VELS
INSTITUTE OF SCIENCE, TECHNOLOGY & ADVANCED STUDIES (VISTAS)
(Deemed to be University Estd. u/s 3 of the UGC Act, 1956)
PALLAVARAM - CHENNAI

DCMBA-12

Business Statistics

MBA
ODL MODE
(Semester Pattern)

School of Management Studies and Commerce
Centre for Distance and Online Education
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Pallavaram, Chennai - 600117

Vels Institute of Science, Technology and Advanced Studies

Centre for Distance and Online Education

Master of Business Administration (MBA)- ODL Mode

(Semester Pattern)

DCMBA-12: Business Statistics

(4 Credits)

Course Design and Preparation Committee

Dr.P.Ramakrishnan
Dean, School of Management Studies and Commerce, VISTAS, Pallavaram, Chennai

Dr. B.P. Chandramohan
Director, School of Management Studies and Commerce, VISTAS, Pallavaram, Chennai

Dr.G.Rajini
Professor, Department of Management Studies, VISTAS, Pallavaram, Chennai

Course Writer

Mr.K.Sankar Singh
Assistant Professor, Department of Management Studies, VISTAS, Pallavaram, Chennai

Programme Coordinator

Dr.S.Sudha
Professor, Department of Management Studies, CDOE, VISTAS, Pallavaram, Chennai

Content Editing

Dr.L.Sudha
Assistant Professor, Department of Management Studies, CDOE, VISTAS, Pallavaram, Chennai

Language Editing

Dr.A.A.Jayashree Prabhakar
Professor, Department of English , VISTAS, Pallavaram, Chennai

Printing and Distribution

Ms.M.S.Vijayalakshmi
Deputy Registrar, CDOE, VISTAS, Pallavaram, Chennai

Mr.V.Kumar
Section Officer, CDOE, VISTAS, Pallavaram, Chennai

September 2023 (First Edition)

Further information on the VISTAS ODL Academic Programmes may be obtained from VISTAS-CDOE at Velan Nagar, P.V.Vaithiyalingam Road, Pallavaram, Chennai-600117 [or] www.vistas.ac.in.

Printed at:

FOREWORD

Dr.Ishari K Ganesh
Chancellor

Vels Institute of Science, Technology and Advanced Studies (VISTAS), deemed to be a university, was established in 2008 under section 3 of the Act of 1956 of the University Grants Commission, Government of India, New Delhi.

VISTAS has blossomed into a multi-disciplinary Institute offering more than 100 UG & PG Programmes, besides Doctoral Programmes, through 18 Schools and 46 Departments. All the Programmes have the approval of the relevant Statutory Regulating Authorities such as UGC, UGC-DEB, AICTE, PCI, BCI, NCTE and DGS.

The deemed to be University aims to provide innovative syllabi and industry-oriented courses, and hence, the revision of curricula is a continuous and ongoing process. The revision is initiated by the faculty depending on the requirement and approved by the Board of Studies of the concerned Department/School. The courses are under Choice Based Credit Systems that enable students to get adequate freedom in choosing subjects.

I am pleased to inform you that VISTAS has been rendering its services to society to democratize the opportunities of higher education for those who are in need through Open and Distance Learning (ODL) mode.

VISTAS ODL Programmes offered have been approved by the University Grants Commission (UGC) – Distance Education Bureau (DEB), New Delhi.

The curriculum and syllabi have been approved by the Board of Studies, Academic Council, and the Executive Committee of the VISTAS, and they are designed to help provide employment opportunities to the students.

The ODL Programme (B.Com., BBA and MBA) study material have been prepared in the Self Instructional Mode (SIM) format as per the UGC-DEB (ODL & OL) Regulations 2020. It is highly helpful to the students, faculties and other professionals. It gives me immense pleasure to bring out the ODL programme with a noble cause of enriching learners' knowledge. I extend my congratulations and appreciation to the Programme Coordinator and the entire team for bringing up the ODL Programme in an elegant manner.

At this juncture, I am glad to announce that the syllabus of this ODL Programme has been made available on our website, www.vistas.ac.in, for the benefit of the student fraternity and other knowledge seekers. I wish that this Self Learning Materials (SLM) would be a nice treatise to the academic community and everyone.

CHANCELLOR

FOREWORD

Dr.S.Sriman Narayanan
Vice-Chancellor

My Dear Students!

Open and Distance Learning (ODL) of VISTAS gives you the flexibility to acquire a University degree without the need to visit the campus often. VISTAS-CDOE involves the creation of an educational experience of qualitative value for the learner that is best suited to the needs outside the classroom. My wholehearted congratulations and delightful greetings to all those who have availed themselves of the wonderful leveraged opportunity of pursuing higher education through this Open and Distance Learning Programme.

Across the world, pursuing higher education through Open and Distance Learning Systems is on the rise. In India, distance education constitutes a considerable portion of the total enrollment in higher education, and innovative approaches and programmes are needed to improve it further, comparable to Western countries where close to 50% of students are enrolled in higher education through ODL systems.

Recent advancements in information and communications technologies, as well as digital teaching and e-learning, provide an opportunity for non-traditional learners who are at a disadvantage in the conventional system due to age, occupation, and social background to upgrade their skills.

VISTAS has a noble intent to take higher education closer to the oppressed, underprivileged women and the rural folk to whom higher education has remained a dream for a long time.

I assure you all that the Vels Institute of Science, Technology and Advanced Studies would extend all possible support to every registered student of this deemed to be university to pursue her/his education without any constraints. We will facilitate an excellent ambience for your pleasant learning and satisfy your learning needs through our professionally designed curriculum, providing Open Educational Resources, continuous mentoring and assessments by faculty members through interactive counselling sessions.

This university brings to reality the dreams of the great poet of modern times, Mahakavi Bharathi, who envisioned that all our citizens be offered education so that the globe grows and advances forever.

I hope that you achieve all your dreams, aspirations, and goals by associating yourself with our ODL System for never-ending continuous learning.

With warm regards,

VICE-CHANCELLOR

Course Introduction

This course **DCMBA-12: Business Statistics** mainly focuses on improving knowledge of basic statistical tools and techniques with emphasis on their application in Business decision process and Management and also focus on more practical than theoretical. Practices to do statistical analysis informs the judgment of the ultimate decision-maker-rather than replaces it-some key conceptual under pinning of statistical analysis and covered to insure the understandability of its proper usage.

Block-1: Introduction to Business Statistics deals with Introduction to Statistics, Definitions of statistics and limitations of statistics, Collection of Data, Primary data and Secondary data, Measures of Central Tendency, Mean, Median and Mode, Measures of Dispersions, Range, Quartile deviation, Standard deviation and Coefficient of variation.

Block-2: **Probability Theory** explains about the Introduction to Probability, Definition of probability, Types of events, Addition Theorem of probability, Conditional and Multiplication Theorem of probability, Bayes' Theorem of probability and Application problems.

Block-3: **Correlation and Regression** presents with Correlation, Definition of correlation, Types of correlation, Scatter diagram, Karl Pearson correlation and Spearman correlation, Partial correlation and Multiple correlation, Regression Analysis and also the Regression equation of X on Y and Y on X.

Block-4: Testing of Hypothes is describes about Introduction to testing of hypothesis, Types of hypothesis, level of significance, test statistic, one tail test and two tail test, Type-I error and Type-II error, Student t test for single mean and double mean for small sample data, Chi-square test and Analysis of Variance.

Block-5: Index numbers and Time Series Analysis deals with Introduction to Index Numbers, definitions and applications of index numbers, Construction of Price Index Numbers, Laspyre, Paasche and Fisher price index numbers, Introduction to Time Series, Components of Time Series, Trend, Seasonal variation, cyclical variation, random variation and Methods for finding the trend values.

This Course DCMBA-12: Business Statistics has been divided in to 20 Units

DCMBA-12: Business Statistics

Block-1: Introduction

Block-1: Introduction to Business Statistics has been divided into four Units (Unit-1 to Unit-4).

Unit-1: Introduction to Statistics deals with Origin and growth of statistics, Definition of Statistics, Characteristics of statistics, Importance and Scope of Statistics, Limitations of Statistics, Descriptive and Inferential statistics.

Unit-2: Collection of Data explains about Introduction, Types of Data, Primary and Secondary Data, Methods of Collecting Primary Data and Sources of Secondary Data.

Unit-3: Measures of Central Tendency describes about the **Introduction**, Characteristics for a good statistical average, Arithmetic mean, Median and Mode.

Unit-4: Measures of Dispersions presents with Introduction, Range, Quartile Deviation, Standard Deviation and Co-efficient of Variation.

In all the Units of Block -1: **Introduction to Business Statistics**, the Check your progress, Glossary, Answers to Check your progress and Suggested Reading has been provided and the Learners are expected to attempt all the Check your progress as part of study.

Unit -1
Introduction to Statistics

STRUCTURE

Overview

In this unit the origin and growth of statistics, definition of statistics and limitations statistics along with the details about descriptive statistics and inferential statistics has been clearly explained.

Objectives

After completion of this unit, you will be able to:

- Highlights the origin and growth of statistics
- Introduces the meaning and definitions of statistics
- Presents the scope and functions of statistics
- Explains the applications of statistics in different fields

1.1. Origin and growth of statistics

The origin of statistics can be traced back to the primitive man, who put notches on trees to keep an account of his belongings. During 5000 BCE, kings used to carry out census of populations and resources of the state. Kings of olden days made their crucial decisions on wars, based on statistics of infantry, cavalry and elephantary units of their own and that of their enemies.

Later it enhanced its scope in their kingdoms' tax management and administrative domains. Thus, the word 'Statistics' has its root either to Latin word 'Status' or Italian word 'Statista' or German word 'Statistik' each of which means a 'political state'.

The word 'Statistics' was primarily associated with the presentation of facts and figures pertaining to demographic, social and political situations prevailing in a state/government. Its evolution over time formed the basis for most of the science and art disciplines. Statistics is used in the developmental phases of both theoretical and applied areas, encompassing the field of Industry, Agriculture, Medicine, Sports and Business analytics.

In olden days statistics was used for political- war purpose. Later, it was extended to taxation purposes. This is evident from Kautilya's Arthasastra (324 – 300 BCE). Akbar`s finance minister Raja Thodarmall collected information regarding agricultural land holdings. During the seventeenth century, statistics entered in vital statistics, which is the basis for the modern day Actuarial Science. Gauss introduced the theory of errors in physical sciences at the end of eighteenth century.

Statistics is used in two different forms-singular and plural. In plural form it refers to the numerical figures obtained by measurement or counting in a systematic manner with a definite purpose such as number of accidents in a busy road of a city in a day, number of people died due to a chronic disease during a month in a state and so on. In its singular form, it refers to statistical theories and methods of collecting, presenting, analyzing and interpreting numerical figures.

Though the importance of statistics was strongly felt, its tremendous growth was in the twentieth century. During this period, lot of new theories, applications in various disciplines were introduced. With the contribution of renowned statisticians several theories and methods were introduced, naming a few are Probability Theory, Sampling Theory, Statistical Inference, Design of Experiments, Correlation and Regression Methods, Time Series and Forecasting Techniques.

In early 1900s, statistics and statisticians were not given much importance but over the years due to advancement of technology it had its wider scope and gained attention in all fields of science and management. We also tend to think statistician as a small profession but a steady growth in the last century is impressive.

It is pertinent to note that the continued growth of statistics is closely associated with information technology. As a result several new inter-

disciplines have emerged. They are Data Mining, Data Warehousing, Geographic Information System, Artificial Intelligence etc. Now-a-days, statistics can be applied in hardcore technological spheres such as Bioinformatics, Signal processing, Telecommunications, Engineering, Medicine, Crimes, Ecology, etc.

Today's business managers need to learn how analytics can help them make better decisions that can generate better business outcomes. They need to have an understanding of the statistical concepts that can help analyze and simplify the flood of data around them. They should be able to leverage analytical techniques like decision trees, regression analysis, clustering and association to improve business processes.

1.2. Definition of Statistics

1. "Statistics are the classified facts representing the conditions of the people in a State…specially those facts which can be stated in number or in tables of numbers or in any tabular or classified arrangement." - Webster.
2. "Statistics are numerical statements of facts in any department of enquiry placed in relation to each other." -Bowley.
3. "By statistics we mean quantitative data affected to a marked extent by multiplicity of causes". - Yule and Kendall.
4. 'Statistics is the science of counting' -A. L .Bowley
5. 'Statistics is the science which deals with the collection, presentation, analysis and interpretation of numerical data' - Croxton and Cowden
6. Wallist and Roberts defines statistics as "Statistics is a body of methods for making decisions in the face of uncertainty"
7. Ya-Lun-Chou slightly modifies Wallist and Roberts's definition and come with the following definition: "Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risk."

1.3 Characteristics of statistics

(i) **Aggregate of Facts.** Simple or isolated items cannot be termed as Statistics unless they are a part of aggregate of facts relating to any particular field of enquiry. For instance, the height of an individual or the prices of a particular commodity do not form Statistics as such figures are unrelated and incomparable.

(ii) **Affected by Multiplicity of Causes**: Numerical figures should be affected by multiplicity of factors. In physical sciences, it is possible to isolate the effect of various factors on a single item but it is very difficult to do so in social sciences, particularly when the effect of some of the factors cannot be measured quantitatively.

(iii) **Numerically Expressed**: Only numerical data constitute Statistics. Thus the statements like 'the standard of living of the people in Delhi has improved' or 'the production of a particular commodity is increasing' do not constitute Statistics.

(iv) **Enumerated or Estimated According to Reasonable Standard of Accuracy**: The numerical data pertaining to any field of enquiry can be obtained by completely enumerating the underlying population. In such a case data will be exact and accurate (but for the errors of measurement, personal bias, etc.).

(v) **Collected in a Systematic Manner:** The data must be collected in a very systematic manner. Thus, for any socio-economic survey, a proper schedule depending on the object of enquiry should be prepared and trained personnel (investigators) should be used to collect the data by interviewing the persons. An attempt should be made to reduce the personal bias to the minimum.

(vi) **Collected for a Pre-determined Purpose**: It is of utmost importance to define in clear and concrete terms the Objectives or the purpose of the enquiry and the data should be collected keeping in view these Objectives. An attempt should not be made to collect too many data some of which are never examined or analyzed.

(vii) **Comparable:** From practical point of view, for statistical analysis the data should be comparable. They may be compared with respect to some unit, generally time (period) or place. For example, the data relating to the population of a country for different years or the population of different countries in some fixed year constitute Statistics, since they are comparable.

Check Your Progress-1

True/False

a. During 5000 BCE, kings used to carry out census of populations and resources of the state.

b. 'Statistics is the science of counting' -A. L .Bowley

c. Though the importance of statistics was strongly felt, its tremendous growth was in the twentieth century.
d. In old days, Economic Theories were based on deductive logic only.
e. In ancient times the scope of statistics was not limited.

1.4. Importance and Scope of Statistics

In ancient times the scope of statistics was limited. When people hear the word 'Statistics' they think immediately of either sports related numbers or a subject they have studied at college and passed with minimum marks. While statistics can be thought in these terms there is a wide scope for statistics.

Today, there is no human activity which does not use statistics. There are two major divisions of statistical methods called descriptive statistics and inferential statistics and each of the divisions are important and satisfies different objectives. The descriptive statistics is used to consolidate a large amount of information. For example, measures of central tendency, like mean are descriptive statistics. Descriptive statistics just describes the data in a condensed form for solving some limited problems. They do not involve beyond the data at hand.

Statistics in Planning: Statistics is indispensable in planning – may it be in business, economics or government level. The modern age is termed as the 'age of planning' and almost all organizations in the government or business or management are resorting to planning for efficient working and for formulating policy decisions.

Statistics in State: As has already been pointed out, in the old days Statistics was the science of State- craft and its Objectives was to collect data relating to manpower, crimes, income and wealth, etc., for formulating suitable military and fiscal policies. With the inception of the idea of Welfare State and its taking deep roots in almost all the countries, today statistical data relating to prices, production, consumption, income and expenditure, investments and profits, etc.,

Statistics in Economics: In old days, Economic Theories were based on deductive logic only. Moreover, the statistical techniques were not that much advanced for applications in other disciplines. It gradually dawned upon economists of the Deductive School to use Statistics effectively by making empirical studies. In 1871, W.S. Jevons, wrote that: "The deductive science of economy must be verified and rendered useful from the purely inductive science of Statistics. Theory must be invested with the reality of life and fact." These views were supported by

Roscher, Kines and Hildebrand of the Historical School (1843 – 1883), Alfred Marshall, Pareto, Lord Keynes. The following quotation due to Prof. Alfred Marshall in 1890 amply illustrates the role of Statistics in Economics: “Statistics are the straws out of which I, like every other economist, have to make bricks.”

Statistical data and advanced techniques of statistical analysis have proved immensely useful in the solution of a variety of economic problems such as production, consumption, distribution of income and wealth, wages, prices, profits, savings, expenditure, investment, unemployment, poverty, etc. Use of Statistics in Economics has led to the formulation of many economic laws some of which are mentioned below for illustration:

A detailed and systematic study of the family budget data which gives a detailed account of the family budgets showing expenditure on the main items of family consumption together with family structure and composition, family income and various other social, economic and demographic characteristics led to the famous Engel’s Law of Consumption in 1895. Vilfredo Pareto in 19th-20th century propounded his famous Law of Distribution of Income by making an empirical study of the income data of various countries of the world at different times. Time Series Analysis, Index Numbers, Forecasting Techniques and Demand Analysis are some of the very powerful statistical tools which are used immensely in the analysis of economic data and also for economic planning.

The demand analysis consists in making an economic study of the market data to determine the relation between:

(i) The prices of a given commodity and its absorption capacity for the market i.e., demand; and

(ii) The price of a commodity and its output i.e., supply. Forecasting techniques based on the method of curve fitting by the principle of least squares and exponential smoothing are indispensable tools for economic planning.

Statistics in Business and Management: Prior to the Industrial Revolution, when the production was at the handicraft stage, the business activities were very much limited and were confined only to small units operating in their own areas. The owner of the concern personally looked after all the departments of business activity like sales, purchase, production, marketing, finance and so on. According to Wallis and Roberts: “Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty.” A refinement over this

definition is provided by Prof. Ya-Lun-Chou as follows: “Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks.” These definitions reflect the applications of Statistics in business since modern business has its roots in the accuracy and precision of the estimates and statistical forecasting regarding the future demand for the product, market trends and so on. Wrong expectations which might be the result of faulty and inaccurate analysis of various factors affecting a particular phenomenon might lead to his disaster.

