many fields of statistics by Galton, Pearson, Yule, Hooker, Elderton and others. The experience of these writers

Notes

warrants the adoption of the coefficient of correlation by economists as one of their standard averages.

9.4 Key-Words

1. Correlation : The correlation coefficient a concept from statistics is a measure of how well trends in the predicted values follow trends in past actual values. It is a measure of how well the predicted values from a forecast model "fit" with the real-life data.
2. Galton : An explorer and anthropologist, Francis Galton is known for his pioneering studies of human intelligence. He devoted the latter part of his life to eugenics, i.e. improving the physical and mental makeup of the human species by selected parenthood.

9.5 Review Questions

1. Define correlation. What is its utility ?
2. Explain the meaning of the term 'correlation'. Does it always signify cause and effect relationship?
3. Discuss the various types of correlation.
4. Describe the application of correlation for economists.
5. Explain, how correlation is a powerful statistical tool. Can it be used to establish cause and effect relationship ?

Answers: Self-Assessment

1. (*i*) $\frac{1}{12}\left(m^3 - m\right)$ (*ii*) $\frac{1-r^2}{\sqrt{N}}$ (*iii*) r^2

   (*iv*) r + P.E, to r – P.E (*v*) multiple

9.6 Further Readings



1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods – An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods – Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Pavitar Parkash Singh, Lovely Professional University

Notes

Unit 10: Correlation: Scatter Diagram Method, Karl Pearson's Coefficient of Correlation

CONTENTS

Objectives

Introduction

10.1 Scatter Diagram Method

10.2 Karl Pearson's Coefficient of Correlation

10.3 Summary

10.4 Key-Words

10.5 Review Questions

10.6 Further Readings

Objectives

After reading this unit students will be able to:

- Discuss Scatter Diagram Method.
- Explain Karl Pearson's Coefficient of Correlation.

Introduction

A scatter diagram is used to show the relationship between two kinds of data. It could be the relationship between a cause and an effect, between one cause and another, or even between one cause and two others. To understand how scatter diagrams work, consider the following example.

Suppose you have been working on the process of getting to work within a certain time period. The control chart you constructed on the process shows that, on average, it takes you 25 minutes to get to work. The process is in control. You would like to decrease this average to 20 minutes. What causes in the process affect the time it takes you to get to work ? There are many possible causes, including traffic, the speed you drive, the time you leave for work, weather conditions, etc. Suppose you have decided that the speed you drive is the most important cause. A scatter diagram can help you determine if this is true.

In this case, the scatter diagram would be showing the relationship between a "cause" and an "effect." The cause is the speed you drive and the effect is the time it takes to get to work. You can examine this cause and effect relationship by varying the speed you drive to work and measuring the time it takes to get to work. For example, on one day you might drive 40 *mph* and measure the time it takes to get to work. The next day, you might drive 50 *mph*. After collecting enough data, you can then plot the speed you drive versus the time it takes to get to work. Figure 1 is an example of a scatter diagram for this case. The cause (speed) is on the *x*-axis. The effect (time it takes to get to work) is on the *y*-axis. Each set of points is plotted on the scatter diagram.

In statistics, the Pearson product-moment correlation coefficient (*r*) is a common measure of the correlation between two variables X and Y. When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (?). When computed in a sample, it is designated by the letter "*r*" and is sometimes called "Pearson's *r*." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from + 1 to – 1. A correlation of + 1 means that there is a perfect positive linear relationship between variables. A correlation of – 1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is

Notes

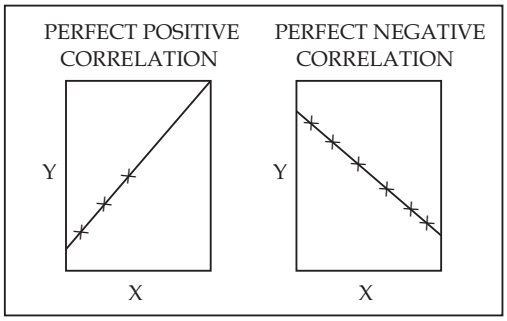
no linear relationship between the two variables. Correlations are rarely if ever 0, 1, or – 1. If you get a certain outcome it could indicate whether correlations were negative or positive.

Mathematical Formula

The quantity *r*, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of its developer Karl Pearson.

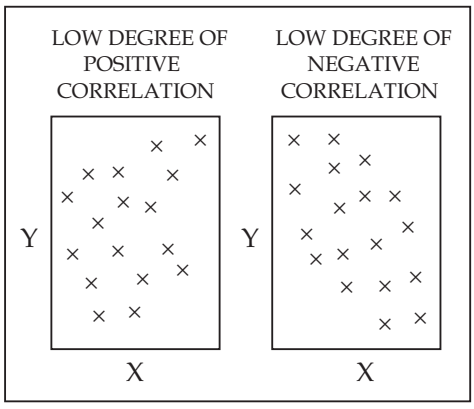
10.1 Scatter Diagram Method

The simplest device for determining relationship between two variables is a special type of dot chart called scatter diagram. When this method is used the given data are plotted on a graph paper in the form of dots, *i.e.,* for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the less relationship there is between the two variables. The more nearly the points come to falling on a line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right corner, correlation is said to be perfectly positive *(i.e., r* = + l) (diagram I).



I **II**

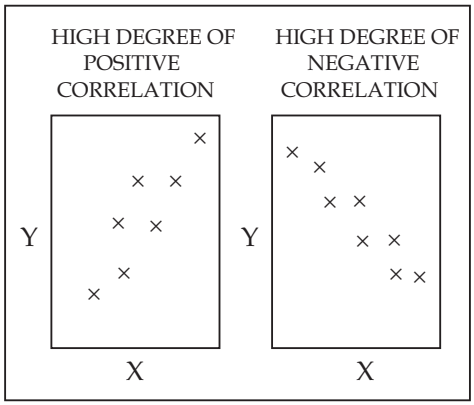
On the other hand, if all the points are lying on a straight line rising from the upper left hand corner to the lower right-hand corner of the diagram correlation is said to be perfectly negative, (*i.e.*, *r* = –1) (diagram II). If the plotted points fall in a narrow band there would be a high degree of correlation between the variables—correlation shall be positive if the points show a rising tendency from the lower left-hand corner to the upper right-hand corner (diagram III) and negative if the points show



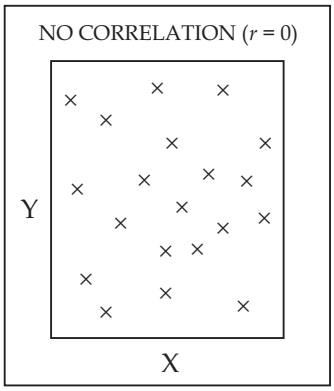
III **IV**

a declining tendency from the upper left-hand corner to the lower right-hand corner of the diagram (diagram IV). On the other hand, if the points are widely scattered over the diagram it is the indication

of very little relationship between the variables—correlation shall be positive if the points are rising from the lower left-hand corner to the upper right hand corner (diagram V) and negative if the points are running from the upper left-hand side to the lower right-hand side of the diagram (diagram VI). If the plotted points lie on a straight line parallel to the *x*-axis or in a haphazard manner, it shows the absence of any relationship between the variables, *(i.e., r* = 0) as shown by diagram VII.



V **VI**



VII

Example 1: Given the following pairs of value of the variables X and Y*:*

X:	2	3	7	6	8	9
Y*:*	6	5	5	8	12	11

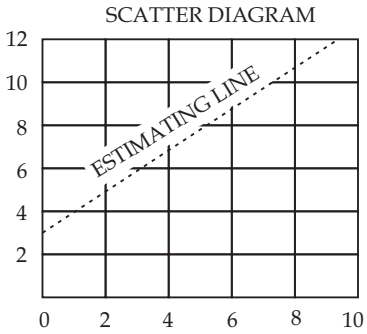
(a) Make a scatter diagram.

(b) Do you think that there is any correlation between the variables X and Y ? Is it positive or negative ? Is it high or low ?

(c) By graphic inspection, draw an estimating line.

Solution: By looking at the following scatter diagram we can say that the variables X and Y are correlated. Further, correlation is positive because the trend of the points is upward rising from the lower left-hand corner to the upper right-hand corner of the diagram. The diagram also indicates that the degree of relationship is high because the plotted points are near to the line which shows perfect relationship between the variables.

Notes



Merits and Limitations of the Method

Merits

1. It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.
2. It is not influenced by the size of extreme items whereas most of the mathematical methods of finding correlation are influenced by extreme items.
3. Making a scatter diagram usually is the first step in investigating the relationship between two variables.

Limitations

By applying this method we can get an idea about the direction of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.

10.2 Karl Pearson's Coefficient of Correlation

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The Pearsonian coefficient of correlation is denoted by the symbol *r*. It is one of the very few symbols that is used universally for describing the degree of correlation between two series. The formula for computing Pearsonian *r* is:

$$r = \frac{\sum xy}{N\sigma_x\sigma_y} \quad \text{... (i)}$$

Hence $x = \left(X - \overline{X}\right)$, $y = \left(Y - \overline{Y}\right)$

σ_x = Standard deviation of series X

σ_y = Standard deviation of series Y

N = Number of paired observations.

This method is to be applied only when the deviations of items are taken from *actual* means and *not* from assumed means.

The value of the coefficient of correlation as obtained by the above formula shall always lie between ± 1. When *r* = + 1, it means there is perfect positive correlation between the variables. When *r* = – 1, it means there is perfect negative correlation between the variables. When *r* = 0, it means there is no relationship between the two variables. However, in practice, such values of *r* as + 1, – 1 and 0 are rare. We normally get values which lie between + 1 and – 1 such as + .1, – .4, etc. The coefficient of

Notes

correlation describes not only the magnitude of correlation but also its direction. Thus, + .8 would mean that correlation is positive because the signs of r is + and the magnitude of correlation is .8.

The above formula for computing Pearsonian coefficient of correlation can be transformed in the following form which is easier to apply:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}} \quad \text{... (ii)}$$

where $x = \left(X - \bar{X}\right)$ and $y = \left(Y - \bar{Y}\right)$.

It is obvious that while applying this formula we have not to calculate separately the standard deviation of X and Y series as is necessary while applying formula (i). This simplifies greatly the task of calculating correlation coefficient.

Steps

(i) Take the deviation of X series from the mean of X and denote the deviations by x.

(ii) Square these deviations and obtain the total, *i.e.*, $\sum x^2$.

(iii) Take the deviations of Y series from the mean of Y and denote these deviations by y.

(iv) Square these deviations and obtain the total, *i.e.*, $\sum y^2$.

(v) Multiply the deviation of X and Y series and obtain the total, *i.e.*, $\sum xy$.

(vi) Substitute the values of $\sum xy$, $\sum x^2$ and $\sum y^2$ in the above formula.

The following examples will illustrate the procedure:

Example 2: Calculate Karl Pearson's coefficient of correlation from the following data:

X:	6	8	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

Solution:

Calculation of Karl Pearson's Correlation Coefficient

X	(X – 18) x	x^2	Y	(Y – 19) y	y^2	xy
6	– 12	144	10	– 9	81	+ 108
8	– 10	100	12	– 7	49	+ 70
12	– 6	36	15	– 4	16	+ 24
15	– 3	9	15	– 4	16	+ 12
18	0	0	18	– 1	1	0
20	+ 2	4	25	+ 6	36	+ 12
24	+ 6	36	22	+ 3	9	+ 18
28	+ 10	100	26	+ 7	49	+ 70
31	+ 13	169	28	+ 9	81	+ 117
$\sum X = 162$	$\sum x = 0$	$\sum x^2 = 598$	$\sum Y = 171$	$\sum y = 0$	$\sum y^2 = 338$	$\sum xy = 431$

Notes

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

$$\Sigma xy = 431, \Sigma x^2 = 598, \Sigma y^2 = 338$$

$$r = \frac{431}{\sqrt{598 \times 338}} = \frac{431}{449.582} = +0.959.$$

Example 3: Find coefficient of correlation for the following:

Cost (Rs.)	39	65	62	90	82	75	25	98	36	78
Sales (Rs.)	47	53	58	86	62	68	60	91	51	84

Solution: **Calculation of Karl Peason's Correlation Coefficient**

X	**(X – 65)** x	x^2	**Y**	**(Y – 66)** y	y^2	xy
39	– 26	676	47	– 19	361	+ 494
65	0	0	53	– 13	169	0
62	– 3	9	58	– 8	64	+ 24
90	+ 25	625	86	+ 20	400	+ 500
82	+ 17	289	62	– 4	16	– 68
75	+ 10	100	68	+ 2	4	+ 20
25	– 40	1600	60	– 6	36	+ 240
98	+ 33	1089	91	+ 25	625	+ 825
36	– 29	841	51	– 15	225	+ 435
78	+ 13	169	84	+ 18	324	+ 234
$\Sigma X = 650$	$\Sigma x = 0$	$\Sigma x^2 = 5398$	$\Sigma Y = 660$	$\Sigma y = 0$	$\Sigma y^2 = 2224$	$\Sigma xy = 2704$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{650}{10} = 65, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{660}{10} = 66$$

Since actual means of x and y are whole numbers, apply the actual mean method of finding correlation

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$r = \frac{2704}{\sqrt{5398 \times 2224}} = \frac{2704}{3464.85} = +0.78$$

We can also solve the question with the help of Logarithms. This method is easy where calculators are not allowed.

$$r = \frac{2704}{\sqrt{5398 \times 2224}}$$

Notes

$$\log r = \log 2704 - \frac{1}{2}\ [\log 5398 + \log 2224]$$

$$= 3.4320 - \frac{1}{2}\ [3.7322 + 3.3472]$$

$$= 3.4320 - 3.5397 = \bar{1}.8923 = 0.78$$

Thus the answer is the same.