The time series analysis is a very important statistical tool which is used in business for the study of:

(i) Trend (by method of curve fitting by the principle of least squares) in order to obtain the estimates of the probable demand of the goods; and

(ii) Seasonal and Cyclical movements in the phenomenon, for determining the ‘Business Cycle’ which may also be termed as the four-phase cycle composed of prosperity (period of boom), recession, depression and recovery. The upswings and downswings in business depend on the cumulative nature of the economic forces (affecting the equilibrium of supply and demand) and the interaction between them.

The studies of Economic Barometers (Index Numbers of Prices) enable the businessman to have an idea about the purchasing power of money. The statistical tools of demand analysis enable the businessman to strike a balance between supply and demand. Statistical tools of probability and expectation are extremely useful in Life Insurance which is one of the pioneer branches of Business and Commerce to use Statistics since the end of the seventeenth century. Statistical techniques have also been used very widely by business organizations in:

(i) Marketing Decisions (based on the statistical analysis of consumer preference studies – demand analysis).

(ii) Investment (based on sound study of individual shares and debentures).

(iii) Personnel Administration (for the study of statistical data relating to wages, cost of living, incentive plans, effect of labour dispute/unrest on the production, performance standards, etc.).

From the above discussion it is obvious that the use of statistical data and techniques is indispensable in almost all the branches of business activity.

Statistics in Accountancy and Auditing: Today, the science of Statistics has assumed such unprecedented dimensions that even the subjects like Accountancy and Auditing have not escaped its domain. The ever-increasing applications of the statistical data and the advanced statistical techniques in Accountancy and Auditing are well supported by the inclusion of a compulsory paper on Statistics both in the Chartered Accountants (Foundation) and Cost and Works Accountants (Intermediate) examinations curriculum. Statistics has innumerable applications in accountancy and auditing.

Statistics in Industry: In industry, Statistics is extensively used in 'Quality Control'. The main Objectives in any production processes it to control the quality of the manufactured product so that it conforms to specifications. This is called process control and is achieved through the powerful technique of control charts and inspection plans.

Statistics in Physical Sciences: The applications of Statistics in Astronomy, which is a physical science, have already been discussed above. In physical sciences, a large number of measurements are taken on the same item. There is bound to be variation in these measurements.

Statistics in Social Sciences: According to Bowley, "Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations." In the words of W.I. 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 of estimates." These words of Bowley and King amply reflect upon the importance of Statistics in social sciences.

Statistics in Biology and Medical Sciences: Sir Francis Galton (1822 – 1911), a British Biometrician pioneered the use of statistical methods with his work on 'Regression' in connection with the inheritance of stature. According to Prof. Karl Pearson (1857 – 1936) who pioneered the study of 'Correlation Analysis', the whole theory of heredity rests on statistical basis.

In medical sciences also, the statistical tools for the collection, presentation and analysis of observed factual data relating to the causes and incidence of diseases are of paramount importance.

Statistics in Psychology and Education: Statistics has been used very widely in education and psychology too e.g., in the scaling of mental tests and other psychological data; for measuring the reliability and validity of test scores; for determining the Intelligence Quotient (I.Q.); in Item Analysis and Factor Analysis.

1.5. Limitations of Statistics

Although Statistics is indispensable to almost all sciences - social, physical and natural, and is very widely used in almost all spheres of human activity, it is not without limitations which restrict its scope and utility.

1. **Statistics does not Study Qualitative Phenomenon:** 'Statistics are numerical statements in any department of enquiry placed in relation to each other'. Since Statistics is a science dealing with a set of numerical data, it can be applied to the study of only those phenomena which can be measured quantitatively.
2. **Statistics does not Study Individuals:** According to Prof. Horace Secrist, "By Statistics we mean aggregate of facts affected to a marked extent by multiplicity of factors…and placed in relation to each other." Thus a single or isolated figure cannot be regarded as Statistics unless it is a part of the aggregate of facts relating to any particular field of enquiry. Thus statistical methods do not give any recognition to an object or a person or an event in isolation. This is a serious limitation of Statistics. For instance, the prices of a single commodity, the profit of a particular concern or the production of a particular business house do not constitute statistics since these figures are unrelated and incomparable.
3. **Statistical Laws are not exact:** Since the statistical laws are probabilistic in nature, inferences based on them are only approximate and not exact like the inferences based on mathematical or scientific (physical and natural sciences) laws.
4. **Statistics is Liable to be misused:** Perhaps the most significant limitation of Statistics is that it must be used by experts. According to Bowley, "Statistics only furnishes a tool though imperfect which is dangerous in the hands of those who do not know its use and deficiencies." Statistical methods are the most dangerous tools in the hands of the in experts. Statistics is one of those sciences whose adepts must exercise the self-restraint of an artist.

1.6. Descriptive and Inferential Statistics

Descriptive statistics give information that describes the data in some manner. For example, suppose a pet shop sells cats, dogs, birds and fish. If 100 pets are sold, and 40 out of the 100 were dogs, then one description of the data on the pets sold would be that 40% were dogs. This same pet shop may conduct a study on the number of fish sold each day for one month and determine that an average of 10 fish were sold each day. The average is an example of descriptive statistics.

Some other measurements in descriptive statistics answer questions such as 'How widely dispersed is this data?', 'Are there a lot of different values?' or 'Are many of the values the same?', 'What value is in the middle of this data?', 'Where does a particular data value stand with respect with the other values in the data set?'

A graphical representation of data is another method of descriptive statistics. Examples of this visual representation are histograms, bar graphs and pie graphs, to name a few. Using these methods, the data is described by compiling it into a graph, table or other visual representation.

Inferential statistics is a way of making inferences about populations based on samples. With inferential statistics, you take data from samples and make generalizations about a population. For example, you might stand in a mall and ask a sample of 100 people if they like shopping at Sears.

Let us sum up

The word statistics can be used either plural sense or in singular sense. When used in plural sense, the word statistics refers to numerical statements of facts or data. To be called statistics, numerical data should possess the following characteristics:

1) they must be aggregate of facts,

2) they must be affected by multiplicity of factors,

3) they must be numerically expressed,

4) they must be enumerated or estimated according to a reasonable standard of accuracy,

5) they must be collected in a systematic manner for a predetermined purpose, and

6) they must be placed in relation to each other.

The word statistics, when used in singular sense, refers to a body of knowledge which provides methods and techniques required for:

1) collection of data,

2) classification and tabulation of data,

3) presentation of data,

4) analysis of data, and

5) interpretation of data. In this unit you have learned about introductory concept of business statistics.

Check your progress-2

1. The collection of numerical information is called_____.

2. Statistics always deals with_____data.

3. In statistics all qualitative data converted into _____data.

Glossary

Data: A collection of numerical information.

Descriptive Statistics: Refers to methods and techniques of summarizing and describing the characteristics of the data.

Inferential Statistics: Refers to those methods which are helpful in drawing inferences about the characteristics of the population on the basis of sample data.

Statistical Data: Information expressed in quantitative or numerical form is called statistical data. All statistical data is numerical statements of facts but all numerical statements of facts are nor statistics.

Statistical Methods: A body of methods and principles that are helpful in the collection, summarization, description, analysis and interpretation of numerical data.

Statistics: When used in plural sense, refer to numerical statements of facts or data. When used in singular sense, refers to a body of methods which provides.

Answers to check your progress-1

a-True

b-True

c-True

d-True

e-False

Answers to check your progress-2

1. Statistics
2. Quantitative data
3. Quantitative

Suggested Readings

1. McEvoy, David M. A Guide to Business Statistics, John Wiley and Sons(2020)
2. N.D. Vohra, “Business Statistics”, Tata McGraw-Hill Education, 2ndEd,2018
3. Berenson M., Levine D., Szabat K.A. and Krehbiel T.C. Basic Business Statistics: Concepts and Applications, Pearson Higher Education AU (2012).

Unit -2
Collection of Data

STRUCTURE

Overview

In this unit the different types of data and what are the different methods for data collection, primary data, secondary data and sources for each method has been clearly explained.

Objectives

After completion of this unit, you will be to :

- Emphasis the necessity of data collection
- Distinguishes between primary and secondary data
- Introduces methods of collecting primary data with their advantages and disadvantages
- Designs a questionnaire for the collection of data.
- Describes Secondary data

2.1. Introduction

Statistics are a set of numerical data. In fact only numerical data constitute Statistics. This means that the phenomenon under study must be capable of quantitative measurement. Thus the raw material of Statistics always originates from the operation of counting (enumeration) or measurement. For any statistical enquiry, whether it is in business,

economic or social sciences, the basic problem is to collect facts and figures relating to particular phenomenon under study. The person who conducts the statistical enquiry i.e., counts or measures the characteristics under study for further statistical analysis is known as investigator.

The process of counting or enumeration or measurement together with the systematic recording of results is called the collection of statistical data. The entire structure of the statistical analysis for any enquiry is based upon systematic collection of data. We embark upon the collection of data for a given statistical enquiry, it is imperative to examine carefully the following points which may be termed as preliminaries to data collection:

i. Objectives and scope of the enquiry
ii. Statistical units to be used
iii. Sources of information (data)
iv. Method of data collection
v. Degree of accuracy aimed at in the final results.
vi. Type of enquiry

We shall discuss these points briefly in the following sections.

Objectives and Scope of the Enquiry: The first and foremost step in organizing any statistical enquiry is to define in clear and concrete terms the Objectives of the enquiry. This is very essential for determining the nature of the statistics (data) to be collected and also the statistical techniques to be employed for the analysis of the data.

Statistical Units to be used: A well-defined and identifiable object or a group of objects with which the measurements or counts in any statistical investigation are associated is called a statistical unit. For example, in a socio-economic survey the unit may be an individual person, a family, a household or a block of locality. A very important step before the collection of data begins is to define clearly the statistical units on which the data are to be collected.

Requisites of a Statistical Unit

The following points might serve as guidelines for deciding about the unit in any statistical enquiry.

1. It should be unambiguous. A statistical unit should be rigidly defined so that it does not lead to any ambiguity in its interpretation. The units must cover the entire population and they should be distinct and non-overlapping in the sense that

every element of the population belongs to one and only one statistical unit.

2. It should be specific. The statistical unit must be precise and specific leaving no chance to the investigators. Quite often, in most of the socio-economic surveys the various concepts/ characteristics can be interpreted in different variant forms and accordingly the variable used to measure it may be defined in several different ways.

Types of Statistical Units

The statistical units may be broadly classified as follows:

(i) Units of collection.

(ii) Units of analysis and interpretation.

(i) **Units of Collection:** The units of collection may further be sub-divided into the following two classes:

(a) **Units of Enumeration:** In any statistical enquiry, whether it is conducted by 'sample' method or 'census' method, unit of enumeration is the basic unit on which the observations are to be made and this unit is to be decided in advance before conducting the enquiry keeping in view the Objectives of the enquiry. The unit of enumeration may be a person, a household, a family, a farm (in land experiments), a shop, a livestock, a firm, etc. As has been pointed out earlier, this unit should be very clearly defined in terms of shape, size, etc. For instance, for the construction of cost of living index number, the proper unit of enumeration is household.

(b) **Units of Recording:** The units of recording are the units in terms of which the data are recorded or in other words they are the units of quantification. For instance, in the construction of cost of living index number (consumer price index) the data to be collected from each household, among other things, include the retail prices of various commodities together with the quantities consumed by the class of people for whom the index is meant.

(ii) **Units of Analysis and Interpretation:** As the name implies, the units of analysis and interpretation are those units in the form of which the statistical data are ultimately analyzed and interpreted. It should be decided whether the results would be expressed in absolute figures or relative figures.

Sources of Information (Data): Having decided about the Objectives and scope of the enquiry and the statistical units to be used, the next problem is to decide about the sources from which the information (data) can be obtained or collected. For any statistical enquiry, the investigator may collect the data first hand or he may use the data from other published sources such as the publications of the government/semi-government organizations, periodicals, magazines, newspapers, research journals, etc.

Method of Data Collection: The problem does not arise if secondary data are to be used.

However, if primary data are to be collected a decision has to be taken whether census method

Degree of Accuracy Aimed at in the Final Results: A decision regarding the degree of accuracy or precision desired by the investigator in his estimates or results is essential before starting any statistical enquiry. An idea about the precision aimed at is extremely helpful in deciding about the method of data collection and the size of the sample (if the enquiry is to be on the basis of a sample study).

The information gained from any previous completed sample study on the subject in the form of precision achieved for a given sample size may serve as a useful guide in this matter provided there is no fundamental reason to change this empirical basis. In any statistical enquiry perfect accuracy in final results is practically impossible to achieve because of the errors in measurement, collection of data, its analysis and interpretation of the results.

Types of Enquiry: Another important point one has to bear in mind before embarking upon the process of collection of data is to decide about the type of enquiry. The statistical enquiries may be of different types as outlined below:

(i) Official, Semi-official or Un-official
(ii) Initial or Repetitive
(iii) Confidential or Non-confidential
(iv) Direct or Indirect
(v) Regular or Ad-hoc
(vi) Census or Sample
(vii) Primary or Secondary

(i) Official, Semi-official or Un-official Enquiry: A very important factor in the collection of data is 'the sponsoring agency of the

survey or enquiry'. If an enquiry is conducted by or on behalf of the central, state or local governments it is termed as official enquiry. A semi-official enquiry is one that is conducted by organizations enjoying government patronage like the Indian Council of Agricultural Research (I.C.A.R.), New Delhi; Indian Agricultural Statistics Research Institute (I.A.S.R.I.), New Delhi; Indian Statistical Institute (I.S.I.), Calcutta and New Delhi; and so on.

An un-official enquiry is one which is sponsored by private institutions like the F.I.C.C.I., trade unions, universities or the individuals.

(ii) **Initial or Repetitive Enquiry:** As the name suggests, an initial or original enquiry is one which is conducted for the first time while a repetitive enquiry is one which is carried on in continuation or repetition of some previously conducted enquiry (enquiries).

(iii) **Confidential or Non-confidential Enquiry:** In a confidential enquiry, the information collected and the results obtained are kept confidential and they are not made known to the public. The findings of such enquiries are meant only for the personal records of the sponsoring agency.

(iv) **Direct or Indirect Enquiry:** An enquiry is termed as direct if the phenomenon under study is capable of quantitative measurement such as age, weight, income, prices, quantities consumed and so on. However, if the phenomenon under study is of a qualitative nature which is not capable of quantitative measurement like honesty, beauty, intelligence, etc., the corresponding enquiry is termed as indirect one.

(v) **Regular or Ad-hoc Enquiry:** If the enquiry is conducted periodically at equal intervals of time (monthly, quarterly, yearly, etc.), it is said to be regular enquiry. For example, the census is conducted in India periodically every 10 years.

2.2. Types of Data

Data can be classified as Qualitative and Quantitative data

Quantitative data are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject.

These data may be represented by ordinal, interval or ratio scales and lend themselves to most statistical manipulation.

Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, socio economic status, religious preference are usually considered to be qualitative data.

Both types of data are valid types of measurement, and both are used in education journals. Only quantitative data can be analyzed statistically, and thus more rigorous assessments of the data are possible.

Check your progress-1

True/False

a. The entire structure of the statistical analysis for any enquiry is based upon systematic collection of data.
b. The first and foremost step in organizing any statistical enquiry is to define in clear and concrete terms the Objectives of the enquiry.
c. Primary data are the first hand information which is collected, compiled and published by organizations for some purpose.
d. Statistics are not a set of numerical data.
e. Only quantitative data can be analyzed statistically, and thus more rigorous assessments of the data are possible.

2.3. Primary and Secondary Data

Primary Data

Primary data are the first hand information which is collected, compiled and published by organizations for some purpose. They are the most original data in character and have not undergone any sort of statistical treatment.
Example: Population census reports are primary data because these are collected, complied and published by the population census organization.

Secondary Data

The secondary data are the second hand information which is already collected by an organization for some purpose and are available for the present study. Secondary data are not pure in character and have undergone some treatment at least once.

Example: An economic survey of England is secondary data because the data are collected by more than one organization like the Bureau of Statistics, Board of Revenue, banks, etc.

Difference between Primary data and Secondary Data

Primary	Secondary
It is collected for the first time	Compiled from already existing sources
It is collected directly by the investigator or by his team	Compiled by persons other than the persons who collected the data
It costs more	It costs less
It requires more time	It requires considerably less time
Possibility of having personal bias	Personal bias is minimized

2.4. Methods of Collecting Primary Data

Primary data are collected using the following methods:

1. Personal Investigation: The researcher conducts the survey him/herself and collects data from it. The data collected in this way are usually accurate and reliable. This method of collecting data is only applicable in case of small research projects.
2. Through Investigation: Trained investigators are employed to collect the data. These investigators contact the individuals and fill in questionnaires after asking for the required information. Most organizations utilize this method.
3. Collection through Questionnaire: Researchers get the data from local representations or agents that are based upon their own experience. This method is quick but gives only a rough estimate.
4. Through the Telephone: Researchers get information from individuals through the telephone. This method is quick and gives accurate information.

2.5. Sources of Secondary Data

The chief sources of secondary data may be broadly classified into the following two groups:

(i) Published sources.
(ii) Unpublished sources.

Published Sources: There are a number of national (government, semi-government and private) organizations and also international agencies which collect statistical data relating to business, trade, labour,

prices, consumption, production, industries, agriculture, income, currency and exchange, health, population and a number of socio-economic phenomena and publish their findings in statistical reports on a regular basis (monthly, quarterly, annually, ad-hoc). These publications of the various organizations serve as a very powerful source of secondary data.

Unpublished Sources: The statistical data need not always be published. There are various sources of unpublished statistical material such as the records maintained by private firms or business enterprises who may not like to release their data to any outside agency ; the various departments and offices of the Central and State Governments ; the researches carried out by the individual research scholars in the universities or research institutes.

Let us sum up

In this unit, you have learned about the followings:

- Data are the ingredients on which statistics works.
- Data type may be primary data or secondary data.
- Source of getting data depends on the problem of study.
- Each method of collection of data has its own advantage and disadvantages.
- Hence an appropriate method should be used in data collection.

Check your progress-2

1. Face to face interview method is sources of _____data.

2. Collecting information from websites is called _____ data.

3. Data collection through Google form is called _____ data.

Glossary

Quantitative data: Quantitative data are those that can be quantified in definite units of measurement. These refer to characteristics whose successive measurements yield quantifiable observations.

Qualitative data: Qualitative data refer to qualitative characteristics of a subject or an object.

A characteristic is qualitative in nature when its observations are defined and noted in terms of the presence or absence of a certain attribute in discrete numbers.

Answers to check your progress-1

a-True

b-True

c-True

d-False

e-True

Answers to check your progress-2

1. Primary data

2. Secondary data

3. Primary data

Suggested Readings

1. R.S.N. Pillai, V. Bagavathi," Statistics", S.Chand Limited, 7thEd,2008
2. Elhance , D.N. Fundamentals of Statistics. Allahabad: KitabMahal, (2007).

Unit-3
Measures of Central Tendency

STRUCTURE

Overview

In this unit contains characteristics of a good statistical average and discussion about different measures of central tendencies like mean, median and mode and the same has been clearly explained.

Objectives

After completion of this unit, you will be able to:

- Knows the average as the representation of the entire group.
- To know the Characteristics for a good statistical average.
- Calculates the mathematical averages and the positional averages
- Understands the relationships among the averages and stating their uses.

3.1. Introduction

Human mind is incapable of remembering the entire mass of unwieldy data. Having learnt the methods of collection and presentation of data, one has to condense the data to get representative numbers to study the characteristics of data. The characteristics of the data set is explored with some numerical measures namely measures of central tendency, measures of dispersion, measures of skewness, and measures of

kurtosis. This unit focuses on “Measure of central tendency”. The measures of central tendency are also called “the averages”.

In practical situations one need to have a single value to represent each variable in the whole set of data. Because, the values of the variable are not equal, however there is a general tendency of such observations to cluster around a particular level.

In this situation it may be preferable to characterize each group of observations by a single value such that all other values clustered around it. That is why such measure is called the measure of central tendency of that group. A measure of central tendency is a representative value of the entire group of data. It describes the characteristic of the entire mass of data.

It reduces the complexity of data and makes them amenable for the application of mathematical techniques involved in analysis and interpretation of data.

3.2. Characteristics for a good statistical average

The following properties should be possessed by an ideal average.

- It should be well defined so that a unique answer can be obtained.
- It should be easy to understand, calculate and interpret.
- It should be based on all the observations of the data.
- It should be amenable for further mathematical calculations.
- It should be least affected by the fluctuations of the sampling.
- It should not be unduly affected by the extreme values.

3.3. Arithmetic mean

Problem-1

The following data represent the number of books issued in a school library on selected from 7 different days 7, 9, 12, 15, 5, 4, 11 find the mean number of books.

$$\textbf{Mean (or) Arithmetic Mean (or) Average } (\bar{X}) = \frac{\sum X}{N}$$

$$= \frac{63}{7}$$

$$\textbf{Mean} = 9$$

Problem-2

Calculate arithmetic mean for the following frequency distribution.

C.I:	100-110	110-120	120-130	130-140	140-150
F:	45	20	35	40	20

C.I	**F**	**X=mid C.I = $\frac{(UL+LL)}{2}$**	**FX**
100-110	45	105	4725
110-120	20	115	2300
120-130	35	125	4375
130-140	40	135	5400
140-150	20	145	2900
Total	**160**	**-**	**19700**

$(\bar{X})$ = $\frac{\sum FX}{N}$; here N=∑F

$(\bar{X})$ = $\frac{19700}{160}$

$(\bar{X})$ = 123.12

Problem-3

The following table gives Return on Investment of 60 companies. Analyze and Calculate Arithmetic mean & Median.

(Return on Investment) Class interval	**(No. of companies) Frequency (f)**
0- 5	10
5-10	25
10-15	12
15-20	8
20-25	5

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

Class Interval	X =MID C.I = (U+L)/2	F	FX
0 - 5	**2.5**	10	**25**
5 – 10	**7.5**	25	**187.5**
10 – 15	**12.5**	12	**150**
15 – 20	**17.5**	8	**140**
20 - 25	**22.5**	5	**112.5**
-	**-**	**60**	**615**

Mean $\bar{X}$ = $\frac{\sum FX}{N}$ = $\frac{615}{60}$ = 10.25

3.4. Median

Example 1: Find the Median of the following data.

70, 60, 75, 90, 65, 80, 42, 65, 72

Solution:

Arrange the data in ascending order, 42, 60, 65, 65, 70, 72, 75, 80, 90

Formula:

$$\text{Median} = \frac{\text{Size of (N+1)}}{\text{2th item}}$$

Here N = 9

$$\text{Median} = \frac{\text{Size of (9+1)}}{\text{2th item}}$$

Median = size of (5) th item

Median = 70

Example2. Find the Median of the following data.

13, 17, 11, 19, 9, 16, 14

Solution:

Arrange the data in ascending order, 9, 11, 13, 14, 16, 17, 19

Formula:

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

Here N = 7

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

Median = size of (4) th item

Median = 14

Example 3

Calculate median for the following frequency distribution.

X:	12	14	17	20	23
F:	6	5	9	4	8

Median = $\frac{(N+1)}{2}$th value in cumulative frequency

X	F	CF
12	6	6
14	5	11
17	9	20
20	4	24
23	8	32

Median = $\frac{(N+1)}{2}$th value in cumulative frequency

Median = $\frac{(32+1)}{2}$th value in cumulative frequency

Median = 16.5 th value in cumulative frequency

Median= 17

Example 4

Calculate median for the following frequency distribution.

Marks(C.I):	0-10	10-20	20-30	30-40	40-50	50-60
No of students (F):	10	18	12	24	16	20

Class interval	**Frequency**	**Cumulative frequency**
0-10	10	10
10-20	18	28
20-30	12	40
30-40	24	64
40-50	16	80
50-60	20	100
Total	100	-

$$\frac{\text{Size of (N+1)}}{\text{2th item}} \quad \frac{N}{2} \quad \frac{100}{2} = 50$$

Median = L $+\frac{\left(\frac{N}{2}\right)-Before\ cummulative\ frequency}{f}$ X C

Median = 30 $+\frac{50-40}{24}$ X 10

Median = 30 + 4.16

Median = 34.16

Example 5

The following table gives Return on Investment of 60 companies. Analyze and Calculate Arithmetic mean & Median.

(Return on Investment) Class interval	**(No. of companies) Frequency (f)**
0- 5	10
5– 10	25
10– 15	12
15– 20	8
21- 25	5

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

CLASS INTERVAL	**X =MID C.I =** $\frac{(U+L)}{2}$	**F**	**FX**
0 - 5	**2.5**	10	**25**
5 – 10	**7.5**	25	**187.5**
10 – 15	**12.5**	12	**150**
15 – 20	**17.5**	8	**140**
20 - 25	**22.5**	5	**112.5**
-	**-**	**60**	**615**

Mean $\bar{X}$ = $\frac{\sum FX}{N}$ = $\frac{615}{60}$ = 10.25

Calculation of Median

Frequency (f)	**Cumulative Frequency(CF)**
10	10
25	35
12	47
8	55
5	60

$\frac{N}{2} = \frac{60}{2} = 30$

L=5 F=25 CF=35 BCF=10 C= CLASS WIDTH =5

Median = L + $\frac{\left(\frac{N}{2}\right) - Before\ cummulative\ frequency}{f}$ X C

= 5 + $\frac{30-10}{25}$ X 5

= 5 +4

MEDIAN = 9

Example: 6

Calculate Median for the following frequency distribution.

C.I :	100-110	110-120	120-130	130-140	140-150
F:	46	20	35	40	20

Solution:

S-1: Cumulative Frequency

S-2: N/2

S-3: Marking from CF

S-4: Formula

Median = L + $\frac{\frac{N}{2} - bcf}{f}$ X C

C.I	F	CUMULATIVE FREQUENCY(CF)
100-110	45	45
110-120	20	65
120-130	35	100 *
130-140	40	140
140-150	20	160

N=160

$\frac{N}{2} = \frac{160}{2} = 80$

L=120, F=35, CF=100, BCF=65, C=10

Median = L + $\frac{\frac{N}{2} - bcf}{f}$ X C = 120 + $\frac{80-65}{35}$ X 10

$$\text{Median} = \frac{120 + 150}{35}$$

Median = 120 + 4.29

Median = 124.29

Check your progress-1

True/False

a. Human mind is incapable of remembering the entire mass of unwieldy data.

b. A measure of central tendency is not a representative value of the entire group of data.

c. In practical situations one need to have a single value to represent each variable in the whole set of data.

d. Statistical average should be unduly affected by the extreme values.

e. Statistical average should easy to understand, calculate and interpret

3.5. Mode

Problem-1

Calculate mode for the following data:45,56,78,36,25,36,67 and 100

Mode= the maximum repeated value

Mode=36.

Problem-2

Calculate mode for the following data.45,56,55,36,67,67,78,36,25,36,67 and 100

Mode= the maximum repeated value

Mode= 36 and 67 (Bi-modal)

Problem-3

Calculate mode for the following data.

X :	56	45	53	30	100	25
F:	19	26	35	23	10	15

Mode= The X value with highest frequency.

Mode= 53.

Problem-4

Calculate mode for the following data.

Marks(C.I):	0-10	10-20	20-30	30-40	40-50	50-60
No of students F:	10	18	12	24	16	20
				f0	f1	f2

The maximum frequency lies between 30 and 40.

So

Mode= L + $\frac{f1-f0}{2f1-f0-f2}$ X C

Mode= 30 + $\frac{24-12}{2(24)-12-16}$ X 10

Mode= 30 + $\frac{120}{20}$

Mode= 30+ 6

Mode= 36

Problem-5

Calculate Mode for the following frequency distribution.

C.I :	100-110	110-120	120-130	130-140	140-150
F:	10	18	12	24	16
			F0	F1	F2

Solution:

Select maximum frequency and give name F1

F1= 50

Predecessor of F1 is F0

Successor of F1 is F2

Mode = L + $\frac{f1-f0}{2f1-f0-f2}$ X C

Mode = 130 + $\frac{50-35}{100-35-20}$ X 10

Mode = 130 + $\frac{150}{45}$

Mode = 130 + 3.33

Mode= 133.33

3.6. Empirical Relationship among mean, median and mode

A frequency distribution in which the values of arithmetic mean, median and mode coincide is known of symmetrical distribution, when the values of mean, median and mode are not equal the distribution is known as asymmetrical or skewed. In moderately skewed asymmetrical distributions a very important relationship exists among arithmetic mean, median and mode.

Karl Pearson has expressed this relationship as follows

Mode = 3 Median – 2 Arithmetic Mean

Example

In a moderately asymmetrical frequency distribution, the values of median and arithmetic mean are 72 and 78 respectively; estimate the value of the mode.

Solution:

The value of the mode is estimated by applying the following formula:

Mode = 3 Median – 2 Mean = 3 (72) – 2 (78)

= 216 - 156 = 60

Mode = 60

Let us sum up

Measures of central tendency are measures used to represent a midline of a data set. The most commonly used measures of central tendency are mean, mode and median. Mean is a calculated value that lies at the centre of a data set. In this unit you have learned about the characteristics of a good statistical average and discussion about different measures of central tendencies like mean, median and mode.

Check your progress-2

1. The median of the variate values 11, 7, 6, 9, 12, 15, 19 is_____.

2. The maximum repeated value is called_____.

3. For symmetrical data mean, median and mode are_____.

Glossary

Mean : The most popular and widely used measure of representing the entire data by one value is known as arithmetic mean. Its value is obtained by adding together all the items and by dividing this total by the number of items.

Median: Median is that value of the variable which divides the group into two equal parts, one part comprising all values greater than, and the other all values less than the median

Mode : Mode is the value that appears most frequently in a series i.e. it is the value of the item around which frequencies are most densely concentrated.

Model Questions

1. What is Arithmetic mean?
2. Find the AM of the following data 25,32,28,34,24,31,36,27,29,30.
3. Find the Mean to the following distribution.

Marks(C.I):	0-20	20-40	40-60	60-80	80-100
Frequency :	3	17	27	20	9

4. Calculate Median for the given data

Value	10	12	15	20	50
Frequency	2	3	10	8	2

5. Estimate Arithmetic mean, Median & Mode from the following data:

200	210	208	160	220	250

6. Determine Arithmetic mean, Median & Mode from the following data:

9	27	18	54	45	72	36	63	81

7. From the following data Calculate Arithmetic mean, Median and mode.

Marks	0-10	10-20	20-30	30-40	40-50	50-60
No. of students (f)	5	10	25	30	20	10

8. Compute Arithmetic mean, Median and mode.

Wages (Rs)	20-30	30-40	40-50	50-60	60-70
No. of labourers (f)	3	5	20	10	5

9. The following data relates to the IQ level of 100 students in a test as follows. Estimate Arithmetic Mean and Median.

(IQ level) Class interval	(No. of students) Frequency (f)
40 – 50	10
50 – 60	20
60 – 70	20
70 – 80	15
80 - 90	15
90 - 100	20

Answer to check your progress-1

a-True

b-False

c-True

d-False

e-True

Answer to check your progress-2

(1). 11

(2). Mode

(3). Equal

Suggested Readings

1. R.S.N. Pillai, V. Bagavathi," Statistics", S.Chand Limited, 7thEd,2008
2. Elhance , D.N. Fundamentals of Statistics. Allahabad: KitabMahal, (2007).

Unit -4
Measures of Dispersions

STRUCTURE

Overview

In this unit the different measures of dispersions like range, quartile deviation, standard deviation and coefficient of variation and also the problems related to the measures of dispersions has been clearly explained.

Objectives

After completion of this unit, you will be able to:

- Provide the importance of the concept of variability (dispersion)
- Describe the spread - range and standard deviations
- Calculate the coefficient of variation

4.1. Introduction

The measures of central tendency describes the central part of values in the data set appears to concentrate around a central value called average. But these measures do not reveal how these values are dispersed (spread or scattered) on each side of the central value. Therefore while describing data set it is equally important to know how

for the item in the data are close around or scattered away from the measures of central tendency.

4.1.1. Characteristics of a good Measure of Dispersion

An ideal measure of dispersion is to satisfy the following characteristics.

(i) It should be based on all observations in the data set.

(ii) It should be well defined without any ambiguity.

(iii) It should be easy to understand and compute.

(iv) It should be capable of further mathematical treatment.

(v) It should not be affected by fluctuations of sampling.

4.2. Range

Range is defined as difference between the largest and smallest observations in the data set. Range(R) = Largest value in the data set (L) –Smallest value in the data set(S)

Range = L – S

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

Example: 1

Calculate Range for the following data: 45,56,78,36,25,36,67 and 100.

Largest value=100

Smallest value=25

R = L – S

R = 100 – 25

Range (R) = 75

Example: 2

Calculate Range for the following data:

X:	56	89	100	62	98	50
F:	45	60	10	62	20	45

Largest value= 100

Smallest value=50

R = L – S

R = 100 - 50

Range (R) = 50

Example: 3

Calculate Range.

C.I:	65-75	75-85	85-95	55-65	45-55	95-105	105-115
F:	23	14	50	33	28	36	40

Largest value = 115

Smallest value = 45

Range = L – S

Range = 115-45

Range = 70

4.3. Quartile Deviation

The Quartile Deviation can be defined mathematically as half of the difference between the upper and lower quartile. Here, quartile deviation can be represented as Q.D; Q3 denotes the upper quartile and Q1 indicates the lower quartile.

$$\text{Formula for Q.D} = \frac{(Q3 - Q1)}{2}$$

$$\text{Coefficient of Q.D} = \frac{(Q3 - Q1)}{(Q3 + Q1)}$$

To find $Q1$

$$\text{Q1} = \frac{\text{Size of (N+1)}}{\text{4th item}}$$

To find $Q3$

$$\text{Q3} = \frac{\text{Size of 3 (N+1)}}{\text{4th item}}$$

Problem: 1

From the following data, compute the Quartile deviation and coefficient of quartile deviation.20, 28, 40, 12, 30, 15, 50.

Solution:

Arrange the data in ascending order, 12, 15, 20, 28, 30, 40, 50

$$\text{Formula for Q.D} = \frac{\text{(Q3-Q1)}}{2}$$

$$\text{Coefficient Q.D} = \frac{\text{(Q3-Q1)}}{\text{(Q3+Q1)}}$$

Here N= 7

$$Q1 = \frac{\text{Size of } (7+1)}{4^{th} \text{ item}} = \text{size of (2)nd item} = 15$$

To find $Q3$

$$Q3 = \frac{\text{size of } 3(7+1)}{4 \text{ th item} = \text{size}} = \text{size of (6) th item} = 40$$

Inter Quartile range = 40 - 15 = 25

$$Q.D = \frac{(40-15)}{2} = \frac{25}{2} = 12.5$$

$$\text{Coefficient of Q.D} = \frac{(40-15)}{(40+15)} \quad \frac{25}{55} = 0.45$$

4.4. Standard Deviation

'Standard deviation is the positive square root of average of the deviations of all the observation taken from the mean.' It is denoted by a Greek letter σ.

For raw data

Standard deviation (σ) = $\sqrt{\frac{\sum(X-\overline{X})^2}{n}}$

For discrete data

Standard deviation (σ) = $\sqrt{\frac{\sum f(X-\overline{X})^2}{N}}$

For continuous data

Standard deviation (σ) = $\sqrt{\frac{\sum f(M-\overline{X})^2}{N}}$

Check your progress-1

True/False

a. The measures of central tendency describes the central part of values in the data set appears to concentrate around a central value called average.

b. Range is defined as difference between the largest and smallest observations in the data set.

c. The Quartile Deviation can be defined mathematically as half of the difference between the upper and lower quartile.

d. Coefficient of variation is not a type of relative measure of dispersion.

e. Standard deviation is not the positive square root of average of the deviations of all the observation taken from the mean.

4.5. Coefficient of Variation

Coefficient of variation is a type of relative measure of dispersion. It is expressed as the ratio of the standard deviation to the mean. The coefficient of variation is a dimensionless quantity and is usually given as a percentage. It helps to compare two data sets on the basis of the degree of variation.