Example 4: Making use of the data summarised 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	$x_1 x_2$
A	10	0	0	9	+ 1	1	0
B	6	– 4	16	4	– 4	16	16
C	9	– 1	1	6	+ 2	1	2
D	10	0	0	9	+ 1	1	0
E	12	+ 2	4	11	+ 3	+ 3	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
N = 8	$\Sigma X_1 = 80$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 32$	$\Sigma X_2 = 64$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 72$	$\Sigma x_1 x_2 = 43$

$$\bar{X}_1 = \frac{\Sigma X_1}{N} = \frac{80}{8} = 10; \qquad \bar{X}_2 = \frac{\Sigma X_2}{N} = \frac{64}{8} = 8.$$

$$r_{12} = \frac{\Sigma x_1 x_2}{\sqrt{\Sigma x_1^2 \times \Sigma x_2^2}}$$

$$\Sigma x_1 x_2 = 43, \Sigma x_1^2 = 32, \Sigma x_2^2 = 72.$$

Substituting the values

$$r_{12} = \frac{43}{\sqrt{32 \times 72}} = \frac{43}{\sqrt{2304}} = \frac{43}{48} = 0.896.$$

Note: It should be noted that the above formula is the same as given earlier, *i.e.*,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

Notes

The only difference is that of the symbols. Since in this question we were given series $\bar{X}_1$ and $\bar{X}_2$ we changed the symbols in the formula accordingly.

Calculation of Correlation Coefficient when Change of Scale and Origin is made

Since *r* is a pure number, shifting the origin and changing the scale of series do not affect its value.

Example 5: Find the coefficient of correlation from the following data:

X :	300	350	400	450	500	550	600	650	700
Y :	800	900	1000	1200	1300	1400	1500	1600	

Solution: In order to simplify calculations, let us divide each value of the variable X by 50 and each value of variable Y by 100.

CALCULATION OF CORRELATION COEFFICIENT

X	$\frac{X}{50}$ X_1	$(X_1 - \bar{X}_1)$ $\bar{X}_1 = 10$ x	x^2	Y	$\frac{Y}{100}$ Y_1	$(Y_1 - \bar{Y}_1)$ $\bar{Y}_1 = 12$ y	y^2	xy
300	6	– 4	16	800	8	– 4	16	16
350	7	– 3	9	900	9	– 3	9	9
400	8	– 2	4	1,000	10	– 2	4	4
450	9	– 1	1	1,100	11	– 1	1	1
500	10	0	0	1,200	12	0	0	0
550	11	+ 1	1	1,300	13	+ 1	1	1
600	12	+ 2	4	1,400	14	+ 2	4	4
650	13	+ 3	9	1,500	15	+ 3	9	9
700	14	+ 4	16	1,600	16	+ 4	16	16
	$\Sigma X_1 = 90$	$\Sigma x = 0$	$\Sigma x^2 = 60$		$\Sigma Y_1 = 108$	$\Sigma y = 0$	$\Sigma y^2 = 60$	$\Sigma xy = 60$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}}$$

$$\Sigma xy = 60, \Sigma x^2 = 60, \Sigma y^2 = 60$$

$$r = \frac{60}{\sqrt{60 \times 60}} = \frac{60}{60} = 1.$$

When Deviations are taken from an Assumed Mean

When actual means are in fractions, say the actual means of X and Y series are 20.167 and 29.23, the calculation of correlation by the method discussed above would involve too many calculations and would take a lot of time. In such cases we make use of the assumed mean method for finding out correlation. When deviations are taken from an assumed mean the following formula is applicable:

Notes

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_X)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

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); $\Sigma d_x d_y$ = sum of the product of the deviations of X and Y series from their assumed means; Σd_x^2 = sum of the squares of the deviations of X series from an assumed means; Σd_y^2 = sum of the squares of the deviations of Y series from an assumed mean; Σd_x = sum of the deviations of X series from an assumed mean; Σd_y = sum of the deviations of Y series from an assumed mean.

It may be pointed out that there are many variations of the above formula. For example, the above formula may be written as:

$$r = \frac{N\Sigma d_x d_y - \{(\Sigma d_x)(\Sigma d_y)\}}{\sqrt{N\ \Sigma d_x^2 - (\Sigma d_x)^2}\sqrt{N\ \Sigma d_y^2 - (\Sigma d_y)^2}}$$

But the form given above is the easiest to apply.

Note: While applying assumed mean method, any value can be taken as the assumed mean and the answer will be the same. However, the nearer the assumed mean to the actual mean, the lesser will be the calculations.

Steps

(i) Take the deviations of X series from an assumed mean, denote these deviations by d_x and obtain the total, *i.e.*, Σd_x.

(ii) Take the deviations of Y series from an assumed mean, denote these deviations by d_y and obtain the total, *i.e.*, Σd_y.

(iii) Square d_x and obtain the total Σd_x^2.

(iv) Square d_y and obtain the total Σd_y^2.

(v) Multiply d_x and d_y and obtain the total $\Sigma d_x d_y$.

(vi) Substitute the values of $\Sigma d_x d_y$, Σd_x, Σd_y, Σd_x^2 and Σd_y^2 in the formula given above.

The following examples shall illustrate the procedure:

Notes

Example 6: Compute Karl Pearson's correlation coefficient for the data given below:

X :	45	55	56	58	60	65	68	70	75	80	85
Y :	56	50	48	60	62	64	65	70	74	82	90

Solution: Since means of X and Y are in fractions we will apply the assumed mean method of calculating correlation. Taking 65 as the assumed mean in case of X and 66 in case of Y:

Calculation of Coefficient of Correlation

X	(X – 65) d_x	d_x^2	Y	(Y – 66) d_y	d_y^2	$d_x d_y$
45	– 20	400	56	– 10	100	+ 200
55	– 10	100	50	– 16	256	+ 160
56	– 9	81	48	– 18	324	+ 162
58	– 7	49	60	– 6	36	+ 42
60	– 5	25	62	– 4	16	+ 20
65	0	0	64	– 2	4	0
68	+ 3	9	65	– 1	1	– 3
70	+ 5	25	70	+ 4	16	+ 20
75	+ 10	100	74	+ 8	64	+ 80
80	+ 15	225	82	+ 16	256	+ 240
85	+ 20	400	90	+ 24	576	+ 480
$\Sigma X = 717$	$\Sigma d_x = +2$	$\Sigma d_x^2 = 1414$	$\Sigma Y = 721$	$\Sigma d_y = -5$	$\Sigma d_y^2 = 1649$	$\Sigma d_x d_y = 1401$

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$\Sigma d_x d_y = 1401, \Sigma d_x = +2, \Sigma d_y = -5, \Sigma d_x^2 = 1414, \Sigma d_y^2 = 1649$, N = 11

$$r = \frac{1401 - \frac{(2)(-5)}{11}}{\sqrt{1414 - \frac{(2)^2}{11}}\sqrt{1649 - \frac{(-5)^2}{11}}}$$

$$= \frac{1401.91}{\sqrt{1414 - .364}\sqrt{1649 - 2.273}}$$

$$= \frac{1401.91}{\sqrt{1413.636}\sqrt{1646.727}} = \frac{1401.91}{37.598 \times 40.5799} = \frac{1401.91}{1525.723} = +0.919$$

Note: We can simplify considerably the calculation by using logarithms.

Notes

Let $$r = \frac{1401.91}{\sqrt{1413.636}\sqrt{1646.727}}$$

$$\log r = \log 1401.91 - \frac{1}{2}[\log 1413.636 + \log 1646.727]$$

$$= 3.1467 - \frac{1}{2}[3.1504 + 3.2167]$$

$$= 3.1467 + \frac{1}{2}[6.3671] = 3.1467 - 3.1836 = 1.9631$$

r = AL 1.9631 = + 0.919.

Example 7: The following table gives the distribution of the total population and those who are wholly or partially blind among them. Find out if there is any relation between age and blindness.

Age	No. of persons (in thousands)	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

Solution: For facilitating comparison we must determine the number of blinds in terms of a common denominator, say 1 lakh. The first figure would remain as it is because 55 persons are blind out of 100 thousand, *i.e.*, 1 lakh. The second value would be obtained like this.

Out of 60,000 persons number of blinds = 40

Out of 1,00,000 persons number of blinds = $\frac{40}{60,000} \times 1,00,000 = 67$

and so on.

Age	Mid-points X	(X – 35)/10 d_x	d_x^2	Blind persons per lakh Y	(Y – 185) d_y	d_y^2	$d_x d_y$
0—10	5	– 3	9	55	– 130	16,900	+ 390
10—20	15	– 2	4	67	– 118	13,924	+ 236
20—30	25	– 1	1	100	– 85	7,225	+ 85
30—40	35	0	0	111	– 74	5,476	0
40—50	45	+ 1	1	150	– 35	1,225	– 35
50—60	55	+ 2	4	200	+ 15	225	+ 30

Notes

60—70	65	+ 3	9	300	+ 115	13,225	+ 345
70—80	75	+ 4	16	500	+ 315	99,225	+ 1,260
N = 8		Σd_x	Σd_x^2		Σd_y	Σd_y^2	$\Sigma d_x\, d_y$
		= 4	= 44		= +3	= 1,57,425	= 2,311

$$r = \frac{\Sigma d_x\, d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$N = 8,\ \Sigma d_x\, d_y = 2,311,\ \Sigma d_x = 4,\ \Sigma d_y = 3,\ \Sigma d_x^2 = 44,$

$\Sigma d_y^2 = 1,57,425$

Substituting these values

$$r = \frac{2311 - (4)(3)}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{157425 - \frac{(3)^2}{8}}} = \frac{2309.5}{\sqrt{42}\sqrt{157423.88}}$$

$$= \frac{2309.5}{6.4807 \times 396.77} = \frac{2309.5}{2571.34} = 0.898.$$

Correlation of Grouped Data

When the number of observations of X and Y variables is large, the data are often classified into two-way frequency distribution called a correlation table. The class intervals for Y are listed in the captions or column headings, and those for X are listed in the stubs at the left of the table (the order can also be reversed). The frequencies for each cell of the table are determined by either tallying or sorting just as in the case of a frequency distribution of a single variable.

The formula for calculating the coefficient of correlation is:

$$r = \frac{\Sigma f d_x\, d_y - \frac{(\Sigma f d_x)(\Sigma f d_y)}{N}}{\sqrt{\Sigma f d_x^2 - \frac{(\Sigma f d_x)^2}{N}}\sqrt{\Sigma f d_y^2 - \frac{(\Sigma f d_y)^2}{N}}}$$

Note: The formula is the same as the one discussed above for assumed mean. The only difference is that here the deviations are also multiplied by the frequencies.

Steps

(i) Take the step deviations of variable X and denote these deviations by d_x.

(ii) Take the step deviations of the variable Y and denote these deviations by d_y.

(iii) Multiply $d_x\, d_y$ and the respective frequency of each cell and write the figure obtained in the right hand upper corner of each cell.

(iv) Add together all the cornered values as calculated in step (iii) and obtain the total $\Sigma f_x d_y$.

(v) Multiply the frequencies of the variable X by the deviations of X and obtain the total $\Sigma f d_x$.