The coefficient of variation can be determined for both a sample as well as a population. In industries such as finance, the coefficient of variation is used to help investors assess the risk to reward ratio.

Coefficient of variation (C.V) = $\frac{\sigma}{\bar{X}}$ X 100 %

Problem: 1

Calculate mean and standard deviation for the following student's height (cm). Also find coefficient of variation.

140,150,155,160 and 165.

Mean $\bar{X}$ = $\frac{\sum X}{n}$

Mean $\bar{X}$= $\frac{770}{5}$

Mean $\bar{X}$ = 154

X	$(X - \overline{X})$ $(X - 154)$	$(X - \overline{X})^2$
140	-14	196
150	-4	16
155	1	1
160	6	36
165	11	121
Total	-	370

Standard deviation (σ) = $\sqrt{\frac{370}{5}}$

Standard deviation (σ) = $\sqrt{74}$

Standard deviation (σ) = 8.60

Coefficient of variation (c.v) = $\frac{\sigma}{\bar{X}}$ X 100 %

Coefficient of variation (c.v) = $\frac{8.60}{154}$ X 100 %

Coefficient of variation (c.v) = 5.58 %

Problem: 2

Calculate mean, standard deviation and coefficient of variation for the following discrete frequency distribution.

X: 10 15 25 40 50 60

F: 6 5 10 11 8 10

X	F	fx	$(X-\overline{X})$ $(X-36.5)$	$(X-\overline{X})^2$	$f(X-\overline{X})^2$
10	6	60	-26.5	702.25	4213.5
15	5	75	-21.5	462.25	2311.25
25	10	250	-11.5	132.25	1322.5
40	11	440	3.5	12.25	134.75
50	8	400	13.5	182.25	1458
60	10	600	23.5	552.25	5522.5
Total	50 = N	1825	-	-	14962.5

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

Mean $\bar{X} = \frac{1825}{50}$

Mean $\bar{X}$ = 36.5

Standard deviation (σ) = $\sqrt{\frac{\sum f(X-\overline{X})^2}{N}}$

Standard deviation (σ) = $\sqrt{\frac{14962.5}{50}}$

Standard deviation (σ) = $\sqrt{299.25}$

Standard deviation (σ) = 17.29

Coefficient of variation (c.v) = $\frac{\sigma}{\bar{X}}$ X 100 %

Coefficient of variation (c.v) = $\frac{17.29}{36.5}$ X 100 %

Coefficient of variation (c.v) = 47.36 %

Problem: 3

Calculate mean, standard deviation and coefficient of variation for the following frequency distribution.

C.I: 0-10 10-20 20-30 30-40 40-50

F: 4 6 10 20 10

C.I	X	F	FX	$f(X-\overline{X})^2$ $f(X-30.2)^2$
0-10	5	4	20	2540.16
10-20	15	6	90	1386.24
20-30	25	10	250	270.40
30-40	35	20	700	460.80
40-50	45	10	450	2190.40
Total	-	50	1510	6848.00

Mean $\bar{X} = \frac{\sum fX}{N} = \frac{1510}{50} = 30.2$

Standard deviation (σ) = $\sqrt{\frac{\sum f(X-\overline{X})^2}{N}} = \sqrt{\frac{6848}{50}} = \sqrt{136.96} = 11.70$

Coefficient of variation (c.v) = $\frac{\sigma}{\bar{X}}$ X 100 % = $\frac{11.70}{\overline{30.2}}$ X 100 = 38.74%

Let us sum up

Dispersion in statistics refers to the measure of the variability of data or terms. Such variability may give random measurement errors where some of the instrumental measurements are found to be imprecise. It is a statistical way of describing how the terms are spread out in different data sets.

In this unit you have learned about the different measures of dispersions like range, quartile deviation, standard deviation and coefficient of variation and also the problems related to the measures of dispersions.

Check your progress-2

(1).The difference between highest value and lowest value is called_____.

(2). The expansion of QD is_____.

(3). The square root of variance is called_____.

Glossary

Range: Difference between largest and smallest data.

Coefficient of range: The measure of the distribution based on range is the coefficient of range also known as range coefficient of dispersion.

Quartile Deviation: Is the half of the difference between the first and third quartiles.

Coefficient of quartile Deviation: The coefficient of quartile deviation allows you to compare dispersion for two or more sets of data.

Standard Deviation: The standard deviation is simply the square root of the variance.

Model Questions

1. Calculate the Range, standard deviation and Mean deviation from the following data.

X	0-5	5-10	10-15	15-20	20-25	25-30	30-35
f	1	2	5	14	10	9	2

2. Interpret by calculating Range, standard deviation and Mean deviation from the following data.

Class	0-10	10-20	20-30	30-40	40-50
Frequency (f)	5	8	15	16	6

3. The market capitalization of top companies were given. Calculate Arithmetic Mean, Median, Mode, Range, Standard deviation and mean deviation

Company	**Market capitalization (Rs. In thousand Cr)**
RIL	163
ONGC	162
INFOSYS	112
NTPC	108

TCS	106
Bharti Airtel	90
Wipro	78
ITC	70
Indian oil	66
ICICI bank	62

4. Calculate Arithmetic Mean, Median, Range, Standard deviation and Mean deviation,

Marks	0-10	10-20	20-30	30-40	40-50
No. of students (f)	3	5	9	3	2

5. Estimate Arithmetic Mean, Median, Range, Standard deviation and Mean deviation

Marks	0-5	5-10	10-15	15-20	20-25	25-30	30-35
No. of students (f)	1	2	5	14	10	9	2

6. Calculate Arithmetic Mean, Median, Range, Standard deviation and Mean deviation

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency (f)	14	24	38	20	4

7. Calculate Arithmetic Mean, Median, Range, Standard deviation and mean deviation for the following data related to equity holdings of 20 billionaires. (Rs. Crores)

Equity holdings Class interval	**Frequency (f)**
2000 – 3000	2
3000 – 4000	5
4000 – 5000	6
5000 – 6000	4
6000– 7000	3

Answer to check your progress-1

a-True

b-True

c-True

d-False

e-False

Answer to check your progress-2

(1). Range

(2). Quartile Deviation

(3). Standard deviation

Suggested Readings

1. Richard I. Levin and David S. Rubin, Statistics for Management. Prentice Hall of India Pvt. Ltd., New Delhi(1996).
2. Bernsen M.L.And Levine D.MBasic Business Statistics, Prentice-Hall Englewood Cliffs New jersey(1996).

Block-2 : Introduction

Block-2: Probability Theory has been divided in to four Units (Unit-5 to Unit-8)

Unit- 5: Introduction to Probability deals with Introduction and Definition of probability, Axioms on probability, Sample Space and the Types of Events.

Unit-6: AdditionTheorem of Probability explains about Introduction and Addition rule on probability and also the Application problems.

Unit-7: Conditional and Multiplication Theorem of Probability describes about the Introduction and Conditional rule on probability, Multiplication rule on probability and also the Application problems.

Unit-8: Bayes' Theorem of Probability presents about the Introduction, Conditional rule on probability, Bayes' Theorem of Probability, Multiplication rule on probability and also the Application problems.

In all the units of Block -2: **Probability Theory**, the Check your progress, Glossary, Answers to Check your progress and Suggested Reading has been provided and the Learners are expected to attempt all the Check your progress as part of study.

Unit-5
Introduction to Probability

STRUCTURE

Overview

In this unit the different definitions of probability, rules and regulation will be followed in probability and the different types of events like independent events, dependent events etc. has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To know the definition of probability
- To understand the axioms on probability
- To learn about the elements of probability

5.1. Introduction

Every scientific experiment conducted to investigate the patterns in natural phenomenon may result with “events” which may or may not happen. Most of the events in real life have uncertainty in their happening.

For example, Disintegration of a given atom of radium in a given time interval may or may not disintegrate; A plant may or may not be infected by species during rainy season; The event of increase in the gold price

under an economic condition in a country; A drug administered to a cancer patient for curing a disease in a period of time.

In all these cases, there is an amount of uncertainty prevails. Even for a student, asking a particular question in the examination from a particular portion of a subject is uncertain. Yet, the student is compelled to take a decision during the preparation of examination, whether to go for an in-depth study or leaving the question in choice.

In a nutshell, one has to take a wise decision under the conditions of uncertainty. In such a situation, knowledge about the chance or probability for occurrence of an event of interest is vital and calculation of probability for happening of an event is imperative.

It is very much essential to determine a quantitative value to the chance or probability for the occurrence of random events in many real life situations.

Experiments and Events.

- An EXPERIMENT is any activity with an observable result. Tossing a coin, rolling a die or choosing a card are all considered experiments.
- An OUTCOME (or SAMPLE POINT) is the result of a the experiment. The set of all possible outcomes or sample points of an experiment is called the SAMPLE SPACE.
- An EVENT is a subset of the sample space. Example - A coin is tossed and the side noted and a card is drawn and the color noted. In probability, the set of outcomes from an experiment is known as an Event. So say for example you conduct an experiment by tossing a coin. The outcome of this experiment is the coin landing ‘heads’ or ‘tails’. These can be said to be the events connected with the experiment. So when the coin lands tails, an event can be said to have occurred.

5.2. Definition of probability

“Probability of a given event is defined as the expected frequency of occurrence of the event among events of a like sort.”

A probability gives the likelihood that a defined event will occur. It is quantified as a positive number between 0 (the event is impossible) and 1 (the event is certain). Thus, the higher the probability of a given event, the more likely it is to occur.

If the sample space S, of an experiment is finite with all its elements being equally likely, then the probability for the occurrence of any event, A, of the experiment is defined as

$$\text{Probability} = \frac{N(E)}{N(S)}$$

Here N (E) = the no of likelihood events

N (S) = the no of sample space.

Example:

A coin tossed at once. What is the probability for getting a head?

Sample space (S) = { head, tail}

N(S) = 2 , N (E) = 1

Probability = $\frac{1}{2}$

5.3. Axioms on probability

Probability –the chance of happening an event

$$\text{Probability } P(E) = \frac{N(E)}{N(S)}$$

Here N (E) = the no of likelihood events

N (S) = the no of sample space.

The probability value always lies between 0 and 1.

The probability of sure event is one.

The probability of an impossible event is zero.

$P(E_1 \cup E_2 \cup \ldots..\cup E_N) = P(E_1) + P(E_2) + \ldots.. + P(E_N)$

5.4. Sample Space

Sample space: The set of all possible outcomes of a random experiment is called the sample space of the experiment and is usually denoted by S (or Ω). If S contains only finite number of elements, it is termed as finite sample space. If S contains countable number of elements, S may be called as countable sample space or discrete sample space. Otherwise, S is called an uncountable sample space.

Consider the random experiment of tossing a coin once “Head” and “Tail” are the two possible outcomes. The sample space is S = {H, T}.

Suppose that a study is conducted on all families with one or two children. The possible outcomes, in the order of births, are: boy only, girl

only, boy and girl, girl and boy, both are boys and both are girls. Then, the sample space is S = {b,g,bg,gb,bb,gg}. It is also a finite sample space. Here, 'b' represents the child is a boy and 'g' represents the child is a girl.

Check your progress-1

True/False

a. A probability gives the likelihood that a defined event will occur.

b. As opposed to a simple event, if there is more than one sample point on a sample space, such an event is called Compound Event.

c. Most of the events in real life have not uncertainty in their happening.

d. Exhaustive events may or may not be equally likely and mutually exclusive.

e. Mutually exclusive events always have a different outcome.

5.5. Types of Events

Mutually exclusive event

Two events are said to be mutually exclusive events when both cannot occur at the same time. Mutually exclusive events always have a different outcome. Two simple events are always mutually exclusive, whereas two compound events may or may not be.

If A and B are two events, then

$(A \cap B) = \emptyset$

$P (A \cap B) = 0$

$P (A \cup B) = P (A) + P (B)$

Two events are said to be mutually exclusive events when both cannot occur at the same time. Mutually exclusive events always have a different outcome. Such events are so that when one happens it prevents the second from happening. For example, if the coin toss gives you a "Head" it won't give you a "Tail". These are mutually exclusive events.

Exhaustive event

Exhaustive Events: A set of events is said to be exhaustive, if it includes all the possible events. For example, in tossing a coin there are two exhaustive cases either head or tail and there is no other possibility.

In probability, exhaustive is a condition of two or more events which serves a great role in finding the probability as it changes if the events are exhaustive or not. Two or more events are said to be exhaustive if there is a certain chance of occurrence of at least one of them when they are all considered together.

Exhaustive events may or may not be equally likely and mutually exclusive. There is no any particular formula to find the exhaustive events, but sure we are able to tell that events are exhaustive or not after go through below examples.

Consider the experiment of throwing a fair die and the event of getting a number less than or equal to 6. Now this particular event is certain in spite of being a single event whenever the die is thrown and hence this single event is also exhaustive.

Consider the die throwing experiment. Let the events be getting a number which is multiple of 2 {2, 4, 6} and event of getting factors of 6 {1, 2, 3, 6} and event of getting a {5}. Clearly all these events together are exhaustive.

Equally likely events

Equally likely events are events that have the same theoretical probability (or likelihood) of occurring. Getting a 3 on the toss of a die and getting a 5 on the toss of a die are equally likely events, since the probabilities of each event are equal.

Getting an even number on the toss of a die and getting an odd number on the toss of a die are equally likely events, since the probabilities of each event are equal.

Getting a 1, 2 or 3 on the toss of a die and getting a 4, 5 or 6 on the toss of a die are equally likely events, since the probabilities of each event are equal.

Equally Likely Events Outcomes of a trial are said to be equally likely if taking in to consideration all the relevant evidences, there is no reason to expect one in preference to the others. (i.e.) Two or more events are said to be equally likely if each one of them has an equal chance of occurring

Simple and compound event

- **Simple Event**

If the event *E* has only one sample point of a sample space, it is called a simple event or an Elementary Event. It is an event that consists of

exactly one outcome. Let us understand this with an example. Say you throw a die, the possibility of 2 appearing on the die is a simple event and is given by E = {2}.

- **Compound Event**

As opposed to a simple event, if there is more than one sample point on a sample space, such an event is called Compound Event. It involves combining two or more events together and finding the probability of such a combination of events.

For example, let us take another example. When we throw a die, the possibility of an even number appearing is a compound event, as there is more than one possibility, there are three possibilities i.e. E = {2,4,6}.

Independent and Dependent events

- Independent Events Several events are said to be independent if the happening of an event is not affected by the happening of one or more events. Example 1. When two seeds are sown in a pot, one seed germinates. It would not affect the germination or non-germination of the second seed. One event does not affect the other event.
- Dependent Events If the happening of one event is affected by the happening of one or more events, then the events are called dependent events. Example If we draw a card from a pack of well shuffled cards, if the first card drawn is not replaced then the second draw is dependent on the first draw

Let us sum up

In this unit, you have learned about the following:

- “Probability of a given event is defined as the expected frequency of occurrence of the event among events of a like sort.”
- The probability value always lies between 0 and 1.
- The probability of sure event is one.
- The probability of an impossible event is zero.
- Different types of events: Equally likelihood events, Independent events, Dependent events, and mutually exclusive events.

Check your progress-2

1. Which one of the following is not related to random experiment?
   (a) Outcomes are predictable in advance
   (b) outcomes is unknown, in advance
   (c) Experiment repeatable finite number of times
   (d) experiment is repeatable any number of times.
2. Mathematical probability may also be termed as
   (a) Statistical probability
   (b) classical probability
   (c) Empirical probability
   (d) None of the above
3. In rolling of a die until 4 appears, the sample space is
   (a) A null set
   (b) a countable finite set
   (c) a countable infinite set
   (d) an uncountable set

Glossary

Experiment – Events - Sample space - Equally likelihood events - Independent events - Dependent events - Mutually exclusive events.

Model Questions

1. Define the term probability and discuss the axioms on probability.
2. Explain the different types of events.
3. If E1 and E2 are two mutually exclusive events and Given that
   P(E2) = 0.5 and P(E1UE2) = 0.7 then find P(E1).
4. Find the probability that a leap year selected at random will contain 53 Fridays?
5. Two coins are tossed simultaneously. What is the probability of getting exactly two heads?

Answers to check your progress-1

a-True

b-True

c-False

d-True

e-True

Answers to check your progress-2

1. Outcomes are predictable in advance
2. Classical probability
3. A countable infinite set

Suggested readings

1. Keller. G, ”Statistics for Management”, Cengage Learning, 1st Ed,2009.
2. J. K Sharma, “Business Statistics”, Pearson, 2nd Ed,2010.

Unit-6

Addition Theorem of probability

STRUCTURE

Overview

In this unit the concept of additional rule of probability and related problems for independent, dependent and mutually exclusive events and the application problems has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the addition theorem of probability.
- To identify the addition theorem of probability for mutually exclusive events and independent events.
- To know the applications of addition theorem.

6.1. Introduction

In the 16^{th} Century, a gambler named Chevalier de Mere wanted to find out about the chances of a number appearing on the roll of dice, so he decided to approach a French Philosopher and Mathematician Blaise Pascal to solve the dice problem.

Blaise Pascal got interested in the concept of possibility and so he discussed it with another French Mathematician, Pierre de Fermat. Both the Mathematicians started working on the concept of probability separately.

Later, J. Cardan, an Italian Mathematician, wrote the first book named 'Book on Games of Chance' in 1663 that deals with the inception of probability. This caught the attention of some of the great Mathematicians J. Bernoulli, P. Laplace, A.A Markov and A.N.Kolmogorov.

Out of all the Mathematicians, A.N.Kolmogorov, a Russian mathematician, treated probability as a function of outcomes of the experiment. With the help of this concept, we can find the probability of events allied with discrete sample spaces. This also establishes the concept of conditional probability, which is important for the perception of Bayes' Theorem, multiplication rule, and independence of events. In 1812, Laplace also came up with 'Theory Analytique des Probabilities', which is considered as the greatest contribution by an individual to the theory of probability. The deductions and reasoning introduced by these mathematicians related to probability are now being used in Biology, economics, genetics, physics, sociology, etc.

6.2. Addition rule on probability

If A and B are any two events in a random experiment, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

If A, B and C are any three events, then

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

(i) If A and B are independent event $P(A \cap B) = P(A) \cdot P(B)$

$$P(A \cup B) = P(A) + P(B) - [P(A) \cdot P(B)]$$

(ii) If A and B are mutually exclusive event $P(A \cap B) = 0$

$$P(A \cup B) = P(A) + P(B)$$

6.3. Application problems

Problem: 1

A card is drawn at random from a well shuffled pack of cards. What is the probability that it is heart or queen?

Let A be the event of getting a heart card.

Let B be the event of getting a queen card.

$$P(A) = \frac{13C1}{52C1} = \frac{13}{52} = ¼ = 0.25$$

$$P(B) = \frac{4C1}{52C1} = \frac{4}{52} = \frac{1}{13} = 0.076$$

$$P(A \cap B) = \frac{1C1}{52C1} = \frac{1}{52} = 0.019$$

$P(A \cup B) = P(A) + P(B) – P(A \cap B)$

$P(A \cup B) = 0.25 + 0.076 – 0.019$

$P(A \cup B) = 0.307$

Problem: 2

One card is drawn from a standard pack of 52. What is the probability of getting either a king or queen? Use addition law of probability for independent events.

A= getting a king

$$P(A) = \frac{4C1}{52C1} = \frac{1}{13} = 0.077$$

B= getting a queen

$$P(B) = \frac{4C1}{52C1} = \frac{1}{13} = 0.077$$

If A and B are independent event $P(A \cap B) = P(A) \cdot P(B)$

$P(A \cup B) = P(A) + P(B) – P(A \cap B)$

$P(A \cup B) = P(A) + P(B) – P(A \cap B)$

$= 0.077 + 0.077 – (0.077 \times 0.077)$

$= 0.154 – 0.059$

$P(A \cup B) = 0.095$

Problem: 3

Use addition law of probability for independent events. First bag contains 4 white balls and 2 black balls and second bag contains 3 white balls and 3 black balls. A bag is selected at random, and a ball is then drawn at random from a bag chosen. What is the probability that the ball drawn is white?

I bag = 4w,2b

II bag = 3w,3b

If A and B are independent event P(A∩B) = P(A) . P(B)

P (AUB) = P(A) + P(B) – P(A∩B)

$$\mathrm{P\ (AUB)} = \left(\frac{4c1}{6c1}\right) + \left(\frac{3c1}{6c1}\right) - \left\{\left(\frac{4c1}{6c1}\right) \times \left(\frac{3c1}{6c1}\right)\right\}$$

P (AUB) = 0.67 + 0.5 – { 0.67 X 0.5 }

P (AUB) = 1.17 – 0.335

P (AUB) = 0.835

P (AUB) = P (A) + P (B) – P (A∩B)

P (AUBUC) = P (A) + P (B) + P (C) – P (A∩B) – P (A∩C) – P (B∩ C) + P (A ∩ B ∩ C)

Let us sum up

In this unit, you have learned about the concept of additional rule of probability and related problems for independent, dependent and mutually exclusive events and also the application problems.

Check your progress

1. The probability of the entire sample space is _____
2. On throwing the single die, then the event of getting odd number or even number are _____ event.
3. Probability of getting a Monday in a week is _____

Glossary

Experiment – Events - Sample space – Addition theorem - Independent events - Dependent events - Mutually exclusive events.

Model Questions

1. Explain addition theorem of probability.
2. A herd contains 30 cows numbered from 1 to 30. One cow is selected at random. Find the probability that number of the selected cow is a multiple of 5 or 8.
3. A number was drawn at random from the number 1 to 50. What is the probability that it will be a multiple of 2 or 3 or 10?
4. A bag contains 30 balls numbered from 1 to 30. One ball is drawn at random; find the probability that the number of the ball will be multiple of 5 or 9.

5. A person can hit a target in 3 out 4 shots, whereas another person can hit the target in 2 out of 3 shots. Find the probability of the targets being hit at all when they both try.

Answer to Check your progress

1. One

2. Mutually exclusive

3. 1/7

Suggested readings

1. Bernsen M.L. and Levine D.M, Basic Business Statistics, Prentice-Hall Englewood Cliffs New jersey(1996).
2. Webster. A Applied Statistics For Business and Economics, Homewood, IL: Irwin(1992).

Unit-7

Conditional and Multiplication Theorem of Probability

STRUCTURE

Overview

In this unit the meaning of conditional events and conditional probability and also the application problems from conditional probability and multiplication theorem has been clearly explained.

Objectives

After completion of this unit you will be able:

- To understand the multiplication theorem of probability.
- To identify the multiplication theorem of probability for independent events.
- To know the applications of conditional probability and multiplication theorem.

7.1. Introduction

Probability is a measure for quantifying the likelihood that events will occur. Probability of an event to happen lies between 0 and 1, where, 0 indicates impossibility and 1 indicates certainty. The higher the probability of an event, the more likely it is that the event will occur. Decisions are easy to take when we have several options and we choose only one on the basis of higher probability. Even if anybody is

not able to change his/she chance, he/she is able to guess which option is more likely to happen over the others. Probability is all around us and it does influence our lives every day. We should not ignore analysing small probabilities, they seem insignificant but oversight these may deprive us of happy ending. 'Life is a Probability and Probability is life'.

7.2. Conditional rule on probability

If A and B are the two events,

(i) P(A/B) is the conditional probability of A when B has already happened

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

(ii) P(B/A) is the conditional probability of B when A has already happened

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

7.3. Multiplication rule on probability

- Let A and B are dependent events

(i) $P(A \cap B) = P(A).\ P(B/A)$

P(B/A) is the conditional probability of B when A has already happened

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

(i.e) $P(A \cap B) = P(A).\ P(B/A)$

(ii) $P(A \cap B) = P(B).\ P(A/B)$

P(A/B) is the conditional probability of A when B has already happened

$$P(B/A) = \frac{P(A \cap B)}{P(B)}$$

(i.e) $P(A \cap B) = P(B).\ P(A/B)$

- If A and B are independent events, the multiplication theorem is

$$P(A \cap B) = P(A).P(B)$$

7.4. Application problems

Problem: 1

A bag contains 3 pink balls and 7 green balls. Two balls are taken out from the bag with replacement. Calculate the probability that both balls are pink.

Let A be the getting a pink ball from first drawn

Let B be the getting a pink ball from second drawn.

Here A and B are independent events.(with replacement)

$$P(A) = \frac{3C1}{10C1} = \frac{3}{10} = 0.3$$

$$P(B) = \frac{3C1}{10C1} = \frac{3}{10} = 0.3$$

the probability that both balls are pink

(i.e) $P(A \cap B) = P(A).P(B)$

$= 0.3 \text{ X } 0.3$

$P(A \cap B) = 0.09$

Problem: 2

A bag has 4 white cards and 5 blue cards. Two cards are drawn from the bag one by one without replacement. Estimate the probability of getting both cards white.

Let A be the event of getting a white card from I drawn.

Let B be the event of getting a white card from II drawn.

Here A and B are dependent events.(without replacement)

$$P(A) = \frac{4C1}{9C1} = \frac{4}{9} = 0.44$$

$$P(B) = \frac{3C1}{8C1} = \frac{3}{8} = 0.38$$

$P(A \cap B) = P(A).P(B) = 0.44 \text{ X } 0.38 = 0.17$

Problem: 3

One bag contains 4 white balls and 2 black balls. Another contains 3 white balls and 5 black balls. If one ball is drawn from each bag, compute the probability that both are white.

Let A be the event of getting a white ball from I bag.

Let B be the event of getting a white ball from II bag.

Here A and B are independent events.

$$P(A) = \frac{4C1}{6C1} = \frac{4}{6} = \frac{2}{3} = 0.67$$

$$P(B) = \frac{3C1}{8C1} = \frac{3}{8} = 0.38$$

$P(A \cap B) = P(A) \cdot P(B) = 0.67 \times 0.38 = 0.25$

Problem: 4

One bag contains 4 white balls and 2 black balls. Another contains 3 white balls and 5 black balls. If one ball is drawn from each bag, Estimate the probability that both are black.

Let A be the event of getting a black ball from I bag.

Let B be the event of getting a black ball from II bag.

Here A and B are independent events.

$$P(A) = \frac{2C1}{6C1} = \frac{2}{6} = \frac{1}{3} = 0.33$$

$$P(B) = \frac{5C1}{8C1} = \frac{5}{8} = 0.63$$

$P(A \cap B) = P(A) \cdot P(B) = 0.33 \times 0.63 = 0.208$

Problem: 5

One bag contains 4 white balls and 2 black balls. Another contains 3 white balls and 5 black balls. If one ball is drawn from each bag, find the probability that one is white and one is black.

I bag = 4w,2b

II bag = 3w,5b

Let A be the event of getting a White ball

Let B be the event of getting a black ball

$P(A \cap B)$ = (White from I bag& black from II bag) or (White from II bag& black from I bag)

$$= \left(\frac{4C1}{6C1} \times \frac{5C1}{8C1}\right) + \left(\frac{3C1}{8C1} \times \frac{2C1}{6C1}\right)$$

= (0.67 X 0.63) + (0.38 X 0.33)

= 0.4221 + 0.1254

P(A∩B)= 0.5475

Problem: 6

If a player plays a game of chance where he can win Rs.1000 with probability 0.5, win Rs.500 with probability 0.3 and lose Rs.3000 with probability 0.2, what is his expected gain in one play of the game?

Mathematical expectation E(X) = ∑X .P(X)

Mathematical expectation E(X) = (1000 X 0.5)+(500 x 0.3) – (3000 x 0.2)

E(X) = 500 + 150 – 600

E(X) = Rs. 50

Problem: 7

Use multiplication theorem of probability for independent events. A bag contains 8 white balls and 10 black balls. Two balls are drawn in succession. What is the probability that first is white and second is black?

- If A and B are independent events, the multiplication theorem is

P(A∩B) = P(A).P(B)

$$P(A) = \frac{8C1}{18C1} = 0.44$$

$$P(B) = \frac{10C1}{17C1} = 0.59$$

P (A∩B) = P(A).P(B) = 0.44 X 0.59 = 0.2596

Thus

- Conditional probability

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

- Multiplication rule on probability

P (A∩B) = P (A). P (B/A)

$$P(A \cap B) = P(B).\ P(A/B)$$

For independent events, the multiplication theorem is

$$P(A \cap B) = P(A).P(B)$$

Let us sum up

In this unit you have learned about the meaning of conditional events and conditional probability and also the application problems from conditional probability and multiplication theorem.

Check your progress

1. The sample size is denoted as

   (a) S (b) P (c) N (d) n

2. If A and B are mutually exclusive events them P(AUB) is equal to

   (a) P(A) + P(B) (b) P(A) – P(B)

   (c) P(A) + P(B) − P(A∩B) (d) P(A)P(B)

3. If A= {1, 2}; B = {3, 4, 5}; C= {5, 6} are events in S. Then the number sample point in S is

   (a)1 (b)4 (c) 3 (d) 6

Glossary

Experiment – Events - Sample space – conditional probability – Multiplication rule for probability - Independent events - Dependent events - Mutually exclusive events.

Model Questions

1. Discuss the conditional rule in probability.
2. Explain the multiplication theorem of probability.
3. A box contains 7 red and 3 white marbles. Three marbles are drawn from the box one after the other without replacement. Find the probability of drawing three marbles in the alternate colours with the first marble being red.
4. Three cards are drawn successively from a well-shuffled pack of 52 playing cards. Find the probability all three cards drawn successively is ace without replacing the card after each draw.

Answer to check your progress

1. n
2. P(A) + P(B)
3. 6

Suggested readings

1. N.D. Vohra, “Business Statistics”, Tata McGraw-Hill Education, 2ndEd,2018.
2. Berenson M., Levine D., Szabat K.A. and Krehbiel T.C. Basic Business Statistics: Concepts and Applications, Pearson Higher Education AU (2012).

Unit-8
Bayes' Theorem of probability

STRUCTURE

Overview

In this unit the definitions of Bayes theorem of probability and its application numerical problems has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the concept of Bayes theorem of probability.
- To apply the Bayes theorem in real life situation.

8.1. Introduction

In some cases, probability for the occurrence of an event of interest A may be difficult to compute from the given information. But, it may be possible to calculate its conditional probabilities P (A/B) and P (B/A) for some other event B of the same experiment. Then, P (A) can be calculated applying the law of total probability. This theorem is a prelude for Bayes' theorem.

In real life situations, decision making is an on-going process. Situations may arise where we are interested in an event on an on-going basis. Every time some new information may be available and based on this the probability of the event should be revised. This revision of probability

with additional information is formalized in probability theory in the theorem known as Bayes' Theorem.

8.2 Bayes' Theorem of probability

Let E_1, E_2, E_3 …, B_n be n mutually exclusive events such that where S is the sample space of the random experiment. If P(Ei) > 0 for i = 1, 2, …, n, then for any event A of the same experiment with P(A) > 0,

$$P(Ei / A) = \frac{P(Ei)\, P(A/Ei)}{\sum P(Ei)\, P(A/Ei)}$$

8.3 Application problem

A bag I contains 4 white and 6 black balls while another Bag II contains 4 white and 3 black balls. One ball is drawn at random from one of the bags, and it is found to be black. Find the probability that it was drawn from Bag I.

Solution:

Let E1 be the event of choosing bag I,

E2 be the event of choosing bag II, and A be the event of drawing a black ball.

Then,

P(E1) = P(E2) = ½

Also,

P (drawing a black ball from Bag I) = P (A/E1) = 6/10= 3/5

P (drawing a black ball from Bag II) = P (A/E2) = 3/7

By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(E1 / A) = \frac{P(E1)\, P(A/E1)}{P(E1)\, P(A/E1) + P(E2)\, P(A/E2)} = \frac{7}{12}$$

Thus

Let E_1, E_2, E_3 …, B_n be n mutually exclusive events such that where S is the sample space of the random experiment. If P(Ei) > 0 for i = 1, 2, …, n, then for any event A of the same experiment with P(A) > 0,

$$P(Ei / A) = \frac{P(Ei)\, P(A/Ei)}{\sum P(Ei)\, P(A/Ei)}$$

Let us sum up

In this unit you have learned about the definitions of Bayes theorem of probability and its application numerical problems.

Check your progress

1. If P(A) = 0, then A is called _____ event.
2. If probability for the occurrence of an event is 1, then the event is known as _____ event.
3. On throwing the single die, then the event of getting odd number or even number are _____ event.

Glossary

Experiment – Events - Sample space – conditional probability – Bayes theorem for probability - Independent events - Dependent events.

Model Questions

1. Three identical boxes contain red and white balls. The first box contains 3 red and 2 white balls, the second box has 4 red and 5 white balls, and the third box has 2 red and 4 white balls. A box is chosen very randomly and a ball is drawn from it. If the ball that is drawn out is red, what will be the probability that the second box is chosen?
2. Two urns contain respectively 2 red, 3 white, and 3 red, 5 white balls. One ball is drawn at random from the first urn and transferred into the second one. A ball is then drawn from the second urn and it turns out that the ball is red. What will be the probability that the transferred ball was white?

Answer to check your progress

1. Impossible
2. Sure
3. Mutually exclusive

Suggested readings

1. Sharpe N.R., De Veaux R.D., Velleman P.F., Wright D. and Bock D.E. Business Statistics, Boston, MA: Addison Wesley(2010).
2. Beri, ”Business Statistics” Tata McGraw Hill,2ndEd,2009
3. Keller. G, ”Statistics for Management”, Cengage Learning, 1st Ed,2009.

Block-3: Introduction

Block-3: Correlations and Regression has been divided into four Units (Unit-9 to Unit-12)

Unit-9: Correlation – Introduction, Correlation, Types of Correlation, Correlation Analysis

Unit-10 : Methods of Correlation – Introduction, Simple Correlation, Scatter Diagram, Correlation Coefficient, Karl Pearson Coefficient Correlation, Spearman's Rank Correlation.

Unit-11: Partial and Multiple Correlations - Introduction, Partial Correlation, Multiple Correlation Coefficient.

Unit-12: Regression Analysis – Introduction, the Concept of Regression, Regression Analysis, Regression Coefficient, Difference between Correlation and Regression.

In all the units of Block -3: **Correlations and Regression** the Check your progress, Glossary, Answers to Check your progress and Suggested Reading has been provided and the Learners are expected to attempt all the Check your progress as part of study.

Unit-9
Correlation

STRUCTURE

Overview

In this unit the meaning and definition of correlation, types of correlation and its applications and also the objectives of correlation has been clearly explained.

Objectives

After completion of this unit, you can able to:

- Explain the meaning and definition of correlation
- Describe about types of correlation and its applications
- Define the objectives of correlation
- Use Scatter diagrams to visualize the relationship between two variables
- Compute the simple and rank correlation coefficients between two variables

9.1. Introduction

The goal of statistical data analysis is to understand a complex, real-world phenomenon from partial and uncertain observations. It is important to make the distinction between the mathematical theory underlying statistical data analysis, and the decisions made after

conducting an analysis. Correlation is one which studies the relationship between two or more variables; Correlation quantifies the degree and direction to which two variables are related.

The statistical techniques discussed so far are for only one variable. In many research situations one has to consider two variables simultaneously to know whether these two variables are related linearly. If so, what type of relationship that exists between them? This leads to bivariate (two variables) data analysis namely correlation analysis. If two quantities vary in such a way that movements (upward or downward) in one are accompanied by the movements(upward or downward) in the other, these quantities are said to be co-related or correlated.

The correlation concept will help to answer the following types of questions.

- Whether study time in hours is related with marks scored in the examination?
- Is it worth spending on advertisement for the promotion of sales?
- Whether a woman's age and her systolic blood pressure are related?
- Is age of husband and age of wife related?
- Whether price of a commodity and demand related?
- Is there any relationship between rainfall and production of rice?

9.2. Correlation

Correlation is a statistical measure which helps in analyzing the interdependence of two or more variables. In this chapter the dependence between only two variables are considered.

A.M. Tuttle defines correlation as: "An analysis of the co-variation of two or more variables is usually called correlation"

Ya-kun-chou defines correlation as: "The attempts to determine the degree of relationship between variables".