Notes

(vi) Take the squares of the deviations of the variable X and multiply them by the respective frequencies and obtain Σfd_x^2.

(vii) Multiply the frequencies of the variable Y by the deviations of Y and obtain the total Σfd_y.

(viii) Take the squares of the deviations of the variable Y and multiply them by the respective frequencies and obtain Σfd_y^2.

(ix) Substitute the values of $\Sigma fd_x d_y$, Σfd_x, Σfd_x^2, Σfd_y and Σfd_y^2 in the above formula and obtain the value of *r*.

Example 8: Calculate Karl Pearson's coefficient of correlation and its probable error between the ages of 100 mothers and daughters from the following data:

Age of mothers in years	Age of daughters in years					
	5–10	10–15	15–20	20–25	25–30	Total
15–25	6	3	–	–	–	9
25–35	3	16	10	–	–	29
35–45	–	10	15	7	–	32
45–55	–	–	7	10	4	21
55–65	–	–	–	4	5	9
Total	**9**	**29**	**32**	**21**	**9**	**100**

Solution: Let age of daughters be denoted by X and that of mothers by Y.

Calculation of Coefficient of Correlation

Y \ X		*m*	5–10 7.5	10–15 12.5	15–20 17.5	20–25 22.5	25–30 27.5				
		d_x d_y	– 2	– 1	0	1	2	*f*	fd_y	fd_y^2	fd_xd_y
15–25	*m* 20	– 2	6 |24	3 | 6	—	—	—	9	– 18	36	20
25–35	30	– 1	3 | 6	16 |16	10 | 0	—	—	29	– 29	29	22
35–45	40	0	—	10 | 0	15 | 0	7 | 0	—	32	0	0	0
45–55	50	1	—		7 | 0	10 |10	4 | 8	21	21	21	18
55–65	60	2	—			4 | 8	5 |20	9	18	36	28
Total		*f*	9	29	32	21	9	N = 100	Σfd_y = – 8	Σfd_y^2 = 122	Σfd_xd_y = 98
		fd_x	– 18	– 29	0	21	18	Σfd_x = – 8			
		fd_x^2	36	29	0	21	36	Σfd_x^2 = 122			
		fd_xd_y	30	22	0	18	28	Σfd_xd_y = 98			

Notes

$$r = \frac{\Sigma fd_x\, d_y - \frac{(\Sigma fd_x)(\Sigma fd_y)}{N}}{\sqrt{\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}}\sqrt{\Sigma fdy^2 - \frac{(\Sigma fd_y)^2}{N}}}$$

$\Sigma fd_x\, d_y = 98, \Sigma fd_x = -8, \Sigma fd_y = -8, \Sigma fd_x^2 = 122, \Sigma fd_y^2 = 122, N = 100$

Substituting the values in the above formula

$$r = \frac{98 - \frac{(-8)(-8)}{100}}{\sqrt{122 - \frac{(-8)^2}{100}}\sqrt{122 - \frac{(-8)^2}{100}}}$$

$$= \frac{98 - .64}{\sqrt{121.36}\,\sqrt{121.36}} = \frac{97.36}{121.36} = +0.802$$

P.E. $$r = 0.6745\frac{1 - r^2}{\sqrt{N}}$$

$r = +0.802, N = 100$

$\therefore$ $$P.E._r = 0.6745\frac{1 - (.802)^2}{\sqrt{100}} = 0.6745\frac{1 - 0.6432}{10}$$

$$= 0.6745 \times 0.03568 = 0.024.$$

Assumptions of the Pearsonian Coefficient

Karl Pearson's coefficient of correlation is based on the following assumptions:

1. There is linear relationship between the variables, *i.e.,* when the two variables are plotted on a scatter diagram straight line will be formed by the points so plotted.
2. The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.
3. There is a cause-and-effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, *i.e.,* if the variables are independent, there cannot be any correlation. For example, there is no relationship between income and height because the forces that affect these variables are not common.

Merits and Limitations of the Pearsonian Coefficient

Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is most popular. The correlation coefficient summarises in one figure not only the degree of correlation but also the direction, *i.e.,* whether correlation is positive or negative.

However, the utility of this coefficient depends in part on a wide knowledge of the meaning of this 'yardstick', together with its limitations. The chief *limitations* of the method are:

1. The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is correct or not.
2. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.

Notes

3. The value of the coefficient is unduly affected by the extreme items.
4. As compared with some other methods this method is more time consuming.

Interpreting the Coefficient of Correlation

The coefficient of correlation measures the degree of relationship between two sets of figures. As the reliability estimate depends upon the closeness of the relationship, it is imperative that utmost care is taken while interpreting the value of coefficient of correlation, otherwise fallacious conclusion may be drawn.

Unfortunately, the interpretation of the coefficient of correlation depends very much on experience. The full significance of *r* will only be grasped after working out a number of correlation problems and seeing the kinds of data that give rise to various values of *r*. The investigator must know his data thoroughly in order to avoid errors of interpretation and emphasis. He must be familiar, or become familiar, with all the relationships and theory which bear upon the data and should reach a conclusion based on logical reasoning and intelligent investigation on significantly related matters. However, the following general rules are given which would help in interpreting the value of *r*.

1. When $r = +1$ it means there is perfect positive relationship between the variables.
2. When $r = -1$ it means there is perfect negative relationship between the variables.
3. When $r = 0$ it means that there is no relationship between the variables, *i.e.,* the variables are uncorrelated.
4. 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. Beyond this it is not safe to go. The full interpretation of *r* depends upon circumstances one of which is the size of the sample. All that can really be said is that when estimating the value of one variable from the value of another, the higher the value of *r* the better the estimate.
5. The closeness of the relationship is not proportional to *r*. If the value of *r* is 0.8 it does not indicate a relationship twice as close as one of 0.4. It is in fact very much closer.

Coefficient of Correlation and Probable Error

The probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error it is possible to determine the reliability of the value of the coefficient in so far as it depends on the conditions of random sampling. The probable error of the coefficient of correlation is obtained as follows:

$$\text{P.E.} = 0.6745\,\frac{1-r^2}{\sqrt{N}}$$

where *r* is the coefficient of correlation and N the number of pairs of items.

1. If the value of *r* is less than the probable error there is no evidence of correlation, *i.e.,* the value of *r* is not at all significant.
2. If the value of *r* is more than six times the probable error, the existence of correlation is practically certain, *i.e.,* the value of *r* is significant.
3. By adding and subtracting the value of probable error from the coefficient of correlation we get respectively the upper and lower limits within which coefficient of correlation in the population can be expected to lie. Symbolically,

$$\rho = r \pm \text{P.E.}$$

ρ (rho) denotes correlation in the population.

Carrying out the computation of the probable error, assuming a coefficient of correlation of 0.80 computed from a sample of 16 pairs of items, we have

Notes

$$\text{P.E.}_r = 0.6745\frac{1-.8^2}{\sqrt{16}} = .06$$

The limits of the correlation in the population would be $r \pm \text{P.E.}$, *i.e.*, .8 ± .06 or .74—.86.

Instances are quite common wherein a correlation coefficient of 0.5 or even 0.4 has been considered to be a fairly high degree of correlation by a writer or research worker. Yet a correlation coefficient of 0.5 means that only 25 per cent of the variation is explained. A correlation coefficient of 0.4 means that only 16 per cent of the variations is explained.

Conditions for the Use of Probable Error

The measure of probable error can be properly used only when the following three conditions exist:

1. The data must approximate a normal frequency curve (bell-shaped curve).
2. The statistical measure for which the P.E. is computed must have been calculated from a sample.
3. The sample must have been selected in an unbiased manner and the individual items must be independent.

   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 statistical character.

Example 9: If r = 0.6 and N = 64, find out the probable error of the coefficient of correlation and determine the limits for population r.

Solution:

$$\text{P.E.}_r = 0.6745\frac{1-r^2}{\sqrt{N}}$$

r = 0.6 and N = 64

$$\text{P.E.}_r = 0.6745\frac{1-(.6)^2}{\sqrt{64}} = \frac{0.6745\times 0.64}{8} = 0.054$$

Limits of population correlation

$$= 0.6 \pm 0.054 = 0.546—0.654.$$

Coefficient of Determination

One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals r^2. The coefficient r^2 expresses the proportion of the variance in y determined by x; that is, the ratio of the explained variance to total variance. Therefore, the coefficient of determination expresses the proportion of the total variation that has been 'explained', or the relative reduction in variance when measured about the regression equation rather than about the mean of the dependent variable. If the value of r = 0.9, r^2 will be 0.81 and this would mean that 81 per cent of the variation in the dependent variable has been explained by the independent variable. The maximum value of r^2 is unity because it is possible to explain all of the variation in Y, but it is not possible to explain more than all of it.

It is much easier to understand the meaning of r^2 than r and, therefore, the coefficient of determination is to be preferred in presenting the results of correlation analysis. Tuttle has beautifully pointed out that "the coefficient of correlation has been grossly overrated and is used entirely too much. Its square, the coefficient of determination, is a much more useful measure of the linear covariation of two variables. The reader should develop the habit of squaring every correlation coefficient he finds cited or stated before coming to any conclusion about the extent of the linear relationship between the two correlated variables."

Notes

The relationship between r and r^2 may be noted—as the value of r decreases from its maximum value of 1, the value of r^2 decreases much more rapidly. r will, of course, always be larger than r^2, unless $r^2 = 0$ or 1.

r	r^2
0.90	0.81
0.80	0.64
0.70	0.49
0.60	0.36
0.50	0.25

Thus the coefficient of correlation is 0.707 when just half the variance in Y is due to X.

It should be clearly noted that the fact that a correlation between two variables has a value of $r = 0.60$ and the correlation between two other variables has a value of $r = 0.30$ does not demonstrate that the first correlation is twice as strong as the second. The relationship between the two given values of r can better be understood by computing the value of r^2. When $r = 0.6$, $r^2 = 0.36$ and when $r = 0.30$, $r^2 = 0.09$.

The coefficient of determination is a highly useful measure. However, it is often misinterpreted. The term itself may be misleading in that it implies that the variable X stands in a determining or causal relationship to the variable Y. The statistical evidence itself never establishes the existence of such causality. All that statistical evidence can do is to define covariation, that term being used in a perfectly neutral sense. Whether causality is present or not, and which way it runs if it is present, must be determined on the basis of evidence other than the quantitative observations.

Properties of the Coefficient of Correlation

The following are the important properties of the correlation coefficient r:

1. The coefficient of correlation lies between – 1 and + 1. Symbolically, $-1 \le r \le +1$ or $|r| \le 1$.
2. The coefficient of correlation is independent of change of scale and origin of the variables X and Y.
3. The coefficient of correlation is the geometric mean of two regression coefficients.

   Symbolically,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) There are no limits to the value of r.

(*ii*) If r is negative both the variable are decreasing.

(*iii*) If the values of X variable are 1, 2, 3, 4, 5 and those of Y 4, 6, 8, 10, 12 the Karl Pearson and the Rank method would give the same answer.

(*iv*) Pearsonian coefficient is the best under all situations.

(*v*) Karl Pearson's coefficient of correlation always lies between 0 and + 1.

10.3 Summary

- A scatter diagram is used to show the relationship between two kinds of data. It could be the relationship between a cause and an effect, between one cause and another, or even between one cause and two others.
- In statistics, the Pearson product-moment correlation coefficient (r) is a common measure of the correlation between two variables X and Y. When measured in a population the Pearson Product

Notes

Moment correlation is designated by the Greek letter rho (?). When computed in a sample, it is designated by the letter "*r*" and is sometimes called "Pearson's *r*." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from + 1 to – 1. A correlation of + 1 means that there is a perfect positive linear relationship between variables. A correlation of – 1 means that there is a perfect negative linear relationship between variables. A correlation of 0 means there is no linear relationship between the two variables. Correlations are rarely if ever 0, 1, or – 1. If you get a certain outcome it could indicate whether correlations were negative or positive.