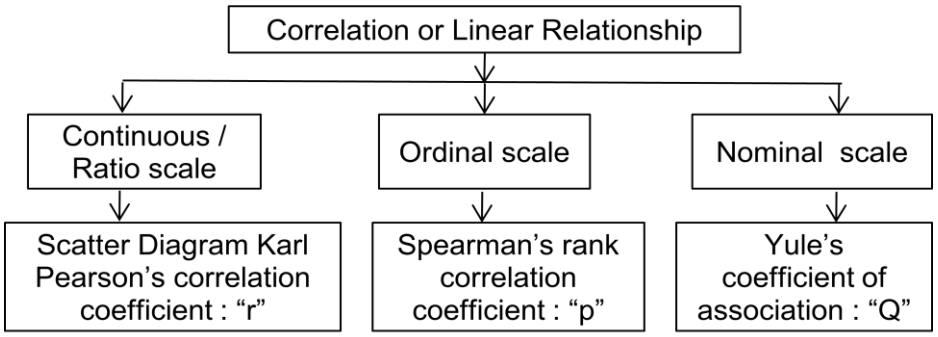
Correlation analysis is the process of studying the strength of the relationship between two related variables. High correlation means that variables have a strong linear relationship with each other while a low correlation means that the variables are hardly related. The type and intensity of correlation is measured through the correlation analysis. The measure of correlation is the correlation coefficient or correlation index. It is an absolute measure.

9.3. Types of Correlation

1. Simple (Linear) correlation (2 variables only): The correlation between the given two variables. It is denoted by rxy
2. Partial correlation (more than 2 variables): The correlation between any two variables while removing the effect of other variables. It is denoted by rxy.z …
3. Multiple correlation (more than 2 variables): The correlation between a group of variables and a variable which is not included in that group. It is denoted by Ry. (xz…)

9.4. Correlation Analysis

The purpose of correlation analysis is to find the existence of linear relationship between the variables. However, the method of calculating correlation coefficient depends on the types of measurement scale, namely, ratio scale or ordinal scale or nominal scale.



(1).According to Karl Pearson the coefficient of correlation in this case is +1. On the other hand, if the variables change in the opposite direction and in the same proportion, the correlation is perfect negative. Its coefficient of correlation is -1. In practice we rarely come across these types of correlations.

(2). Absence of correlation: If two series of two variables exhibit no relations between them or change in one variable does not lead to a change in the other variable, then we can firmly say that there is no correlation or absurd correlation between the two variables.

Let us sum up

In this unit, you have learned about the meaning and definition of correlation, types of correlation and its applications and also the objectives of correlation.

Check your progress

1. “The attempts to determine the degree of relationship between variables is correlation” is the definition given by

   (a) A.M. Tuttle

   (b) Ya-Kun-Chou

   (c) A.L. Bowley

   (d) Croxton and Cowden

2. If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner then it is called

   (a) perfect positive correlation

   (b) perfect negative correlation

   (c) positive correlation

   (d) negative correlation

3. The correlation coefficient lies in the interval

   (a) $-1 \leq r \leq 0$

   (b) $-1 < r < 1$

   (c) $0 \leq r \leq 1$

   (d) $-1 \leq r \leq 1$

Glossary

- Variables
- Relationship
- Simple (Linear) correlation
- Partial correlation
- Multiple correlation

Model Questions

1. What is correlation analysis?
2. Explain applications of correlations.
3. Explain the types of correlation.

Answer to check your progress

1. Ya-Kun-Chou

2. perfect negative correlation

3. $-1 \leq r \leq 1$

Suggested readings

1. J. K Sharma, “Business Statistics”, Pearson, 2nd Ed,2010.
2. Arora PN &others,” Complete Statistical Methods”, S. Chand, 3rd Ed,2010
3. Groebner D.F., Shannon P.W., Fry P.C. and Smith K.D. Business Statistics, Pearson Education(2008).

Unit-10
Methods of Correlation

STRUCTURE

Overview

In this unit the Simple Correlation, Scatter Diagram, Numerical and application problems of Karl Pearson correlation coefficient, Spearman's Rank Correlation and also the Rank correlation problems has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To know the concept of correlation.
- To apply scatter diagrams to visualize the relationship between two variables.
- To calculate the simple and rank correlation coefficients between two variables.

10.1. Introduction

Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. If two series vary in such a way, that fluctuations in one are accompanied by the fluctuations in the other,

these variables are said to be correlated. Like rise in price of a commodity, reduces its demand and vice-versa.

Some relationship exists between age of husband and wife, rainfall and production. Two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable. According to A. M. Tuttle, “Analysis of co-variation of two or more variables is usually called correlation”.

10.2. Simple Correlation

If two variables, say x and y vary or move together in the same or in the opposite directions they are said to be correlated or associated. Thus, correlation refers to the relationship between the variables. Generally, we find the relationship in certain types of variables.

For example, a relationship exists between income and expenditure, absenteeism and production, advertisement expenses and sales etc. Existence of the type of relationship may be different from one set of variables to another set of variables.

Let us discuss some of the relationships with the help of Scatter Diagrams.

Check your progress-1

True/False

a. According to A. M. Tuttle, “Analysis of co-variation of two or more variables is usually called correlation.

b. Karl Pearson is not a great British Biometrician and Statistician.

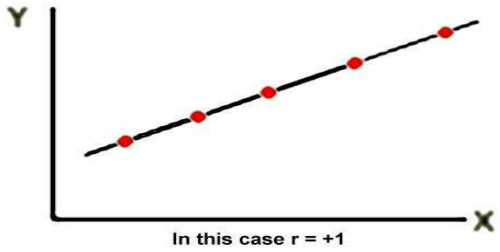
c. Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.

d. Generally, we don’t find the relationship in certain types of variables.

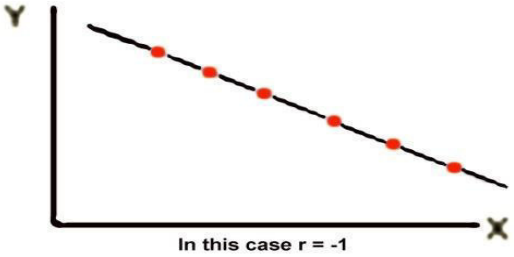
e. The graphical representation of correlation coefficient is called scatter diagram.

10.3. Scatter Diagram

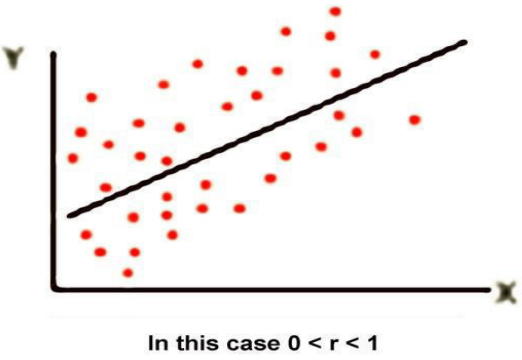
The graphical representation of correlation coefficient is called scatter diagram. In this method we plot the points of (x, y) s in a graph sheet. Then based on the plotted dots we can easily identify the types correlations are as follows.



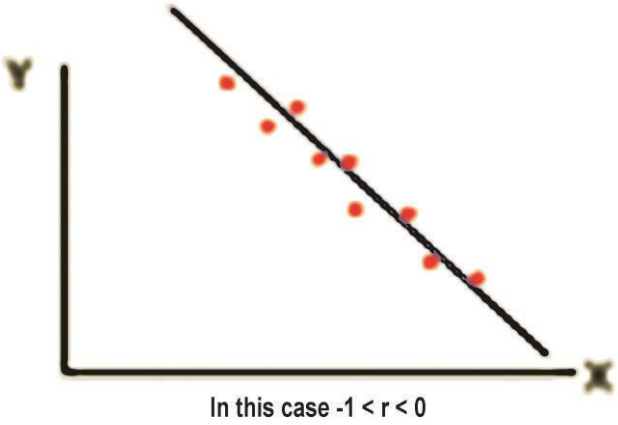
Perfect positively correlated. r=+1



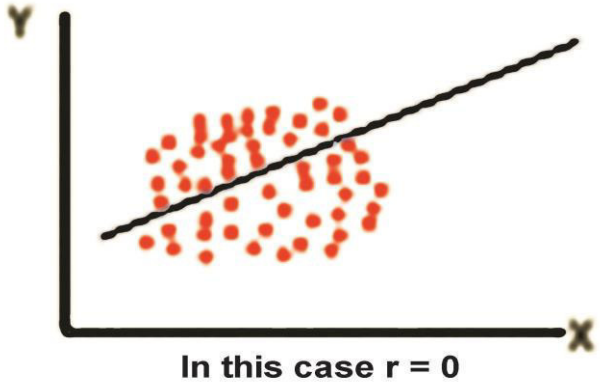
Perfect negatively correlated. =-1



Positively correlated>0



Negatively correlated. R<0



Un correlated. R=0

10.4. Correlation Coefficient

The coefficient of correlation helps in measuring the degree of relationship between two variables, X and Y. The methods which are used to measure the degree of relationship will be discussed below.

10.4.1. Karl Pearson Coefficient Correlation

A mathematical method for measuring the intensity or the magnitude of linear relationship between two variables was suggested by Karl Pearson (1867-1936), a great British Biometrician and Statistician and, it is by far the most widely used method in practice. Karl Pearson's measure, known as Pearson's correlation coefficient between two variables X and Y, usually denoted by r(X,Y) or rxy or simply r is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between X and Y, to the product of the standard deviations of X and Y.

Correlation- the relationship between two variables is called correlation. It was invented by Prof Karl Pearson. The correlation coefficient is denoted as "r"

Karl Pearson correlation coefficient (r) = $\frac{N\sum xy - \sum x \sum y}{\sqrt{[(N\sum X^2 - (\sum X)^2]}\sqrt{[(N\sum y^2 - (\sum y)^2]}}$

Here N = The no of pairs.

Problem: 1

Calculate Karl Pearson **correlation coefficient** for the following pairs of observations.

X	45	60	78	90	95
Y	38	45	90	56	75

Solution:

N=5 pairs.

X	Y	X^2	Y^2	XY
45	38	2025	1444	1710
60	45	3600	2025	2700
78	90	6084	8100	7020
90	56	8100	3136	5040
95	75	9025	5625	7125
368 ∑X	**304 ∑Y**	**28834 $\sum X^2$**	**20330 $\sum Y^2$**	**23595 ∑XY**

Karl Pearson correlation coefficient (r)

$$= \frac{N\sum xy - \sum x \sum y}{\sqrt{[(N\sum X^2 - (\sum X)^2\,]}\sqrt{[(N\sum y^2 - (\sum y)^2]}}$$

$$= \frac{(5X23595) - (368X304)}{\sqrt{[(5X28834) - (368)^2\,]}\sqrt{[(5X20330) - (304)^2]}}$$

$$= \frac{6103}{\sqrt{[(144170) - 135424\,]}\sqrt{[(101650) - 92416]}}$$

$$= \frac{6103}{\sqrt{[8746\,]}X\sqrt{[9234]}}$$

$$= \frac{6103}{93.52X96.09}$$

Karl Pearson correlation coefficient (r) $= \frac{6103}{8986.33}$ = **0.679**

(Positively Correlated)

10.4.2. Spearman's Rank Correlation

If the data are in ordinal scale then Spearman's rank correlation coefficient is used. It is denoted by the Greek letter ρ (rho). Spearman's correlation can be calculated for the subjectivity data also, like competition scores. The data can be ranked from low to high or high to low by assigning ranks. Spearman's rank correlation coefficient is given by the formula.

Spearman Rank correlation coefficient (ρ) = 1 - $\frac{6\sum D^2}{N^3 - N}$

Here N= the no of pairs.
D= R1 – R2
R1= Ranks from X
R2= Ranks from Y

PROBLEM: Calculate spearman **rank correlation coefficient** for the following pairs of observation.

X: 40 100 55 60 35 90 65

Y: 100 65 80 45 55 75 40

Solution: N=7 pairs

X	Y	R1	R2	D=R1-R2	D^2
40	100	6	1	5	25
100	65	1	4	-3	9
55	80	5	2	3	9
60	45	4	6	-2	4
35	55	7	5	2	4
90	75	2	3	-1	1
65	40	3	7	-4	16
$\sum D^2$					68

Spearman Rank correlation coefficient (ρ) = $1 - \frac{6\sum D^2}{N^3 - N}$

$= 1 - \frac{(6X68)}{7^3 - 7} = 1 - \frac{408}{343-7} = 1 - \frac{408}{336}$

$= 1 - 1.214$

Spearman Rank correlation coefficient (ρ) = -0.214 (Negatively correlated)

In this unit, fundamental concepts, meaning and techniques of correlation (or association) have been discussed. Scatter diagrams, which exhibit some typical pattern indicating different kinds of relationships, have been illustrated. A scatter plot of the variables may suggest that the two variables are related but the value of the Karl Pearson's correlation coefficient (r) quantifies the degree of this association. The closer the relation coefficient is to + 1.0, the stronger the linear relationship between the two variables. Spearman's rank correlation for data with rank is outlined. Finally, we discussed the procedure of assigning the ranks to the variables, if the data is in the values for computation of Rank correlation.

Properties of correlation coefficient:

- The relationship between the two variables is called correlation.

- It is lies between -1 and +1.
- It is independent of change of origin and scale.
- Correlation between X and Y is equal to correlation between Y and X.

Let us sum up

In this unit you have learned about Simple Correlation, Scatter Diagram, Numerical and application problems of Karl Pearson correlation coefficient, Spearman's Rank Correlation and also the Rank correlation problems.

Check your progress-2

1. The correlation coefficient lies in the interval

   (a) $-1 \leq r \leq 0$
   (b) $-1 < r < 1$
   (c) $0 \leq r \leq 1$
   (d) $-1 \leq r \leq 1$

2. Rank correlation was developed by

   (a) Pearson
   (b) Spearman
   (c) Yule
   (d) Fisher

3. The purpose of the study of _____ is to identify the factors of influence and try to control them for better performance.

   (a) mean
   (b) correlation
   (c) standard deviation
   (d) skewness

4. The height and weight of a group of persons will have _____ correlation.

   (a) positive
   (b) negative
   (c) zero
   (d) both positive and negative

Glossary

- Scatter diagram
- Karl Pearson correlation coefficient
- Spearman correlation coefficient

Model Questions

1. Define correlation with an example.
2. Define positive and negative correlation with examples.
3. Explain linear correlation with an example.
4. Detail about the characteristics of correlation coefficient.
5. Explain rank correlation with formula.
6. Discuss about scatter diagram.
7. Determine the Karl Pearson correlation coefficient between the following observations to find the relationship.

X	3	5	6	7	9	12
Y	20	14	12	10	9	7

8. Estimate the Karl Pearson correlation coefficient to find the relationship between the following observations.

Age of cars (Years)	2	4	6	7	8	10	12
Annual Maintenance (cost'00)	16	15	18	19	17	21	20

9. The following are the ranks obtained by 10 students in Statistics and Mathematics. Utilize Spearman Rank correlation to find out at what extent the knowledge of students in the two subjects related?

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

10. Compute Spearman rank correlation coefficient from the IQ level and exam marks of students.

I.Q.	110	100	140	120	80	90
Exam marks	70	60	80	55	10	20

Answer to check your progress-1

a-True

b-False

c-True

d-False

e-True

Answer to check your progress-2

1. $-1 \leq r \leq 1$
2. Spearman
3. Correlation
4. Positive

Suggested readings

1. R.S.N. Pillai, V. Bagavathi," Statistics", S.Chand Limited, 7thEd,2008
2. Elhance , D.N. Fundamentals of Statistics. Allahabad: KitabMahal, (2007).
3. G.V. Shenoy, Uma K. Srivastava, S.C. Sharma," Business Statistics", New Age International,2nd Ed, 2005

Unit-11
Partial and Multiple Correlations

STRUCTURE

Overview

In this unit the concept of multiple correlation and partial correlation and also the applications and formula for partial along with multiple correlation coefficient has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the concept of partial correlation.
- To Explain, the difference between partial and semi-partial correlation.
- To Describe and explain concept of multiple correlation.

11.1. Introduction

While learning about correlation, we understood that it indicates relationship between two variables. Indeed, there are correlation coefficients that involve more than two variables. It sounds unusual and you would wonder how to do it? Under what circumstance it can be done? Let me give you two examples. The first is about the correlation between cholesterol level and bank balance for adults.

Let us say that we find a positive correlation between these two factors. That is, as the bank balance increases, cholesterol level also increases. But this is not a correct relationship as Cholesterol level can also

increase as age increases. Also as age increases, the bank balance may also increase because a person can save from his salary over the years. Thus there is age factor which influences both cholesterol level and bank balance.

Suppose we want to know only the correlation between cholesterol and bank balance without the age influence, we could take persons from the same age group and thus control age, but if this is not possible we can statistically control the age factor and thus remove its influence on both cholesterol and bank balance. This if done is called partial correlation. That is, we can use partial and part correlation for doing the same.

Sometimes in psychology we have certain factors which are influenced by large number of variables. For instance academic achievement will be affected by intelligence, work habit, extra coaching, socio economic status, etc. To find out the correlation between academic achievements with various other factors ad mentioned above can be done by Multiple Correlation.

11.2. Partial Correlation

Two variables, A and B, are closely related. The correlation between them is partialled out, or controlled for the influence of one or more variables is called as partial correlation. So when it is assumed that some other variable is influencing the correlation between A and B, then the influence of this variable(s) is partialled out for both A and B. Hence it can be considered as a correlation between two sets of residuals. Here we discuss a simple case of correlation between A and B is partialled out for C. This can be represented as RAB.C which is read as correlation between A and B partialled out for C. the correlation between A and B can be partialled out for more variables as well.

11.3. Multiple Correlation Coefficient

The multiple correlation coefficient denoting a correlation of one variable with multiple other variables. The multiple correlation coefficient is denoted as RA.BCD…k which denotes that A is correlated with B, C, D, up to k variables.

For example, we want to compute multiple correlation between A with B and C then it is expressed as RA.BC. In this case we create a linear combination of the B and C which is correlated with A. We continue with the same example which we have discussed for partial and semi-partial correlations. This example has academic achievement, anxiety and intelligence as three variables. The correlation between academic

achievement with the linear combination of anxiety and intelligence is multiple correlation. This denotes the proportion of variance in academic achievement explained by intelligence and anxiety. We denote this as R (Academic Achievement. Intelligence, Anxiety), which is a multiple correlation. Often, it is used in the context of regression, where academic achievement is a criterion variable and intelligence and anxiety are called as predictors.

Let us sum up

In this unit you have learned about the interesting procedures of computing the correlations. Especially, when we are interested in controlling for one or more variable. The multiple correlations provide us with an opportunity to calculate correlations between a variable and a linear combination of other variable.

Check your progress

1. The statistical device which helps in analyzing the co-variation of two or more variables is

   (a) variance
   (b) probability
   (c) correlation coefficient
   (d) coefficient of skewness

2. "The attempts to determine the degree of relationship between variables is correlation" is the definition given by

   (a) A.M. Tuttle
   (b) Ya-Kun-Chou
   (c) A.L. Bowley
   (d) Croxton and Cowden

3. If all the plotted points lie on a straight line falling from upper left hand corner to lower right hand corner then it is called

   (a) perfect positive correlation
   (b) perfect negative correlation
   (c) positive correlation
   (d) negative correlation

Glossary

- Partial correlation
- Multiple correlation

Model Questions

1. Write the applications partial correlation.
2. Write the applications of multiple correlation.
3. Which correlation coefficient she should compute, if she wants to know the relationship between health (Y) and linear combination of stress (X1) and coping (X2).

Answer to Check your Progress

1. Correlation coefficient
2. Ya-Kun-Chou
3. Perfect negative correlation

Suggested readings

1. Vittal.P.R ‘’Mathematical Statistics’’, Margham Publications, Chennai (2002).
2. Gupta, S.P. and M.P. Gupta, Business Statistics, Sultan Chand & Sons: New Delhi(2000).
3. Richard I. Levin and David S. Rubin, Statistics for Management. Prentice Hall of India Pvt. Ltd., New Delhi(1996).

Unit-12
Regression Analysis

STRUCTURE

Overview

In this unit the concept and applications of regression and how can construct the two regression equations like X on Y and Y on X and also the difference between correlation and regression has been clearly explained.

Objectives

After completion of this unit, you will able to:

- Know the concept of regression, its types and their uses.
- Calculate the regression coefficient and interpret the same.
- Know the uses of regression coefficients.
- Distinguish between correlation analysis and regression analysis.

12.1. Introduction

The correlation coefficient is a useful statistical tool for describing the type (positive or negative or uncorrelated) and intensity of linear relationship (such as moderately or highly) between two variables. But it fails to give a mathematical functional relationship for prediction purposes. Regression analysis is a vital statistical method for obtaining functional relationship between a dependent variable and one or more

independent variables. More specifically, regression analysis helps one to understand how the typical value of the dependent variable (or 'response variable') changes when any one of the independent variables (regressor(s) or predictor(s)) is varied, while the other independent variables are held fixed. It helps to determine the impact of changes in the value(s) of the independent variable(s) upon changes in the value of the dependent variable. Regression analysis is widely used for prediction.

12.2. The Concept of Regression

The two regression lines coincide when there is a perfect positive or perfect negative correlation between the two variables. If there is zero correlation between two variables (i.e.)

If the variables are uncorrelated then the two regression lines are perpendicular. These regression lines are expressed algebraically as regression equations.

Thus, there are two regression equations, one regression equation of Y and X and another regression equation of X on Y.

Let us consider two variables X and Y. We shall have regression lines-one regression line of X on y which gives the most probable values of X for given values of Y and another regression line of Y on X which gives the most probable values of Y for given values of X.

Thus, in the case of two variables X and Y we have two regression line one to estimate X given Y and another to estimate Y given X.

12.3. Regression Analysis

The regression analysis is widely used in all the scientific disciplines. In economics, it plays a significant role in measuring or estimating the relationship among the economic variables.

For example, the two variables – price (X) and demand (Y) are closely related to each other, so we can find out the probable value of X from the given value of Y and similarly the probable value of Y can be found out from the given value of X. The Regression Line is the line that best fits the data, such that the overall distance from the line to the points (variable values) plotted on a graph is the smallest.

In other words, a line used to minimize the squared deviations of predictions is called as the regression line. There are as many numbers of regression lines as variables. Suppose we take two variables, say X and Y, then there will be two regression lines: ▪ Regression line of Y on

X: This gives the most probable values of Y from the given values of X. Regression line of X on Y: This gives the most probable values of X from the given values of Y.

12.4. Regression Coefficient

Regression equation of X on Y

X- $\bar{X}$ = bxy (Y- $\bar{Y}$)

Regression equation of Y on X

Y- $\bar{Y}$ = byx (X- $\bar{X}$)

Here $\bar{X}$ = mean of X = $\frac{\sum X}{n}$, Here n is no of pairs.

$\bar{Y}$ = mean of Y = $\frac{\sum Y}{n}$

Here bxy= regression coefficient of X on Y = $\frac{N\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$

Here byx= regression coefficient of Y on X = $\frac{N\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$

Problem: 1

Calculate the two regression equations for the following details.

Advt.Exp(Rs in Lakhs):	10	12	15	20	30
Sales(Rs.Crores):	5	6	8	7	10

Also estimate the advertisement expenditure for attaining the sales 15 crores.

Solution:

Let X be the advertisement expenditure.

Let Y be the sales.

X	Y	X^2	Y^2	XY
10	5	100	25	50
12	6	144	36	72
15	8	225	64	120
20	7	400	49	140
30	10	900	100	300
87	**36**	**1769**	**274**	**682**

N= 5

$\bar{X}$ = mean of X = $\frac{\sum X}{n} = \frac{87}{5}$ =17.4

$\bar{Y}$ = mean of Y = $\frac{\sum Y}{n} = \frac{36}{5}$ =7.2

Here bxy= regression coefficient of X on Y = $\frac{N\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$

$= \frac{(5X682) - (87X36)}{(5X274) - 1296} = \frac{278}{74}$ = 3.76

Here byx= regression coefficient of Y on X = $\frac{N\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$

$= \frac{(5X682) - (87X36)}{(5X1769) - 7569} = \frac{278}{1276}$ =0.22

Regression equation of X on Y

X- $\bar{X}$ = bxy (Y- $\bar{Y}$)

X- 17.4 = 3.76 (Y- 7.2)

X= 3.76Y-27.07+17.4

X=3.76Y-9.67

Regression equation of Y on X

Y- $\bar{Y}$ = byx (X- $\bar{X}$)

Y-7.2 = 0.22 (X- 17.4)

Y=0.22X – 3.83+7.2

Y=0.22X+3.37

Estimation of advertisement expenditure(x) when a sale(y) is 15

X=3.76Y-9.67

X= (3.76X15)-9.67

X=56.4-9.67

X= 46.73 Lakhs.

12.5. Difference between Correlation and Regression

Correlation	Regression
• In correlation analysis the degree and direction of relationship between the variables are studied.	• In regression analysis, the nature of relationship is studied.
• If value of one variable is known, the value of other variable cannot be estimated.	• If value of variable is known, the value of other variable can be estimated using the functional relationships.

• Correlation coefficient lies between -1 and 1. • Correlation coefficient is independent of change of origin and scale. • With the help of correlation coefficient and standard deviations of two random variable (X,Y) regression coefficient can be obtained.	• Only one regression coefficient can be greater than 1. • Regression coefficient is independent of change of origin but not of scale.

Thus, fundamental concepts and techniques of simple linear regression have been discussed i.e. in case of two variables only Once it is identified that correlation exists between the variables, an estimating equation known as regression equation could be developed by the least squares method for prediction. Relationship between correlation and regression coefficient and the conceptual differences between correlation and regression have been highlighted. The techniques of regression analysis are widely used in business decision making and data analysis.

Let us sum up

In this unit, you have learned about the concept and applications of regression and how can construct the two regression equations like X on Y and Y on X. And also the difference between correlation and regression.

Check your progress

1. Regression analysis helps in establishing a functional relationship between _____ variables.
2. _____ is widely used for prediction

   a) Regression analysis
   b) correlation analysis
   c) analysis of variance
   d) analysis of covariance

3. The linear equation Y = a + bx is called as regression equation of

   a) X on Y
   b) Y on X
   c) between X and Y
   d) ‘a’ on ‘b’

Glossary

Linear Relationship: The relationship between two variables described by a straight line.

Least Squares Criterion: The criterion for determining a regression line that minimizes the sum of squared errors.

Simple Regression Analysis: A regression model that uses one independent variable to explain the variation in the dependent variable.

Model Questions

1. Outline about regression.
2. Discuss about dependent and independent variables
3. Brief the utility of regression analysis.
4. Differentiate between correlation and regression.
5. The following sample data are price of a product and its demand. Fit the linear regression equation. Find the impact of demand (x) on price (y)

Price (Rs.) (y)	10	18	14	11	16	13
Demand (x)	125	58	90	100	72	85

6. For the following table find and fit the linear regression equation to find the impact of x on y

x	18	23	27	32	38	30	40
y	73	70	78	86	95	105	117

Answer to check your progress

1. Two or more variables

2. Regression analysis

3. Y on X

Suggested readings

1. Bernsen M.L. and Levine D.M, Basic Business Statistics, Prentice-Hall Englewood Cliffs New jersey(1996).
2. Webster. A Applied Statistics For Business and Economics, Homewood, IL: Irwin(1992).

Block-4: Introduction

Block-4: Testing of Hypothesis has been divided in to four Units (Unit-13 to Unit-16)

Unit-13: Introduction to Testing of Hypothesis deals with Introduction, Hypothesis and its types, Steps involved in testing of hypothesis, One tail test and two tail test and Type-I error and Type-II error.

Unit-14: Student 'T' Test explains about Introduction, Student "t" test for single mean and Student "t" test for double mean.

Unit-15: Chi-Square Test presents about Introduction, Applications of Chi-square test and Chi-square test for independence of attributes.

Unit-16: Analysis of Variance describes with Introduction, One way ANOVA and Two way ANOVA.

In all the units of Block -4: **Testing of Hypothesis**, the Check your progress, Glossary, Answers to Check your progress and Suggested Reading has been provided and the Learners are expected to attempt all the Check your progress as part of study.

Unit-13

Introduction to Testing of Hypothesis

STRUCTURE

Overview

In this unit the concept of hypothesis, different types of hypothesis, steps involved in testing of hypothesis and also the concept of Type-I error and Type-II error has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the concept of inferential statistics
- To know the types of hypothesis
- To identify the steps involved in testing of hypothesis
- To know the concept of type I error and type II error.

13.1. Introduction

We concentrated on collection, presentation and analysis of data along with calculation of various measures of central tendency and measures of dispersion. These kinds of describing the data are popularly known as descriptive statistics. Now, we need to understand another dimension of statistical data analysis, which is called inferential statistics.

Inferential Statistics may be described as follows from the statistical point of view:

One of the main objectives of any scientific investigation or any survey is to find out the unknown facts or characteristics of the population under consideration. It is practically not feasible to examine the entire population, since it will increase the time and cost involved. But one may examine a part of it, called sample. On the basis of this limited information, one can make decisions or draw inferences on the unknown facts or characteristics of the population.

Thus, inferential statistics refers to a collection of statistical methods in which random samples are used to draw valid inferences or to make decisions in terms of probabilistic statements about the population under study.

13.2. Hypothesis and its types

The Hypothesis is an assumption which is tested to check whether the inference drawn from the sample of data stand true for the entire population or not.

In many practical studies, as mentioned earlier, it is necessary to make decisions about a population or its unknown characteristics on the basis of sample observations. For example, in biomedical studies, we may be investigating a particular theory that the recently developed medicine is much better than the conventional medicine in curing a disease. For this purpose, we propose a statement on the population or the theory. Such statements are called hypotheses.

Thus, a hypothesis can be defined as a statement on the population or the values of the unknown parameters associated with the respective probability distribution. All the hypotheses should be tested for their validity using statistical concepts and a representative sample drawn from the study population. ‘Hypotheses’ is the plural form of ‘hypothesis’.

A statistical test is a procedure governed by certain determined/derived rules, which lead to take a decision about the null hypothesis for its rejection or otherwise on the basis of sample values. This process is called statistical hypotheses testing. The statistical hypotheses testing plays an important role, among others, in various fields including industry, biological sciences, behavioral sciences and Economics. In each hypotheses testing problem, we will often find as there are two hypotheses to choose between viz., null hypothesis and alternative hypothesis.

Null Hypothesis:

A hypothesis which is to be actually tested for possible rejection based on a random sample is termed as null hypothesis, which will be denoted by H0.

1. Generally, it is a hypothesis of no difference in the case of comparison.
2. Assigning a value to the unknown parameter in the case of single sample problems
3. Suggesting a suitable model to the given environment in the case of model construction.
4. (iv)The given two attributes are independent in the case of Chi-square test for independence of attributes.

Alternative Hypothesis:

A statement about the population, which contradicts the null hypothesis, depending upon the situation, is called alternative hypothesis, which will be denoted by H1.

13.3. Steps involved in testing of hypothesis

Setting the hypothesis

1. **Level of significance (α):** The chance of error in statistical test. Once the hypothesis about the population is constructed the researcher has to decide the level of significance, i.e. a confidence level with which the null hypothesis is accepted or rejected. The significance level is denoted by 'α' and is usually defined before the samples are drawn such that results obtained do not influence the choice. In practice, we either take 5% or 1% level of significance.

   If the 5% level of significance is taken, it means that there are five chances out of 100 that we will reject the null hypothesis when it should have been accepted, i.e. we are about 95% confident that we have made the right decision. Similarly, if the 1% level of significance is taken, it means that there is only one chance out of 100 that we reject the hypothesis when it should have been accepted, and we are about 99% confident that the decision made is correct.

2. **Test statistic:** Here we are going to apply some formula which is related to the hypothesis and data like t, F, χ^2 etc.
3. **Calculated value (CV):** Using the test statistic we can derive the quantitative value, is called CV.

4. **Table value (TV):** This value will be calculated from the respective probability distribution using level of significance and degrees of freedom.
5. **Conclusion:** Here we are comparing CV and TV. If CV is less than TV we accept null hypothesis H_0. If CV is greater than the TV we accept H_1.

13.4. One tail test and two tail test

Definition of One-tailed Test

- One-tailed test alludes to the significance test in which the region of rejection appears on one end of the sampling distribution. It represents that the estimated test parameter is greater or less than the critical value. When the sample tested falls in the region of rejection, i.e. either left or right side, as the case may be, it leads to the acceptance of alternative hypothesis rather than the null hypothesis. It is primarily applied in chi-square distribution; that ascertains the goodness of fit.
- In this statistical hypothesis test, all the critical region, related to α, is placed in any one of the two tails.

  One-tailed test can be:
- Left-tailed test: When the population parameter is believed to be lower than the assumed one, the hypothesis test carried out is the left-tailed test.
- Right-tailed test: When the population parameter is supposed to be greater than the assumed one, the statistical test conducted is a right-tailed test.

Definition of Two-tailed Test

- The two-tailed test is described as a hypothesis test, in which the region of rejection or say the critical area is on both the ends of the normal distribution. It determines whether the sample tested falls within or outside a certain range of values. Therefore, an alternative hypothesis is accepted in place of the null hypothesis, if the calculated value falls in either of the two tails of the probability distribution.
- In this test, α is bifurcated into two equal parts, placing half on each side, i.e. it considers the possibility of both positive and negative effects. It is performed to see, whether the estimated parameter is either above or below the assumed parameter, so

the extreme values, work as evidence against the null hypothesis.

Check Your Progress-1

True/False

a. One of the main objectives of any scientific investigation or any survey is to find out the unknown facts or characteristics of the population under consideration.

b. 'Hypotheses' is the plural form of 'Hypothesis'.

c. The significance level is denoted by 'α' and is usually defined before the samples are drawn such that results obtained do not influence the choice.

d. One-tailed test alludes to the significance test in which the region of rejection appears on one end of the sampling distribution.

e. When the null hypothesis is true and you reject it, you make a Type-II error.

13.5. Type- I error and Type- II error

Type I error

When the null hypothesis is true and you reject it, you make a Type-I error. The probability of making a Type- I error is α, which is the level of significance you set for your hypothesis test.

An "α" of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis. To lower this risk, you must use a lower value for α. However, using a lower value for alpha means that you will be less likely to detect a true difference if one really exists.

Type-II error

When the null hypothesis is false and you fail to reject it, you make a Type-II error. The probability of making a Type-II error is β, which depends on the power of the test. You can decrease your risk of committing a Type-II error by ensuring your test has enough power. You can do this by ensuring your sample size is large enough to detect a practical difference when one truly exists.

Steps involved in testing of hypothesis

Thus, the steps involved in testing of hypothesis are

1. **Setting the hypothesis:** Ho and H1

2. **Level of significance (α) :** 5% or 1%
3. **Test statistic:** here we are going to apply some which is related to the hypothesis and data. T, F and χ^2
4. **Calculated value (CV):** using the test statistic we can derive the quantitative value, is called CV.
5. **Table value (TV):** This value will be calculated from the respective probability distribution using level of significance and degrees of freedom.
6. **Conclusion:** Here we are comparing CV and TV. If CV is less than TV we accept null hypothesis H_0. If CV is greater than the TV we accept H_1.

Let us Sum up

In this unit you have learned about the concept of hypothesis, different types of hypothesis, steps involved in testing of hypothesis and also the concept of Type-I error and Type-II error.

Check your progress-2

1. Probability of type I error is called_____.
2. In conclusion of testing of hypothesis, calculated value is less than the table value we_____ the null hypothesis.
3. We reject the hypothesis when it is true. Such kind of error is called Type_____ error.

Glossary

Parameter – statistic – Hypothesis – null hypothesis – alternative hypothesis – level of significance – test statistic – calculated value – table value –degrees of freedom – conclusion – Type-I error – Type-II error.

Model Questions

1. Discuss about parametric and non-parametric tests
2. Explain the steps in testing of hypothesis.
3. Explain about one tailed and two tailed tests.
4. Brief about type I and type II error.
5. What do you understand by Hypothesis Testing?
6. Explain Null Hypothesis and Alternate Hypothesis.

Answers to check your progress-1

a-True

b-True

c-True

d-True

e-False

Answers to check your progress-2

1. Level of significance
2. Accept
3. One

Suggested readings

1. Richard I. Levin and David S. Rubin, Statistics for Management. Prentice Hall of India Pvt. Ltd., New Delhi (1996).
2. Bernsen M.L.And Levine D.M, Basic Business Statistics, Prentice-Hall Englewood Cliffs New jersey (1996).