- The simplest device for determining relationship between two variables is a special type of dot chart called scatter diagram. When this method is used the given data are plotted on a graph paper in the form of dots, *i.e.,* for each pair of X and Y values we put a dot and thus obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The more the plotted points "scatter" over a chart, the less relationship there is between the two variables. The more nearly the points come to falling on a line, the higher the degree of relationship. If all the points lie on a straight line falling from the lower left-hand corner to the upper right corner, correlation is said to be perfectly positive *(i.e., r* = + l) (diagram I).
- It is a simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.
- Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice. The Pearsonian coefficient of correlation is denoted by the symbol *r*. It is one of the very few symbols that is used universally for describing the degree of correlation between two series.
- When the number of observations of X and Y variables is large, the data are often classified into two-way frequency distribution called a correlation table. The class intervals for Y are listed in the captions or column headings, and those for X are listed in the stubs at the left of the table (the order can also be reversed). The frequencies for each cell of the table are determined by either tallying or sorting just as in the case of a frequency distribution of a single variable.
- The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc., are affected by such forces that a normal distribution is formed.
- There is a cause-and-effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, *i.e.,* if the variables are independent, there cannot be any correlation. For example, there is no relationship between income and height because the forces that affect these variables are not common.
- Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is most popular. The correlation coefficient summarises in one figure not only the degree of correlation but also the direction, *i.e.,* whether correlation is positive or negative.
- The coefficient of correlation measures the degree of relationship between two sets of figures. As the reliability estimate depends upon the closeness of the relationship, it is imperative that utmost care is taken while interpreting the value of coefficient of correlation, otherwise fallacious conclusion may be drawn.
- The probable error of the coefficient of correlation helps in interpreting its value. With the help of probable error it is possible to determine the reliability of the value of the coefficient in so far as it depends on the conditions of random sampling.
- One very convenient and useful way of interpreting the value of coefficient of correlation between

Notes

two variables is to use the square of coefficient of correlation, which is called coefficient of determination. The coefficient of determination thus equals r^2. The coefficient r^2 expresses the proportion of the variance in y determined by x; that is, the ratio of the explained variance to total variance. Therefore, the coefficient of determination expresses the proportion of the total variation that has been 'explained', or the relative reduction in variance when measured about the regression equation rather than about the mean of the dependent variable. If the value of r = 0.9, r^2 will be 0.81 and this would mean that 81 per cent of the variation in the dependent variable has been explained by the independent variable. The maximum value of r^2 is unity because it is possible to explain all of the variation in Y, but it is not possible to explain more than all of it.

- The coefficient of determination is a highly useful measure. However, it is often misinterpreted. The term itself may be misleading in that it implies that the variable X stands in a determining or causal relationship to the variable Y. The statistical evidence itself never establishes the existence of such causality. All that statistical evidence can do is to define covariation, that term being used in a perfectly neutral sense. Whether causality is present or not, and which way it runs if it is present, must be determined on the basis of evidence other than the quantitative observations.

10.4 Key-Words

1. Scatter Diagram : A scatter diagram is a tool for analyzing relationships between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis. The pattern of their intersecting points can graphically show relationship patterns. Most often a scatter diagram is used to prove or disprove cause-and-effect relationships. While the diagram shows relationships, it does not by itself prove that one variable causes the other. In addition to showing possible cause-and-effect relationships, a scatter diagram can show that two variables are from a common cause that is unknown or that one variable can be used as a surrogate for the other.
2. Coefficient determination : In statistics, the coefficient of determination, denoted R2, is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. R2 is most often seen as a number between 0 and 1.0, used to describe how well a regression line fits a set of data. An R2 near 1.0 indicates that a regression line fits the data well, while an R2 closer to 0 indicates a regression line does not fit the data very well. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.

10.5 Review Questions

1. What is Scatter diagram? How do you interprat a Scatter diagram?
2. What is a 'Scatter diagram'? How does it help us in studying the correlation between two variables in respect of both its nature and extent ?
3. How does a scatter diagram help in ascertaining the degree of correlation between two variables?
4. State the properties of Pearson's coefficient of correlation. How do you interpret a calculated value of r ? Explain the term 'Probable error of r'.
5. State the assumptions of Karl Pearson's Correlation.

Notes

Answers: Self-Assessment

1. (*i*) F (*ii*) F (*iii*) T (*iv*) F (*v*) F

10.6 Further Readings


Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 11: Rank Correlation Method

CONTENTS

Objectives
Introduction
11.1 Rank Correlation Method
11.2 Merits and Limitations of the Rank Method
11.3 Summary
11.4 Key-Words
11.5 Review Questions
11.6 Further Readings

Objectives

After reading this unit students will be able to:

- Explain Rank Correlation Method.
- Know the Merits and Limitations of Rank Method.

Introduction

When a group of individuals are arranged according to their degree of possession of a character, they are said to be ranked. The ordinal number of an individual in the arrangement is called its *rank* the arrangement as a whole is called a *ranking*.

When there are two series of ranks for the same set of individuals, corresponding to two different characters or two judges assigning ranks for the same character, one may be interested to know if the two series are associated. The association between two series of ranks for the same set of individuals is called *rank correlation.*

11.1 Rank Correlation Method

This method of finding out covariability or the lack of it between two variables was developed by the British psychologist Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group. In any event, the rank correlation coefficient is applied to a set of ordinal rank numbers, with 1 for the individual ranked first in quantity, or quality, and so on, to *n* for the individual ranked last in a group of *n* individuals (or *n* pairs of individuals). Spearman's rank correlation coefficient is defined as:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

where R denotes rank coefficient of correlation and D refers to the difference of ranks between paired items in two series.



The association between two series of ranks for the same set of individuals is called *rank correlation.*

Notes

The value of this coefficient also ranges between + 1 and – 1. When R is + 1 there is complete agreement in the order of the rank and the ranks are in the same direction. When R is – 1 there is complete agreement in the order of the ranks and they are in opposite directions.

In rank correlation we may have two types of problems:

A. Where actual ranks are given.

B. Where ranks are not given.

A. Where Actual Ranks are Given

Where actual ranks are given to us the steps required for computing rank correlation are:

(i) Take the differences of the two ranks, *i.e.*, $(R_1 - R_2)$ and denote these differences by D.

(ii) Square these differences and obtain the total $\sum D^2$.

(iii) Apply the formula:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

Example 1 : Two judges in a beauty competition rank the 12 entries as follows:

X :	1	2	3	4	5	6	7	8	9	10	11	12
Y:	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the judgment of the two judges ?

Solution:

CALCULATION OF RANK CORRELATION COEFFICIENT

X R_1	**Y** R_2	$(R_1 - R_2)$ **D**	D^2
1	12	– 11	121
2	9	– 7	49
3	6	– 3	9
4	10	– 6	36
5	3	+ 2	4
6	5	+ 1	1
7	4	+ 3	9
8	7	+ 1	1
9	8	+ 1	1
10	2	+ 8	64
11	11	0	0
12	1	+ 11	121
			$\sum D^2 = 416$

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$\sum D^2 = 416$, N = 12

$$R = 1 - \frac{6 \times 416}{12^3 - 12} = 1 - \frac{2496}{1716} = 1 - 1.454 = - 0.454.$$

Notes

Example 2: Ten competitors in a beauty contest are ranked by three judges in the following order:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	9	1	6	9
3rd Judge	6	4	9	8	1	6	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution: In order to find out which pair of judges has the nearest approach to common tastes in beauty we compare Rank Correlation between the judgments of:

(i) 1st Judge and 2nd Judge; (ii) 2nd Judge and 3rd Judge; (iii) 1st Judge and 3rd Judge.

COMPUTATION OF RANK CORRELATION

Rank by 1st Judge R_1	**Rank by 2nd Judge R_2**	**Rank by 3rd Judge R_3**	**$(R_1 - R_2)^2$ D^2**	**$(R_2 - R_3)^2$ D^2**	**$(R_1 - R_3)^2$ D^2**
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	0
4	2	3	4	1	1
9	1	10	64	81	1
7	6	5	1	1	4
8	9	7	1	4	1
N = 10	N = 10	N = 10	$\sum D^2 = 200$	$\sum D^2 = 214$	$\sum D^2 = 60$

Rank correlation between the judgments of 1st and 2nd Judge:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$\sum D^2 = 200$, N = 10

Here we have directly calculated D^2 because D's are not required in applying formula.

$$\therefore \quad R = 1 - \frac{6 \times 200}{10^3 - 10}$$

(I and II)

$$= 1 - \frac{1200}{990} = 1 - 1.212 = -0.212$$

Rank correlation between the judgments of 2nd and 3rd Judge:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

Notes

(II and III)

$$\sum D^2 = 214, N = 10$$

$$= 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{1284}{990} = 1 - 1.297 = -0.297.$$

Rank correlation between the judgments of the 1st and 3rd Judge:

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

(I and III)

$$\sum D^2 = 60, N = 10$$

$$= 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = 1 - .364 = +0.636$$

Thus we find the first and third judges have the nearest approach to common tastes in beauty.

B. Where Ranks are not Given ?

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of both the variables.

Example 3: (a) Calculate Spearman's coefficient of rank correlation for the following data:

X :	53	98	95	81	75	61	59	55
Y :	47	25	32	37	30	40	39	45

Solution:

Calculation of Rank Correlation Coefficient

X	**R_1**	**Y**	**R_2**	**$(R_1 - R_2)^2$ D^2**
53	1	47	8	49
98	8	25	1	49
95	7	32	3	16
81	6	37	4	4
75	5	30	2	9
61	4	40	6	4
59	3	39	5	4
55	2	45	7	25
				$\sum D^2 = 160$

$$R = 1 - \frac{6\sum D^2}{N^3 - N}; \sum D^2 = 160, N = 8$$

$$R = 1 - \frac{6 \times 160}{8^3 - 8} = 1 - \frac{960}{504} = 1 - 1.905 = -0.905.$$

Notes

Example 4: (b) Find the rank correlation coefficient for the following distribution:

Marks in Statistics	48	60	72	62	56	40	39	52	30
Marks in Accountancy	62	78	65	70	38	54	60	32	31

Solution: We first rank the given data:

Calculation of Rank Correlation Coefficient

Marks in Statistics	R_1	Marks is Accountancy	R_2	$(R_1 - R_2)^2$ D^2
48	4	62	6	4
60	7	78	9	4
72	9	65	7	4
62	8	70	8	0
56	6	38	3	9
40	3	54	4	1
39	2	60	5	9
52	5	32	2	9
30	1	31	1	0
				$\sum D^2 = 40$

$$R = 1 - \frac{6\sum D^2}{N^3 - N} = 1 - \frac{6 \times 40}{9^3 - 9} = 1 - \frac{240}{720} = +0.667.$$

Equal Ranks

In some cases it may be found necessary to rank two or more individuals or entries as equal. In such a case it is customary to give each individual an average rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$ that is 5.5 while if three are ranked equal at fifth place they are given the rank $\frac{5+6+7}{3}$ = 6. In other words, where two or more individuals are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have not got had they differed even slightly from each other.

Where equal ranks are assigned to some entries an adjustment in the above formula for calculating the rank coefficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}\left(m^3 - m\right)$ to the value of $\sum D^2$, where *m* stands for the number of items whose ranks are common. If there are more than one such group of items with common rank, this value is added as many times as the number of such groups. The formula can thus be written:

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + ...\right\}}{N^3 - N}$$

Notes

Example 5: (a) Calculate the coefficient of rank correlation from the following data:

X :	48	33	40	9	16	16	65	24	16	57
Y :	13	13	24	6	15	4	20	9	6	19

Solution: **Calculation of Rank Correlation Coefficient**

X	R_1	Y	R_2	$(R_1 - R_2)^2$ D^2
48	8	13	5.5	6.25
33	6	13	5.5	0.25
40	7	24	10	9.00
9	1	6	2.5	2.25
16	3	15	7	16.00
16	3	4	1	4.00
65	10	20	9	1.00
24	5	9	4	1.00
16	3	6	2.5	0.25
57	9	19	8	1.00
				$\sum D^2 = 41$

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right)\right\}}{N^3 - N}$$

Since item 6 is repeated 3 times in series X, $m = 3$. Since items 13 is repeated twice and 6 is repeated twice in case of series Y, m shall be 2 in each case. Hence:

$$R = 1 - \frac{6\left\{41 + \frac{1}{12}\left(3^3 - 3\right) + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(2^3 - 2\right)\right\}}{10^3 - 10}$$

$$= 1 - \frac{6(41 + 2 + 0.5 + 0.5)}{990} = 1 - \frac{264}{990} = 1 - 0.267 = 0.733.$$

Example 6: (b) 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

Solution:

Calculation of Rank Correlation

X	R_x	Y	R_y	$(R_x - R_y)^2$ D^2
80	8	12	1	49.00
78	7	13	2	25.00
75	5.5	14	4	2.25

Notes

75	5.5	14	4	2.25
68	4	14	4	0.00
67	3	16	7	16.00
60	2	15	6	16.00
59	1	17	8	49.00
				$\sum D^2 = 159.5$

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right)\right\}}{N^3 - N}$$

$$= 1 - \frac{6\left\{159.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)\right\}}{8^3 - 8}$$

$$= 1 - \frac{6\{159.5 + 5 + 2\}}{504}$$

$= 1 - 1.929 = - 0.929.$

11.2 Merits and Limitations of the Rank Method

Merits:

1. This method is simpler to understand and easier to apply compared to the Karl Pearson's method. The answer obtained by this method and the Karl Pearson's method will be the same provided no value is repeated, *i.e.*, all the items are different.
2. Where the data is of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and degree of correlation established by applying this method.
3. This is the only method that can be used where we are given the ranks and not the actual data.
4. Even where actual data are given, rank method can be applied for ascertaining degree of correlation.

Limitations:

1. This method cannot be used for finding out correlation in a grouped frequency distribution.
2. Where the number of items exceeds 30 the calculations become quite tedious and require a lot of time. Therefore, this method should not be applied where N is exceeding 30 unless we are given the ranks and not actual values of the variable.

When to use Rank Correlation Coefficient

The rank method has two principal uses:

(1) The initial data are in the form of ranks.

(2) If N is fairly small (say, not large than 25 or 30), rank method is sometimes applied to interval data as an approximation to the more time-consuming *r*. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30, the labour required in ranking the scores becomes greater than is justified by the anticipated saving of time through the rank formula.

Notes

Self-Assessment

1. Indicate whether the following statements are True or False:

(*i*) Rank method can be used for finding correlation coefficient even when actual data is given.

(*ii*) If two items are to be assigned equal ranks, rank method of correlation coefficient cannot be used.

(*iii*) The rank correlation coefficient was developed by spearman.

(*iv*) Rank correlation coefficient is obtained by the formula:

$R = 1 - 6\sum D^2 / N^3 - N$.

(*v*) If difference of ranks in each pair is zero, show that the rank correlation coefifcient is+ 1.

11.3 Summary

- When there are two series of ranks for the same set of individuals, corresponding to two different characters or two judges assigning ranks for the same character, one may be interested to know if the two series are associated. The association between two series of ranks for the same set of individuals is called *rank correlation.*
- This method of finding out covariability or the lack of it between two variables was developed by the British psychologist Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individuals in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group.
- The value of this coefficient also ranges between + 1 and – 1. When R is + 1 there is complete agreement in the order of the rank and the ranks are in the same direction. When R is – 1 there is complete agreement in the order of the ranks and they are in opposite directions.
- When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks can be assigned by taking either the highest value as 1 or the lowest value as 1. But whether we start with the lowest value or the highest value we must follow the same method in case of both the variables.
- If two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$ that is 5.5 while if three are ranked equal at fifth place they are given the rank $\frac{5+6+7}{3}$ = 6. In other words, where two or more individuals are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have not got had they differed even slightly from each other.
- Where the data is of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and degree of correlation established by applying this method.
- If N is fairly small (say, not large than 25 or 30), rank method is sometimes applied to interval data as an approximation to the more time-consuming *r*. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30, the labour required in ranking the scores becomes greater than is justified by the anticipated saving of time through the rank formula.

Notes

11.4 Key-Words

1. Rank correlation : In statistics, a rank correlation is the relationship between different rankings of the same set of items. A rank correlation coefficient measures the degree of similarity between two rankings, and can be used to assess its significance.

11.5 Review Questions

1. Define rank correlation. Derive formula for correlation coefficient.
2. If the difference of rank in each pair is zero, show that the rank correlation coefficient is + 1.
3. State the merits and demerits of rank correlation method.
4. Give the features of spearman's coefficient of correlation.

Answers: Self-Assessment

1. (*i*) T (*ii*) F (*iii*) T (*iv*) F (*v*) T

11.6 Further Readings



Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Notes

Unit 12 : Linear Regression Analysis : Introduction and Lines of Regression

CONTENTS

Objectives
Introduction
12.1 Introduction to Linear Regression Analysis
12.2 Line of Regression
12.3 Summary
12.4 Key-Words
12.5 Review Questions
12.6 Further Readings

Objectives

After reading this unit students will be able to :

- Introduce Linear Regression Analysis
- Discuss Line of Regression.

Introduction

In the previous unit we discussed correlation analysis, which seeks to determine the degree of linear relationship or correlation between two variables in a bivariate distribution. The coefficient of correlation indicates whether the variables are linearly related and if so, how strong the relationship is. In scatter diagram method of determining correlation, we observed that if $r = \pm 1$, then all the points lie exactly on a straight line showing a linear relationship between the two variables. Also for high positive or negative value of correlation coefficient, we observed that the point in a scatter diagram lie near about a straight line. In case $r = 0$, the scatter of points is considerable and the linear trend disappears. Such observations give rise to questions : what is the straight line in the scatter diagram, how can this line be obtained and finally what is the usefulness of this line. *Statistical methods used to answer such questions are the subject matter of* regression analysis. *The* regression analysis *is concerned with the formulation and determination of algebraic expressions for the relationship between the two variables. We use the general form 'regression lines' for these algebraic expressions. These regression lines or the exact algebraic forms of the relationship are then used for predicting the value of one variable from that of the other. Here, the variable whose value is to be predicted is called* **dependent** or **explained variable** *and the variable used for prediction is called* **independent** or **explanatory variable.**



Did u know? **Galton** *termed the line describing the average relationship between the two variables as the line of regression.*

The word regression, which means reversion, was first introduced by **Sir Francis Galton** in the study of heredity. *His study on the heights of fathers and sons revealed an interested relationship. He showed that the heights of sons tended or reverted towards the average rather than two extreme values, i.e., tall fathers tend*

Notes

to have tall sons and short fathers short sons, but the average height of the sons of a group of short fathers is greater than that of fathers. Thus, by regression we mean the average relationship between two or more variables which can be used for estimating the value of one variable from the given values of one or more variables. However, in a bivariate distribution, the analysis is restricted to only two variables only.

12.1 Introduction to Linear Regression Analysis

The study of regression has special importance in statistical analysis. We know that the mutual relationship between two series is measured with the help of correlation. Under correlation, the direction and magnitude of the relationship between two variables is measured. But it is not possible to make the best estimate of the value of a dependent variable on the basis of the given value of the independent variable by correlation analysis. Therefore, to make the best estimates and future estimation, the study of regression analysis is very important and useful.

Meaning and Definition

According to **Oxford English Dictionary,** the word 'regression' means "Stepping back" or "Returning to average value". The term was first of all used by a famous Biological Scientist in 19th century, **Sir Francis Galton** relating to a study of hereditary characteristics. He found out an interesting result by making a study of the height of about one thousand fathers and sons. His conclusion was that (i) Sons of tall fathers tend to be tall and sons of short fathers tend to be short in height (ii) But mean height of the tall fathers is greater than the mean height of the sons, whereas mean height of the short sons is greater than the mean height of the short fathers. The tendency of the entire mankind to **twin** back to average height, was termed by **Galton** 'Regression towards Mediocricity' and the line that shows such type of trend was named as 'Regression Line'.

In statistical analysis, the term 'Regression' is taken in wider sense. **Regression is the study of the nature of relationship between the variables so that one may be able to predict the unknown value of one variable for a known value of another variable.** In regression, one variable is considered as an independent variable and another variable is taken as dependent variable. With the help of regression, possible values of the dependent variable are estimated on the basis of the values of the independent variable. For example, there exists a functional relationship between demand and price, *i.e.*, D = *f* (P). Here, demand (D) is a dependent variable, and price (P) is an independent variable. On the basis of this relationship between demand and price, probable values of demand can be estimated corresponding to the different values of price.

Definition of Regression

Some important definitions of regression are as follows :

1. Regression is the measure of the average relationship between two or more variables.

   — M.M. Blair

2. Regression analysis measures the nature and extent of the relation between two or more variables, thus enables us to make predictions. **— Hirsch**

In brief, regression is a statistical method of studying the nature of relationship between two variables and to make prediction.

Utility of Regression

The study of regression is very useful and important in statistical analysis, which is clear by the following points :

(1) Nature of Relationship : Regression analysis explains the nature of relationship between two variables.

(2) Estimation of Relationship : The mutual relationship between two or more variables can be measured easily by regression analysis.

(3) Prediction : By regression analysis, the value of a dependent variable can be predicated on the basis of the value of an independent variable. For example, if price of a commodity rises, what will be the probable fall in demand, this can be predicted by regression.

Notes

(4) Useful in Economic and Business Research : Regression analysis is very useful in business and economic research. With the help of regression, business and economic policies can be formulated.

12.2 Line of Regression

If the variables in a bivariate frequency distribution are correlated, we observe that the points in a scatter diagram cluster around a straight line called the **line of regression.** In a bivariate study, we have **two lines of regression,** namely :

1. Regression of Y on X.
2. Regression of X on Y.

Regression of Y on X

The line of regression of Y on X *is used to predict or estimate the value of* Y *for the given value of the variable* X. Thus, Y is the dependent variable and X is an independent variable in this case. The *algebraic form of the line* **line of regression of Y on X** is of the form :

$$Y = a + bX \quad \ldots (1)$$

where, *a* and *b* are unknown constants to be determined by observed data on the two variables X and Y. Let $(X_1, Y_1), (X_2, Y_2)..., (X_N, Y_N)$ be N pairs of observations on the variable X and Y. *Then, for determining a and b in equation* (1) *we make use of the following* **normal equations :**

$$\Sigma Y = Na + b\Sigma X \quad \ldots (2)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \quad \ldots (3)$$

The values ΣY, ΣX, ΣX^2 and ΣXY can be obtained from the given data.

These normal equations are obtained by minimising the error sum of squares according to the principle of least squares. Solving equations (2) and (3) for *a* and *b,* the line of regression of Y on X is completely determined.

Alternatively

There is another way of finding the algebraic form of line of the regression of Y on X. Line of regression of Y on X *can also be written in the following form* :

$$\left(Y - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right) \quad \ldots (4)$$

or

$$\left(Y - \overline{Y}\right) = b_{YX}\left(X - \overline{X}\right) \quad \ldots (5)$$

Here,

$\overline{Y}$ = the mean of Y

$\overline{X}$ = the mean of X

σ_Y = the S.D. of Y

σ_X = the S.D. of X

r = the correlation coefficient between X and Y

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$ = the regression coefficient of Y on X

From observed bivariate data $[(X_i, Y_i); i = 1, 2, ... N]$ the regression coefficient of Y on X, b_{YX}, can be computed from any of the following formula :

Notes

$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \quad \ldots (6)$$

or

$$b_{YX} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2} \quad \ldots (7)$$

Here, $d_x = (X - A)$ and $d_y = (Y - B)$ are deviations from assumed means A and B respectively.

Regression of X and Y

The line of regression of X and Y *is used to estimate or predict the value of X for a given value of the variable Y*. In this case X is the dependent variable and Y is the independent variable. The standard algebraic form of the line of regression of X on Y is :

$$X = c + dY \quad \ldots (8)$$

where c and d are unknown constant which are determined from the following two normal equations:

$$\Sigma X = Nc + d\Sigma Y \quad \ldots (9)$$

$$\Sigma XY = c\Sigma Y + d\Sigma Y^2 \quad \ldots (10)$$

The values of ΣX, ΣY, ΣXY and ΣY^2 can be obtained from observed data. The normal equations (9) and (10) are also obtained by minimising error sum of squares according to the method of least squares. Solving (9) and (10) for c and d and putting these values in (8), the form of regression of X on Y is completely determined.

Alternatively

Like regression of Y on X, the line of regression of X on Y also has an alternative form as

$$(X - \bar{X}) = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \quad \ldots (11)$$

or

$$(X - \bar{X}) = b_{XY}(Y - \bar{Y}) \quad \ldots (12)$$

Here, $b_{XY} = r\frac{\sigma_X}{\sigma_Y}$ is called the regression coefficient of X on Y. For observed data, the value of b_{XY} can be computed from any of the two formulae :

$$b_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2} \quad \ldots (13)$$

or

$$b_{XY} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2} \quad \ldots (14)$$

The determination of both lines of regression will be clear from the following examples :

Example 1: From the following data, obtain the two regression equations.

X	6	2	10	4	8
Y	9	11	5	8	7

Solution: First we obtain the regression lines by using the method of least squares.

Notes

Table showing calculations

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$	$\Sigma XY = 214$

Regression of X and Y :

Let line of regression of Y on X be

$$Y = a + bX \quad \text{... (i)}$$

The normal equations giving the values of a and b are

$$\Sigma Y = Na + b\Sigma X \quad \text{... (ii)}$$

$$\Sigma XY = a\Sigma X + \Sigma bX^2 \quad \text{... (iii)}$$

Putting the values from the table, one gets

$$40 = 5a + 30b \quad \text{... (iv)}$$

$$214 = 30a + 220b \quad \text{... (v)}$$

Multiplying equation (iv) by 6, we get

$$240 = 360a + 180b \quad \text{... (vi)}$$

Subtracting (v) from (vi), we have

$$26 = -40b \therefore b = -0.65$$

Thus, from (iv), one gets

$$40 = 5a - 30 \times 0.65 \text{ or } 5a = 40 + 19.5 \therefore a = 11.9$$

Putting the values of a and b in (i), the regression of Y on X becomes

$$Y = 11.9 - 0.65\,X \text{ or } Y + 0.65\,X = 11.9$$

Regression of X on Y :

Let the line of regression of X on Y be

$$X = c + dY \quad \text{... (vii)}$$

The normal equations giving the value of c and d are

$$\Sigma X = Nc + d\Sigma Y \quad \text{... (viii)}$$

$$\Sigma XY = c\Sigma Y + d\Sigma Y^2 \quad \text{... (ix)}$$

Putting the values in (viii) and (ix), one gets

$$30 = 5c + 40d \quad \text{... (x)}$$

$$214 = 40c + 340d \quad \text{... (xi)}$$

Multiplying equation (x) by 8, we have

$$240 = 40c + 320d \quad \text{... (xii)}$$

Subtracting equation (xii) from equation (xi), one gets

Notes

$-26 = 20d \quad \therefore \quad d = -1.3$

Putting $d = -1.3$ in equation (x),

$30 = 5c + 40 \times (-1.3)$ or $5c = 30 + 52 = 82$

$\therefore \quad c = 16.4$

Putting the values of c and d in (vii), the line of regression of X on Y is :

$X = 16.4 - 1.3\,Y$

$\therefore \quad X + 1.3\,Y = 16.4$

Example 2: Obtain regression lines for the data in example 1 by computing regression coefficients.

Solution:

Regression of Y on X :

Computation of regression coefficients

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$	$\Sigma XY = 214$

The line of regression of Y on X using its regression coefficient can be written as :

$$\left(Y - \overline{Y}\right) = b_{YX}\left(X - \overline{X}\right) \quad \text{... (i)}$$

Here,
$$\overline{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

and
$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)}$$

$$= \frac{5 \times 214 - 30 \times 40}{5 \times 220 - (30)^2}$$

$$= \frac{1070 - 1200}{1100 - 900} = \frac{-130}{200} = -0.65$$

Putting the values of $\overline{X}$, $\overline{Y}$ and b_{YX} in equation (i), one gets the line of regression of Y on X as :

$(Y - 8) = -0.65\,(X - 6)$ or $Y = 8 - 0.65\,X + 3.9$

or $Y + 0.65\,X = 11.9$

which is the same as obtained in example 1.

Notes

Regression of X on Y :

The line of regression of X on Y using its regression coefficient is :

$$\left(X - \overline{X}\right) = b_{YX}\left(Y - \overline{Y}\right) \quad \text{... (ii)}$$

Here,
$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{5 \times 214 - 30 \times 40}{5 \times 340 - (40)^2} = \frac{1070 - 1200}{1700 - 1600} = -\frac{130}{100} = -1.30$$

Putting the value of $\overline{X}$, $\overline{Y}$ and b_{YX} in equation (ii), the line of regression of X and Y becomes

$$(X - 6) = -1.30\ (Y - 8) \text{ or } X = 6 - 1.30\ Y + 10.4$$

or
$$X + 1.30\ Y = 16.4$$

which is also the same as obtained in example 1.

Remark

The two expression lines can be obtained by any of the above two methods if the values of the pairs of observations are not very large. *However, when the data are large, short cut method for computing regression coefficient should be applied to avoid huge calculations.* Calculations are further reduced if the deviations of the two variables are taken from their respective means, *i.e.,* when Σd_x and Σd_y are zero. *In this case, the regression coefficients* b_{YX} *and* b_{XY} *are obtained by using formula* (7) *and* (14). The example will on the next page clarify the point.

Example 3: On the basis of following data, obtain regressions of (i) Y on X and (ii) X on Y.

X	15	27	27	30	38	46
Y	12	15	15	18	22	26

Solution:

Calculation for regression equations

X	d_x **= (X – 30)**	d_x^2	**Y**	d_y **= (Y – 18)**	d_y^2	$d_x d_y$
15	– 15	225	12	– 6	36	90
27	– 3	9	15	– 3	9	9
27	– 3	9	15	– 3	9	9
30	0	0	18	0	0	0
38	8	64	22	4	16	32
46	16	256	26	8	64	128
N = 6	$\Sigma d_x = 3$	$\Sigma d_x^2 = 563$		$\Sigma d_y = 0$	$\Sigma d_y^2 = 134$	$\Sigma d_x d_y = 228$

Thus,
$$\overline{X} = A + \frac{\Sigma dX}{N}\ ; \qquad \overline{Y} = B + \frac{\Sigma dY}{N}$$

Notes

$$= 30 + \frac{3}{6} = 30.5 \quad = 18 + \frac{0}{6} = 18.0$$

$$b_{YX} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{6 \times 268 - 3 \times 0}{6 \times 563 - (3)^2} = \frac{1608}{3378 - 9} = \frac{1608}{3369} = 0.4773$$

$$b_{YX} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$= \frac{6 \times 268 - 3 \times 0}{6 \times 134 - (0)^2} = \frac{1608}{804} = 2.00$$

Therefore, the regression of Y on X is :

$$(Y - \overline{Y}) = b_{YX}(X - \overline{X}) \text{ or } (Y - 18) = 0.4773\ (X - 30.5)$$

or $Y = 18 + 0.4773X - 14.5577$

or $Y = 0.4773X + 3.4424$

and regression of X on Y is :

$$(X - \overline{X}) = b_{YX}(Y - \overline{Y}) \text{ or } (X - 30.5) = 2.0\ (Y - 18)$$

or $X = 30.5 + 2.0Y - 36.0$

or $X = 2.0Y - 5.50$

Example 4: In the estimation of regression equations of two variables X and Y, the following results were obtained :

$$\overline{X} = 20,\ \overline{Y} = 30,\ N = 10,\ \Sigma X^2 = 6360,\ \Sigma Y^2 = 9860,\ \Sigma XY = 5900$$

obtain the two equations.

Solution: Given that : $\overline{X} = 20,\ \overline{Y} = 30,\ N = 10,\ \Sigma X^2 = 6360,\ \Sigma Y^2 = 9860,\ \Sigma XY = 5900$

$\therefore \quad \Sigma X = N\overline{X} = 10 \times 20 = 200;\ \Sigma Y = 10 \times 30 = 300$

$\therefore \quad b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$

$$= \frac{10 \times 5900 - 200 \times 300}{10 \times 6360 - (200)^2} = \frac{59000 - 60000}{63600 - 40000}$$

$$= \frac{-1000}{23600} = -0.042$$

and $b_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$

Notes

$$= \frac{10 \times 5900 - 60000}{10 \times 9860 - (300)^2} = \frac{-1000}{8600} = -0.116$$

Regression of Y on X

$$(Y - \bar{Y}) = b_{YX}(X - \bar{X})$$

or $(Y - 30) = -0.042\,(X - 20)$ or $Y = 30 - 0.042X + 0.84$

or $Y + 0.042X = 30.84$

Regression of X on Y

$$(X - \bar{X}) = b_{XY}(Y - \bar{Y})$$

or $(X - 20) = -0.116\,(Y - 30)$ or $X = 20 - 0.116Y + 3.48$

or $X + 0.116Y = 23.48$



Did u know? *Statistical methods used to answer such questions are the subject matter of* regression analysis. *The* regression analysis *is concerned with the formulation and determination of algebraic expressions for the relationship between the two variables.*

Properties of the Regression Lines

1. The regression lines of Y on X is used to estimate or predict the best value (in least squares sense) of Y for a given value of the variable X. Here Y is dependent and X is an independent variable.
2. The regression line of X on Y is used to estimate to best value of X for a given value of the variable Y. Here X is dependent and Y is an independent variable.
3. The two lines of regression cut each other at the points $(\bar{X}, \bar{Y})$. *Thus, on solving the two lines of regression, we get the values of means of the variables in the bivariate distribution.*
4. In a bivariate study, there are two lines of regression. *However, in case of perfect correlation that is when r* = + 1 on – 1 *we have only one regression line as both the regression lines coincide in this case.*
5. When *r* = 0, *i.e.*, if correlation exists between X and Y, the two lines of regression become perpendicular to each other.

Self-Assessment

1. Which of the following statements is true or false :

(*i*) The term 'regression' was first used by Karl Pearson in the year 1900.

(*ii*) Regression analysis reveals average relationship between two variables.

(*iii*) The regression line cut each other at the point of average of X and Y.

(*iv*) In regression analysis b_{xy} stand for regression coefficient of X on Y.

(*v*) The regression line of Y on X minimises total of the squares of the vertical deviations.

12.3 Summary

- *We use the general form 'regression lines' for these algebraic expressions. These regression lines or the exact algebraic forms of the relationship are then used for predicting the value of one variable from that of the other. Here, the variable whose value is to be predicted is called* **dependent** or **explained variable** *and the variable used for prediction is called* **independent** or **explanatory variable.**

Notes

- The word regression, which means reversion, was first introduced by **Sir Francis Galton** in the study of heredity. *His study on the heights of fathers and sons revealed an interested relationship. He showed that the heights of sons tended or reverted towards the average rather than two extreme values, i.e., tall fathers tend to have tall sons and short fathers short sons, but the average height of the sons of a group of short fathers is greater than that of fathers.* **Galton** *termed the line describing the average relationship between the two variables as the line of regression. Thus, by regression we mean the average relationship between two or more variables which can be used for estimating the value of one variable from the given values of one or more variables.*
- Under correlation, the direction and magnitude of the relationship between two variables is measured. But it is not possible to make the best estimate of the value of a dependent variable on the basis of the given value of the independent variable by correlation analysis. Therefore, to make the best estimates and future estimation, the study of regression analysis is very important and useful.
- In statistical analysis, the term 'Regression' is taken in wider sense. **Regression is the study of the nature of relationship between the variables so that one may be able to predict the unknown value of one variable for a known value of another variable.** In regression, one variable is considered as an independent variable and another variable is taken as dependent variable. With the help of regression, possible values of the dependent variable are estimated on the basis of the values of the independent variable. For example, there exists a functional relationship between demand and price, *i.e.*, D = *f* (P). Here, demand (D) is a dependent variable, and price (P) is an independent variable. On the basis of this relationship between demand and price, probable values of demand can be estimated corresponding to the different values of price.
- By regression analysis, the value of a dependent variable can be predicated on the basis of the value of an independent variable. For example, if price of a commodity rises, what will be the probable fall in demand, this can be predicted by regression.
- If the variables in a bivariate frequency distribution are correlated, we observe that the points in a scatter diagram cluster around a straight line called the **line of regression.**
- The line of regression of X and Y *is used to estimate or predict the value of X for a given value of the variable Y*. In this case X is the dependent variable and Y is the independent variable.

12.4 Key-Words

1. Linear Regression Analysis : In statistics, regression analysis is a statistical technique for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables - that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function, which can be described by a probability distribution.

Notes

12.5 Review Questions

1. What is meant by linear regression analysis ? Explain clearly the significance of linear regression analysis.
2. What are regression line ? Explain their uses.
3. Discuss the introduction of linear regression analysis.
4. What are the properties of regression line.
5. Explain the concept of lines of regression. State why there are two lines of regression.

Answers: Self-Assessment

1. (*i*) F (*ii*) T (*iii*) T (*iv*) F (*v*) T

12.6 Further Readings



Books

1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.

Unit 13: Coefficient of Simple Regression Method

Notes

CONTENTS

Objectives
Introduction
13.1 Regression Equations
13.2 Coefficient of Simple Regression Method
13.3 Summary
13.4 Key-Words
13.5 Review Questions
13.6 Further Readings

Objectives

After reading this unit students will be able to:

- Explain Regression Equations
- Discuss the coefficients of Simple Regression Method.

Introduction

After having established the fact that two variables are closely related, we may be interested in estimating (predicting) the value of one variable given the value of another. For example, if we know that advertising and sales are correlated, we may find out the expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales. Similarly, if we know that the yield of rice and rainfall are closely related, we may find out the amount of rain required to achieve a certain production figure. The *statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called* **regression.** With the help of regression analysis,* we are in a position to find out the *average probable* change in one variable given a certain amount of change in another.

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables, price (X) and demand (Y), are closely related, we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of commodity are closely related, we can find out the expected price for a certain amount of tax levy. Thus we find that the study of regression is of considerable help to the economists and businessmen.

13.1 Regression Equations

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations–the regression of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression Equation of Y on X

The regression equation of Y on X is expressed as follows:

$$Y_c = a + bX$$

Notes

In this equation *a* and *b* are two unknown constants (fixed numerical values) which determine the position of the line completely. These constants are called the parameters of the line. If the value of either or both of them is changed, another line is determined. The parameter '*a*' determines the *level* of the fitted line (*i.e.,* the distance of the line directly above or below the origin). The parameter '*b*' determines the *slope* of the line, *i.e.*, the change in Y per unit change in X. The symbol Y_c stands for the value of X computed from the relation for a given X.

If the values of the constants '*a*' and '*b*' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of Least Squares which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the deviation of the actual Y values from the computed Y values is the least or, in other words, in order to obtain a line which fits the points best $\sum (Y - Y_c)^2$ should be minimum. Such a line is known as the line of 'best fit'.

With a little algebra and differential calculus it can be shown that the following two equations, if solved simultaneously, will yield values of the parameters *a* and *b* such that the least squares requirement is fulfilled:

$$\sum Y = Na + b \sum X$$

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

These equations are usually called the *normal equations*. In these equations $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ indicate totals which are computed from the observed pairs of values of two variables X and Y to which the least squares estimating line is to be fitted and N is the number of observed pairs of values.



Notes The dictionary meaning of the term 'regression' is that act of returning or going back. The term 'regression' was first used by Francis Galton towards the end of nineteenth century while studying the relationship between the height of fathers and sons. This term was introduced by him in the paper 'Regression towards Mediocrity in Hereditary Stature'. His study of height of about one thousand fathers and sons revealed a very interesting relationship, *i.e.,* tall fathers tend to have tall sons and short fathers short sons, but 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. The line describing the tendency to regress or going back was called by Galton a 'Regression Line'. The term is still used to describe that line drawn for a group of points to represents the trend present, but it no longer necessarily carries the original implication that Galton intended. These days is growing tendency of the modern writers to use the term *estimating line* instead of *regression line* because the expression estimating line is more clarificatory in character.

Regression Equation of X on Y

The regression equation of X on Y is expressed as follows:

$$X_c = a + bY$$

To determine the value of *a* and *b* the following two normal equations are to be solved simultaneously:

$$\sum X = Na + b \sum Y$$

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

Example 1: From the following data obtain the two regression equations:

Notes

X	6	2	10	4	8
Y	9	11	5	8	7

Solution: **Obtaining Regression equations**

X	**Y**	**XY**	**X^2**	**Y^2**
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\sum X = 30$	$\sum Y = 40$	$\sum XY = 214$	$\sum X^2 = 220$	$\sum Y^2 = 340$

Regression Equation of Y on X

$$Y_c = a + bX$$

To determine the value of a and b the following two normal equations are to be solved:

$$\sum Y = Na + b\sum X$$

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

Substituting the values,

$$40 = 5a + 30b \quad \text{... (i)}$$

$$214 = 30a + 220b \quad \text{... (ii)}$$

Multiplying Eqn. (i) by 6

$$240 = 30a + 180b \quad \text{... (iii)}$$

$$214 = 30a + 220b \quad \text{... (iv)}$$

Subtracting Eqn. (iv) from (iii)

$$-40b = +26$$

$$b = -0.65$$

Substituting the value of b in Eqn. (i)

$$40 = 5a + 30(-0.65)$$

$$5a = 40 + 19.5 = 59.5$$

$$a = 11.9.$$

Putting the values of a and b in the equation, the regression line of Y on X is

$$Y = 11.9 - 0.65X$$

Regression Line of X on Y

$$X_c = a + bY$$

and the two normal equations are:

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2 \quad \text{... (i)}$$

$$30 = 5a + 40b \quad \text{... (ii)}$$

$$214 = 40a + 34b$$

Multiplying Eqn. (i) by 8

$$240 = 40a + 320b \quad \text{... (iii)}$$

Notes

$$214 = 40a + 340b \quad \text{... (iv)}$$

Deducting Eqn. (iv) from (iii)

$$-20b = 26$$

$$\therefore \quad b = -1.3$$

Substituting the value of b in Eqn. (i)

$$30 = 5a + 40(-1.3)$$

$$5a = 30 + 52 = 82$$

$$a = 16.4$$

Putting the values of a and b in the equation, the regression line of X on Y is

$$X = 16.4 - 1.3\,Y.$$

Deviations taken from Arithmetic Means of X and Y

The calculations can be very much simplified if instead of dealing with the actual values of X and Y we take the deviations of X and Y series from their respective means. In such a case the equation $Y_c = a + bX$ is changed to

$$Y - \bar{Y} = b\left(X - \bar{X}\right)$$

or simply $y = bx$

where $y = \left(Y - \bar{Y}\right)$ and $x = \left(X - \bar{X}\right)$

The value of b can be easily obtained as follows:

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

The two normal equations which we had written earlier when changed in terms of x and y become

$$\Sigma y = Na + b\Sigma x \quad \text{... (i)}$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \text{... (ii)}$$

Since $\Sigma x = \Sigma y = 0$ (deviations being taken from means)

Equation (i) reduces to $Na = 0 \quad \therefore \quad a = 0$

Equation (ii) reduces to $\Sigma xy = b\Sigma x^2 \quad \therefore \quad b = \frac{\Sigma xy}{\Sigma x^2}$

After obtaining the value of b the regression equation can easily be written in terms of X and Y by substituting for y, $\left(Y - \bar{Y}\right)$ and for x, $\left(X - \bar{X}\right)$.

Similarly the regression equation $X_c = a + b\,Y$ is reduced to $x = 0$ and the value of b is obtained as follows:

$$b = \frac{\Sigma xy}{\Sigma y^2}$$

Example 2: From the following data obtain the regression equation of X on Y, and also that of Y on X:

X	6	2	10	4	8
Y	9	11	5	8	7

Solution: **Notes**

Calculation of Regression Equations

X	(X – 6) x	x^2	Y	(Y – 8) y	y^2	xy
6	0	0	9	+ 1	1	0
2	– 4	16	11	+ 3	9	– 12
10	+ 4	16	5	– 3	9	– 12
4	– 2	4	8	0	0	0
8	+ 2	4	7	– 1	1	– 2
$\Sigma X = 30$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

Regression Equation of X on Y

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

$$\bar{X} = \frac{30}{5} = 6,\ \bar{Y} = \frac{40}{5} = 8,\ r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

$$X - 6 = -1.3\,(Y - 8)$$

$$X - 6 = -1.3\,Y + 10.4 \text{ or } X = 16.4 - 1.3\,Y$$

Regression Equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma x}\left(X - \bar{X}\right)$$

$$r\frac{\sigma_y}{\sigma x} = \frac{\Sigma xy}{\Sigma x^2} = \frac{-26}{40} = -0.65$$

$$Y - 8 = -0.65\,(X - 6)$$

$$Y - 8 = -0.65\,X + 3.9$$

$$Y = 11.9 - 0.65\,X.$$

Deviations taken from Assumed Means

When actual means of X and Y variables are in fractions, the calculations can be simplified by taking the deviations from the assumed mean. The value of *b*, *i.e.*, the regression coefficient, will be calculated as follows:

Regression Equation of X and Y

$$X - \bar{X} = b_{xy}\left(Y - \bar{Y}\right)$$

$$\sigma_{xy} = \frac{N\,\Sigma\, d_x d_y - \left(\Sigma\, d_x\right)\left(\Sigma\, d_y\right)}{N\Sigma\, d_y^2 - \left(\Sigma\, d_y\right)^2}$$

Notes

Regression Equation of Y on X

$$Y - \bar{Y} = b_{yx}\left(X - \bar{X}\right)$$

$$b_{yx} = \frac{N \Sigma d_x d_y - \left(\Sigma d_x\right)\left(\Sigma d_y\right)}{N \Sigma d_x^2 - \left(\Sigma d_x\right)^2}$$

Once the value of b_{xy} and b_{yx} are determined in the above manner the regression equations can be obtained very easily.

Example 3: From the data of Example 1, obtain regression equations taking deviations from 5 in case of X and 7 in case of Y.

Solution:

X	$(X - 5)$		Y	$(Y - 7)$		
d_x	d_x^2			d_y	d_y^2	$d_x d_y$
6	+ 1	1	9	+ 2	4	+ 2
2	– 3	9	11	+ 4	16	– 12
10	+ 5	25	5	– 2	4	– 10
4	– 1	1	8	+ 1	1	– 1
8	+ 3	9	7	0	0	0
$\Sigma X = 30$	$\Sigma d_x = +5$	$\Sigma d_x^2 = 45$	$\Sigma Y = 40$	$\Sigma d_y = 5$	$\Sigma d_y^2 = 25$	$\Sigma d_x d_y = -21$

Regression Equation of X on Y

$$X - \bar{X} = b_{xy}\left(Y - \bar{Y}\right)$$

$$b_{xy} = \frac{N \Sigma d_x d_y - \left(\Sigma d_x\right)\left(\Sigma d_y\right)}{N \Sigma d_y^2 - \left(\Sigma d_y\right)^2}$$

$$= \frac{(5)(-21) - (5)(5)}{(5)(25) - (5)^2} = \frac{-105 - 25}{125 - 25} = -1.3$$

$$\bar{X} = \frac{30}{5} = 6, \ \bar{Y} = \frac{40}{5} = 8$$

Hence the regression equation is

$$X - 6 = -1.3\,Y\,(-8)$$

$$X - 6 = -1.3\,Y + 10.4$$

or

$$X = 16.4 - 1.3\,Y$$

Regression Equation of Y on X

$$Y - \bar{Y} = b_{yx}\left(X - \bar{X}\right)$$

Notes

$$b_{yx} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{(5)(-21) - (5)(5)}{(5)(45) - (5)^2} = \frac{-105 - 25}{225 - 25} = \frac{-130}{200} = -0$$

So the regression equation is

$$Y - 8 = -0.65 (X - 6)$$
$$Y - 8 = -0.65 X + 3.9$$
$$Y = 11.9 - 0.65 X.$$

13.2 Coefficients of Simple Regression Method

The quantity b in the regression equation is called the regression coefficient. Since there are two regression equations there are also two regression coefficients–regression coefficient of X on Y and regression coefficient of Y on X.

Regression Coefficient of X on Y

The regression coefficient of X and Y is represented by the symbol b_{xy} or b_1. It measure the change in X corresponding to a unit change in Y. The regression coefficient of X on Y is given by

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

where deviations are taken from means of X and Y, the regression coefficient is obtained by:

$$b_{xy} * = \frac{\Sigma xy}{\Sigma y^2}$$

$$* r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}},\ \sigma_x = \sqrt{\frac{\Sigma x^2}{N}} \text{ and } \sigma_y = \sqrt{\frac{\Sigma y^2}{N}}$$

Substituting these values

$$b_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \times \frac{\sqrt{\frac{\Sigma y^2}{N}}}{\sqrt{\frac{\Sigma x^2}{N}}} = \frac{\Sigma xy}{\Sigma y^2}$$

where deviations are taken from assumed means, the value of b_{xy} is obtained as follows:

$$b_{xy} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

Regression Coefficient of Y on X

The regression coefficient of Y on X is represented by b_{yx} or b_2. It measures the change in Y corresponding to unit change in X. The value of b_{yx} is given by

Notes

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

When deviations are taken from actual means of X and Y

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

When deviations are taken from assumed means of X and Y

$$b_{yx} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

Calculating Correlation from Regression Coefficients

It should be interesting to note that the underroot of the product of the two regression coefficients gives us the value of the coefficient of correlation. Symbolically:

$$r = \sqrt{b_1 \times b_2} \text{ or } \sqrt{b_{xy} \times b_{yx}}$$

Proof:

$$b_1 \text{ or } b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$b_2 \text{ or } b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_1 \times b_2 = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

$$\therefore \quad r = \sqrt{b_1 \times b_2}$$

Since the value of the coefficients of correlation (r) cannot exceed one, one of the regression coefficients must be less than one, or in other words, both the regression coefficients cannot be greater than one. For example, if $b_{yx} = 1.2$ and $b_{xy} = 1.4$, r would be $\sqrt{1.2 \times 1.4} = 1.29$ which is not possible. Further, the regression coefficient which may exceed one should also be such in value that when multiplied by the other coefficient the underroot of the product of the two coefficients does not exceed one. Also both the regression coefficients will have the same sign, *i.e.,* they will be either positive or negative. The coefficient of correlation (r) will have the same sign as that of regression coefficients, *i.e.,* if regression coefficients have a negative sign, r will also have negative sign and if regression coefficients have a positive sign, r will also have positive sign.

For example, if $b_{xy} = -0.8$ and $b_{yx} = -1.2$, r would be

$$\sqrt{-0.8 \times -1.2} = -0.98$$

Since

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

we can find out any of the four values, given the other three. For example, if we know that $r = 0.6$, $\sigma_x = 4$ and $\sigma_{xy} = 0.8$, we can find σ_y

Notes

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

$$0.8 = \frac{0.6 \times 4}{\sigma_y}$$

$$\sigma_y = \frac{2.4}{0.8} = 3.$$

Example 4: Obtain the value of the correlation coefficient through the method of regression analysis from the data given below first by taking deviation from the actual means of X and Y and secondly from assumed means 2 and 18 for series X and Y respectively.

X	1	2	3	4	5
Y	10	20	15	25	30

Solution:

(a) Calculation of Regression Coefficient from the actual means

X	(X – 3) x	x^2	Y	(Y – 20) y	y^2	xy
1	– 2	4	10	– 10	100	20
2	– 1	1	20	0	0	0
3	0	0	15	– 5	25	0
4	+ 1	1	25	+ 5	25	5
5	+ 2	4	30	+ 10	100	20
$\Sigma X = 15$	$\Sigma x = 0$	$\Sigma x^2 = 10$	$\Sigma Y = 100$	$\Sigma y = 0$	$\Sigma y^2 = 250$	$\Sigma xy = 45$

$$\bar{X} = \frac{15}{5} = 3, \bar{Y} = \frac{100}{5} = 20$$

Regression Coefficient of X on Y

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{45}{250} = 0.18$$

Regression Coefficient of Y on X

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{45}{10} = 4.5$$

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.18 \times 4.5} = \sqrt{0.81} = 0.9.$$

Notes

(b) Calculation of Regression coefficients from the assumed means 2 and 18

X	$(X-2)$		Y	$(Y-18)$			
d_x	d_x^2			d_y	d_y^2	$d_x d_y$	
1	– 1	1		10	– 8	64	+ 8
2	0	0		20	+ 2	4	0
3	+ 1	1		15	– 3	9	– 3
4	+ 2	4		25	+ 7	49	+ 14
5	+ 3	9		30	+ 12	144	+ 36
$\Sigma X = 15$	$\Sigma d_x = 5$	$\Sigma d_x^2 = 15$		$\Sigma Y = 100$	$\Sigma d_y = 10$	$\Sigma d_y^2 = 270$	$\Sigma d_x d_y = 55$

$$b_{xy} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_y^2 - (\Sigma d_y)^2}$$

$$\Sigma d_x d_y = 55,\ \Sigma d_x = 5,\ \Sigma d_y = 10,\ \Sigma d_y^2 = 270,\ N = 5$$

$$= \frac{(5)(55) - (5)(10)}{(5)(270) - (10)^2} = \frac{275 - 50}{1,350 - 100} = \frac{225}{1,250} = 0.18$$

$$b_{yx} = \frac{N\Sigma d_x d_y - (\Sigma d_x)(\Sigma d_y)}{N\Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{5(55) - (5)(10)}{(5)(15) - (5)^2} = \frac{275 - 50}{75 - 25} = \frac{225}{50} = 4.5$$

Thus the value of regression coefficients is the same by both the methods.

Example 5: Find the regression coefficient of X on Y and Y on X for the following data:

X	3	2	– 1	6	4	– 2	5
Y	5	13	12	– 1	2	20	0

Solution: **Calculation of Regression Coefficients**

X	$(X-2)$		Y	$(Y-7)$		
	d_x	d_x^2		d_y	d_y^2	$d_x d_y$
3	+ 1	1	5	– 2	4	– 2
2	0	0	13	+ 6	36	0
– 1	– 3	9	12	+ 5	25	– 15
6	+ 4	16	– 1	– 8	64	– 32
4	+ 2	4	2	– 5	25	– 10
– 2	– 4	16	20	+13	169	– 52
5	+ 3	9	0	– 7	49	– 21
$\Sigma X = 17$	$\Sigma d_x = +3$	$\Sigma d_x^2 = 55$	$\Sigma Y = 51$	$\Sigma d_y = +2$	$\Sigma d_y^2 = 372$	$\Sigma d_x d_y = -132$

Regression Coefficient of X on Y

Notes

$$b_{xy} = \frac{N \Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N \Sigma d_y^2 - (\Sigma d_y)^2}$$

$$\Sigma d_x d_y = -132,\ \Sigma d_x = +3,\ \Sigma d_y = 2,\ \Sigma d_y^2 = 372,\ N = 7$$

$$b_{xy} = \frac{(7)(-132) - (3)(2)}{(7)(372) - (2)^2} = \frac{-924 - 6}{2,604 - 4} = -0.353$$

Regression Coefficient of Y on X

$$b_{yx} = \frac{N \Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N \Sigma d_x^2 - (\Sigma d_x)^2}$$

$$= \frac{(7)(-132) - (3)(2)}{(7)(55) - (3)^2} = \frac{-924 - 6}{385 - 9} = -2.473.$$

Example 6: Given the bivariate data:

X	1	5	3	2	1	1	7	3
Y	6	1	0	0	1	2	1	3

(a) Fit a regression line of Y on X and thence predict Y if X = 5.

(b) Fit the regression line of X on Y and thence predict X if Y = 2.5.

Solution:

X	(X – 3) d_x	d_x^2	Y	(Y – 2) d_y	d_y^2
1	– 2	4	6	+ 4	16
5	+ 2	4	1	– 1	1
3	0	0	0	– 2	4
2	– 1	1	0	– 2	4
1	– 2	4	1	– 1	1
1	– 2	4	2	0	0
7	+ 4	16	1	– 1	1
3	0	0	5	+ 3	9
$\Sigma X = 23$	$\Sigma d_x = -1$	$\Sigma d_x^2 = 33$	$\Sigma Y = 16$	$\Sigma d_y = 0$	$\Sigma d_y^2 = 36$

Regression Equation of X on Y

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

Notes

$$b_{xy} = \frac{N \Sigma d_x d_y - \Sigma d_x d_y}{N \Sigma d_y^2 - \left(\Sigma d_y\right)^2}$$

$$b_{xy} = \frac{(8)(-10) - (-2)(0)}{(8)(36) - (0)^2} = \frac{-80}{288} = -0.278$$

$$\overline{X} = \frac{23}{8} = 2.875, \ \overline{Y} = \frac{16}{8} = 2, \ b_{xy} = -0.278.$$

Substituting the values in the equation

$$X - 2.875 = -0.278 (Y - 2)$$
$$X - 2.875 = -0.278 Y + 0.556$$
$$X = 3.431 - 0.278 Y$$

If $Y = 2.5$, X is equal to $3.431 + (0.278 \times 2.5)$

$$= 3.431 - 0.695 = 2.736.$$

Regression Equation of Y on X

$$Y - \overline{Y} = b_{yx}\left(X - \overline{X}\right)$$

$$b_{xy} = \frac{N \Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N \Sigma d_y^2 - \left(\Sigma d_y\right)^2}$$

$$= \frac{(8)(-10) - (-1)(0)}{8(33) - (-1)^2} = \frac{-80}{264 - 1} = -0.304$$

$$Y - 2 = -0.304 (X - 2.875)$$
$$Y - 2 = -0.304 X + 0.874$$
$$Y = 2.874 - 0.304 X$$

$X = 5$, Y is equal to $2.874 - (0.304 \times 5) = 2.874 - 1.52$

$$= 1.354$$

$$r = \sqrt{b_{xy} \times b_{yx}} = -\sqrt{(-0.278)(-0.304)} = -0.291.$$

Example 7: Given that the means of X and Y are 65 and 67, their standard deviations are 2.5 and 3.5 respectively, and the coefficient of correlation between them is 0.8.

(i) Write down the two regression lines.

(ii) Obtain the best estimate of X when Y = 70.

(iii) Using the estimated value of X as the given value of X, estimate the corresponding value of Y.

Solution:

(i) ***Regression Line of Y on X***

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}\left(X - \overline{X}\right)$$

$\overline{Y} = 67, \ \overline{X} = 65, \ \sigma_y = 3.5, \ \sigma_x = 2.5, \ r = 0.8$

Notes

$$Y - 67 = 0.8\frac{3.5}{2.5}(X - 65)$$

$$Y - 67 = 1.12 (X - 65)$$

$$Y - 67 = 1.12 X - 72.8$$

$$Y = 1.12 X - 5.8 \text{ or } Y = -5.8 + 1.12 X$$

Regression Line of X on Y

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

$$X = 65 = 0.571 (Y - 67)$$

$$X - 65 = 0.571 Y - 38.257$$

$$X = 0.571 Y + 26.743 \text{ or } X = 26.743 + 0.571 Y$$

(ii) Best estimate of X when Y = 70 can be obtained from the regression equation of X on Y.

$$X = 26.743 + 0.571 (70) = 26.743 + 39.97 = 66.713$$

(iii) When X = 66.713, Y will be

$$X = 1.12 (66.713) - 5.8 = 74.72 - 5.8 = 68.92.$$

Self-Assessment

1. Which of the following statements are True or False (T/F):

(*i*) If both the regression coefficients are negative, the correlation coefficient would be negative.

(*ii*) The under root of two regression coefficients gives us the value of correlation coefficient.

(*iii*) Regression coefficients are independent of change of scale and origin.

(*iv*) Regression coefficient of Y on X measures the change in X corresponding to a unit change in Y.

(*v*) The regression coefficient of Y on X is denoted by the symbol *bxy*.

13.3 Summary

- The *statistical tool with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable is called* **regression.** With the help of regression analysis,* we are in a position to find out the *average probable* change in one variable given a certain amount of change in another.
- Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables, price (X) and demand (Y), are closely related, we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of commodity are closely related, we can find out the expected price for a certain amount of tax levy. Thus we find that the study of regression is of considerable help to the economists and businessmen.
- Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations–the regression of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.
- If the values of the constants '*a*' and '*b*' are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of Least Squares

Notes

which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the deviation of the actual Y values from the computed Y values is the least or, in other words, in order to obtain a line which fits the points best $\sum (Y - Y_c)^2$ should be minimum. Such a line is known as the line of 'best fit'.

- The quantity *b* in the regression equation is called the regression coefficient. Since there are two regression equations there are also two regression coefficients–regression coefficient of X on Y and regression coefficient of Y on X.
- the regression coefficient which may exceed one should also be such in value that when multiplied by the other coefficient the underroot of the product of the two coefficients does not exceed one. Also both the regression coefficients will have the same sign, *i.e.,* they will be either positive or negative. The coefficient of correlation (*r*) will have the same sign as that of regression coefficients, *i.e.,* if regression coefficients have a negative sign, *r* will also have negative sign and if regression coefficients have a positive sign, *r* will also have positive sign.

13.4 Key-Words

1. Regression : A statistical measure that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

13.5 Review Questions

1. What are regression coefficients ? State the properties of regression coefficients
2. What are regression equations ? Explain regression equation of Y on X and X on Y.
3. If one of the regression coefficients is negative what type of variation would you expect in the original series of pairs of observations ?
4. Give the various properties and characteristics of regression coefficients.
5. If two regression coefficients are 0.8 and 0.6, what would be the coefficient of correlation? [r = 0.693]

Answers: Self-Assessment

1. (*i*) T (*ii*) T (*iii*) F (*iv*) F (*v*) F

13.6 Further Readings



1. Elementary Statistical Methods; SP. Gupta, Sultan Chand & Sons, New Delhi - 110002.
2. Statistical Methods — An Introductory Text; Jyoti Prasad Medhi, New Age International Publishers, New Delhi - 110002.
3. Statistics; E. Narayanan Nadar, PHI Learning Private Limied, New Delhi - 110012.
4. Quantitative Methods—Theory and Applications; J.K. Sharma, Macmillan Publishers India Ltd., New Delhi - 110002.