Unit-14
Student “t” test

STRUCTURE

Overview

In this unit the steps involved in one sample “t” test ad two sample “t” test for small sample observations and also application problems solved related to the above two “t” tests has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the concept of t test for single mean
- To compare the two sample means using t test for double mean

14.1. Introduction

The “t” test is one type of inferential statistics. It is used to determine whether there is a significant difference between the means of two groups. With all inferential statistics, we assume the dependent variable fits a normal distribution.

A One-Sample T-Test compares a sample mean and a known population mean to determine whether the difference between the two means is statistically significant or occurred by chance alone.

This example will be comparing the respondents' number of children with the known 2013 United States fertility rate of 2.06 children per woman. The One-Sample T-Test is examining whether the difference between

the sample mean numbers of children per respondent is significantly different from the known population fertility rate.

An Independent-Samples T-Test compares two sample means from different populations regarding the same variable to determine whether the difference between the two means is statistically significant or occurred by chance alone. This example will be comparing the mean number of hours spent emailing per week by married respondents and single respondents.

The 'Email Hours Per Week, emailhr' variable is the test variable, and the 'Not Married, single' variable is the nominal grouping variable. The Independent-Samples T-Test is examining whether the difference between the mean number of hours married respondents spent emailing and the mean number of hours single respondents spent emailing is significantly different or occurred by chance.

14.2. Student t test for single mean

Null hypothesis: sample mean is equal to population mean.

Alternative hypothesis: sample mean is not equal to population mean.

Level of significance: 0.05

Test statistic:

$$t = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

here $\bar{X}$ = sample mean $= \frac{\sum X}{n}$

S^2= sample variance $= \frac{\sum (X - \bar{X})2}{n-1}$

Calculated value ***

Table value ****

Conclusion

If CV is less than TV, accept null hypothesis.

If CV is greater than TV, accept alternative hypothesis.

Problem:

A sample of 10 students from a school was selected. Their scores in a particular subject are 72, 82, 96, 85, 84, 75, 76, 93, 94 and 93. Can we support the claim that the class average scores is 90? (Table value = 2.262)

Solution:

Null hypothesis (H_0): μ = 90

Alternative hypothesis (H_1): μ ≠ 90

Level of significance (α) = 0.05

Test statistic:

$$t = \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

Here $\bar{X}$ = sample mean $= \frac{\sum X}{n}$

S^2= sample variance $= \frac{\sum (X - \bar{X})2}{n-1}$

X	**$(X - \bar{X})^2$** **$(X - 85)^2$**
72	169
82	9
96	121
85	0
84	1
75	100
76	81
93	64
94	81
93	64
Total =850	690

$\bar{X}$ = sample mean = $\frac{\sum X}{n} = \frac{850}{10} = 85$

S^2= sample variance = $\frac{\sum (X - \bar{X})2}{n-1} = \frac{690}{9} = 76.67$

$t = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} = \frac{85 - 90}{\sqrt{76.67/10}} = \frac{5}{\sqrt{7.667}} = \frac{-5}{2.768} = -1.806$

Calculated value = 1.806 (t value is always positive)

Table value = 2.262

Conclusion:

Here CV 1.806 is less than the TV 2.262, so accept null hypothesis.

μ = 90

Yes we can conclude the average score of the class is 90.

14.3. Student t test for double mean

Null hypothesis: There is no significance difference between two sample mean.

Alternative hypothesis: There is significance difference between two sample mean.

Level of significance: 0.05

Test statistic

$t = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2}/(n1+n2)}$

Here $\bar{X} = \frac{\sum X}{n_1}$ = $\bar{Y} = \frac{\sum Y}{n_2}$

$S^2 = \frac{\sum (X - \bar{X})2 + \sum (Y - \bar{Y})2}{n_1+n_2-2}$

Calculated value***

Table value ****

Conclusion:

If CV is less than TV, accept H_0.

If CV is greater than TV, accept H_1.

Problem:

The following table gives the scores of two batches of students in an examination.

Batch I	6	7	9	2	13	3	4	8	7	11
Batch II	5	6	5	7	1	7	2	7		

Test at 5 % level of significance the average performance of the students in Batch-I and Batch-II are equal.

Solution:

Null hypothesis (H_0):

There is no significant difference between two sample mean. ($\bar{X} = \bar{Y}$)

Alternative hypothesis (H_1):

There is significant difference between two sample mean. ($\bar{X} \neq \bar{Y}$)

Level of significance (α): 0.05

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2}(1/n1+1/n2)}$$

Here $\bar{X} = \frac{\sum X}{n_1}$ $\bar{Y} = \frac{\sum X}{n_2}$ $S^2 = \frac{\sum (X - \bar{X})2 + \sum (Y - \bar{Y})2}{n1+n2-2}$

/

X	Y	$(X - \bar{X})^2$ $(X -7)^2$	$(Y - \bar{Y})^2$ $(Y -5)^2$
6	5	1	0
7	6	0	1
9	5	4	0
2	7	25	4
13	1	36	16
3	7	16	4
4	2	9	9
8	7	1	4
7	-	0	-
11	-	16	-
70	**40**	**108**	**38**

$\bar{X} = \frac{70}{10} = 7 = \bar{Y} = \frac{40}{8} = 5, S^2 = \frac{108+38}{17} = \frac{146}{16}$ = 9.13

$t = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2}(1/n1+1/n2)} = \frac{7-5}{\sqrt{9.13}(\frac{1}{10}+\frac{1}{8})} = \frac{2}{\sqrt{9.13}(0.1+0.125)}$

$= \frac{2}{\sqrt{9.13}(0.1+0.125)} = \frac{2}{\sqrt{2.054}} = \frac{2}{1.433}$ = 1.39

Calculated value= 1.39

Table value= 2.120

Conclusion:

CV is less than TV, so accept H_0.

There is no significant difference between two sample mean. ($\bar{X} = \bar{Y}$).

To conclude:

- Test statistic of t test for single mean

  t= $\frac{\bar{X}-\mu}{\sqrt{S^2}/n}$ Here $\bar{X}$ = sample mean = $\frac{\sum X}{n}$

  S^2= sample variance = $\frac{\sum (X-\bar{X})2}{n-1}$

- Test statistic of t test for difference between two sample mean

  t= $\frac{\bar{X}-\bar{Y}}{\sqrt{S^2}/(n1+n2)}$

  Here $\bar{X}$ = $\frac{\sum X}{n_1}$ $\bar{Y}=\frac{\sum Y}{n_2}$ $S^2=\frac{\sum (X-\bar{X})2+\sum (Y-\bar{Y})2}{n1+n2-2}$

Let us sum up

In this unit, you have learned about the steps involved in one sample "t" test ad two sample "t" test for small sample observations and also application problems solved related to the above two "t" tests.

Check your progress

1. In general, large sample theory is applicable when

   (a) n ≥ 100

   (b) n ≥50

   (c) n ≥40

   (d) n ≥ 30

2. In general, small sample theory is applicable when

   (a) n < 100

   (b) n <50

   (c) n <40

   (d) n < 30

3. Critical value at 5% level of significance for two-tailed large sample test is

   (a) 1.645

   (b) 2.33

   (c) 2.58

   (d) 1.96

Glossary

Student t test for single mean – student t test for double mean - null hypothesis – alternative hypothesis – level of significance – test statistic – calculated value – table value –degrees of freedom – conclusion.

Model Questions

1. Explain One-sample t-test and Two-sample t-test.
2. A random sample of 10 packets containing cashew nuts weigh (in grams) 70,120,110,101, 88,83,95,98,107,100 each. Test whether the population mean weight of 100 grams?
3. The average run of cricket player from the past records is 80. The recent scores of the player in 6 test matches are 84, 82, 83, 79, 83 and 85. Test whether the average run is more than 80?
4. The heights (in feet) of 6 rain trees in a town A are 30, 28, 29, 32, 31, 36 and that of 8 rain trees in another town B are 35, 36, 37, 30, 32, 29, 35, 30. Is there any significant difference in mean heights of rain trees?
5. The marks secured by 9 students in Statistics and that of 12 students in Business Mathematics are given below:

Marks in Statistics	65	74	64	58	60	67	71	69	75			
Marks in Business Mathematics	52	45	59	47	53	64	58	62	54	61	57	48

Test whether the mean marks obtained by the students in Statistics and Business mathematics differ significantly at 1% level of significance.

Answers to check your progress

1. $n \geq 30$
2. $n < 30$
3. 1.96

Suggested readings

1. Vittal.P.R ‘’Mathematical Statistics’’, Margham Publications, Chennai(2002).
2. Gupta, S.P. and M.P. Gupta, Business Statistics, Sultan Chand & Sons: New Delhi(2000).

Unit-15
Chi-Square test

STRUCTURE

Overview

In this unit the steps involved in chi-square test for independence of attributes and also the applications of chi-square test has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To know the applications of chi-square test
- To understand the steps involved in chi-square test for independence of attributes
- To practice the problems in chi-square test.

15.1. Introduction

Karl Pearson (1857-1936) was an English Mathematician and Biostatistician. He founded the world's first university statistics department at University College, London in 1911. He was the first to examine whether the observed data support a given specification, in a paper published in 1900. He called it 'Chi-square goodness of fit' test which motivated research in statistical inference and led to the development of statistics as separate discipline. Karl Pearson chi-square test the dawn of Statistical Inference – C.R. Rao. Karl Pearson's famous

chi square paper appeared in the spring of 1900, an auspicious beginning to a wonderful century for the field of statistics - B. Efron.

15.2. Applications of Chi-square test

1. Testing the divergence of observed results from expected results when our expectations are based on the hypothesis of equal probability.
2. Chi-square test when expectations are based on normal distribution.
3. Chi-square test when our expectations are based on predetermined results.
4. Correction for discontinuity or Yates' correction in calculating $\chi 2$.
5. Chi-square test of independence in contingency tables.

15.3. Chi-square test for independence of attributes

To test the association or relationship between the variables or attributes or characteristics.

Null hypothesis: There is no association between the two variables.

Alternative hypothesis: There is association between two variables.

Level of significance (α) = 0.05

Test statistic:

$$\chi 2 = \sum \frac{(O-E)^2}{E}$$

Here O= observed frequency; E = Expected frequency

$$= \frac{Row\ total\ X\ Column\ total}{Grand\ total}$$

Calculated value ***

Table value ****

Conclusion:

Calculated value is less than Table value we accept null hypothesis. Otherwise we reject the null hypothesis and accept alternative hypothesis.

Problem:1

Test whether income and type of schooling is associated or not using chi square test. (Table value at 5% significant level is 3.841).

(2X2) Cross Tabulation (Contingency Table)

Income\Schooling	Private	Govt.	Total
Low	370	430	800
High	130	70	200
Total	500	500	1000

Solution:

Null hypothesis: Income and type of schooling not associated. (Independent)

Alternative hypothesis: Income and type of schooling associated. (Dependent)

Level of significance: 0.05

$\chi^2 = \sum \frac{(O-E)^2}{E}$

here O= observed frequency; E = Expected frequency

$= \frac{Row\ total\ X\ Column\ total}{Grand\ total}$

Income\Schooling	Private	Govt.	Total
Low	370	430	800
High	130	70	200
Total	500	500	1000

O1=370 O2=430 O3=130 O4=70

$E1=\frac{800\ X500}{1000} = 400$

$E2= \frac{800\ X500}{1000} = 400$

$E3 = \frac{200\ X500}{1000} = 100$

$E4 =\frac{200\ X500}{1000} = 100$

O	E	$\frac{(O-E)^2}{E}$
370	400	2.25
430	400	2.25
130	100	9.00
70	100	9.00
-	-	22.50

$\chi^2 = \sum \frac{(O-E)^2}{E} = 22.50$

Calculated value= 22.50

Table value = 3.841

Conclusion: CV 22.50 is greater than TV 3.841. So we accept alternative hypothesis.

Income and type of schooling associated. (Dependent).

Problem: 2

From the following cross tabulation using chi-square test can you conclude whether reference of colour and gender are associated?

Gender \ preference of colour	Pink	Blue	Total
Male	350	250	600
Female	100	300	400
Total	450	550	1000

Null hypothesis (H_0) : There is no association between gender and preference of colour.

Alternative hypothesis (H_1): There is association between gender and preference of colour.

Level of significance (α) : 0.05

$\chi^2 = \sum \frac{(O-E)^2}{E}$

Here O= observed frequency;

E = Expected frequency = $\frac{Row\ total\ X\ Column\ total}{Grand\ total}$

O	E	$\frac{(O-E)^2}{E}$
350	270	23.70
250	330	19.39
100	180	35.55
300	220	29.09
-	-	107.73

$\chi^2 = 107.73$

Calculated Value (CV) = 107.73

Table Value (TV) = 3.841

Conclusion:

Here CV is greater than TV, 107.73> 3.841

So we accept H_1.

There is association between gender and preference of colour.

Thus to conclude, Karl Pearson chi-square test the dawn of Statistical Inference.

Test statistic for chi-square test:

$\chi^2 = \sum \frac{(O-E)^2}{E}$

Here O= observed frequency;

E = Expected frequency = $\frac{Row\ total\ X\ Column\ total}{Grand\ total}$

Let us sum up

In this unit, you have learned about the steps involved in chi-square test for independence of attributes and also the applications of chi-square test

Check your progress

1. If the order of the contingency table is (5 × 4). Then the degree of freedom of the corresponding chi-square test statistic is

   a) 18

   b) 17

   c) 12

   d) 25

2. If n is the degree of freedom of chi-square distribution then its variance is

   a) n

   b) n–1

   c) 2n

   d) n+1

3. The degrees of freedom of chi-square test for R rows and C columns is_____.

Glossary

Chi-square test for independence of attributes - null hypothesis – alternative hypothesis – level of significance – test statistic – calculated value – table value –degrees of freedom – conclusion.

Model Questions

1. Explain the chi-square test for independence of attributes.
2. Write the applications of chi-square test.
3. The following table gives eye colour and hair colour of the customers. Use Chi square test and give statistical inference. You are given Chi-square table value for Degrees of Freedom 4 at 5% Level of Significance as 9.488.

		Hair colour		
		Fair	**Brown**	**Black**
Eye colour	**Blue**	25	5	20
	Grey	20	10	20
	Brown	25	15	20

4. Use Chi square and statistically interpret. Given below the economic conditions of students and their IQ level. You are given Chi-square table value for Degrees of Freedom 1 at 5% Level of Significance as 3.841.

Economic conditions	**IQ level**	
	High	**Low**
Rich	460	140
Poor	240	160

5. Utilize Chi-square test find out whether the opinion towards autonomous status depends upon level of graduation [UG and PG]. You are given Chi-square table value for Degrees of Freedom 1 at 5% Level of Significance as 3.841.

Particulars	**Under Graduates**	**Post Graduates**
Favoring autonomous	190	200
Unfavoring autonomous	110	100

Answer to check your progress

1. 12
2. 2n
3. (R-1) (C-1)

Suggested readings

1. Elhance , D.N. Fundamentals of Statistics. Allahabad: KitabMahal, (2007).
2. G. V. Shenoy, Uma K. Srivastava, S. C. Sharma," Business Statistics", New Age International,2nd Ed, 2005

Unit-16
Analysis of Variance (ANOVA)

STRUCTURE

Overview

In this unit the steps involved in one way ANOVA and two way ANOVA and also the applications of ANOVA and the difference between “t” test for double mean and ANOVA has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To know the applications of ANOVA(Analysis Of Variance)
- To understand the steps involved in one way ANOVA and two way ANOVA
- To practice the problems in ANOVA.

16.1. Introduction

In the previous chapters, we have discussed various concepts used in testing of hypotheses and problems relating to means of the populations. Although many practical problems involve inferences about population means or proportions, the inference about population variances is important and needs to be studied. ANOVA is used when we want to test the equality of means of more than two populations. For example, through ANOVA, one may compare the average yield of several varieties of a crop or average mileages of different brands of cars. In this

chapter we will study (i) one-way ANOVA and (ii) two-way ANOVA, using F-distribution.

16.2. One way ANOVA

In one-way ANOVA the effect of one factor on the mean is tested. It is based on independent random samples drawn from k – different levels of a factor, also called treatments. This unit will provide a brief introduction to the one-way ANOVA, including the assumptions of the test and when you should use this test.

Assumptions:

- Normality – The each sample is taken from a normally distributed population
- Sample independence – the each sample has been drawn independently of the other samples.
- Variance Equality – the variance of data in the different groups should be the same

Problem:

The company wants to compare the sales of three cities. Test the difference in sales through one way ANOVA.

Mumbai	Delhi	Chennai
8	12	11
12	7	8
9	13	14

Solution:

Null hypothesis: There is no significant difference in the sales of three different cities

Alternate hypothesis: There is significant difference in the sales of three different cities

Mumbai (X_1)	**Delhi (X_2)**	**Chennai (X_3)**	**X_1^2**	**X_2^2**	**X_3^2**
8	12	11	64	144	121
12	7	8	144	49	64
9	13	14	81	169	196
$\sum X_1 = 29$	**$\sum X_2 = 32$**	**$\sum X_3 = 33$**	**$\sum X_1^2 = 289$**	**$\sum X_2^2 = 362$**	**$\sum X_3^2 = 381$**

Total sum of all items (T) = $\sum X_1 + \sum X_2 + \ldots\ldots + \sum X_n$

= 29 + 32 + 33

= 94

Correction Factor (CF) = $\frac{T^2}{N}$ = $\frac{94^2}{9}$ = $\frac{8836}{9}$ = 982

Total sum of squares (TSS) = $(\sum X12 + \sum X22 + \ldots\ldots + \sum Xn2 - CF)$

= (289 + 362 + 381)- 982

= 1032 – 982 = 50

Sum of squares between items (SSB)

= $\left(\frac{(\sum X_1)2}{n} + \frac{(\sum X_2)2}{n} + \ldots\ldots + \frac{(\sum X_n)2}{n}\right) - CF$

= $\left(\frac{(29)^2}{3} + \frac{(32)^2}{3} + \frac{(33)^2}{3}\right) - 982$

= (280 + 341 + 363) – 982

= 984 – 982 = 2

Sum of Squares within item (SSW) = TSS – SSB = 50 – 2 = 48

ANOVA TABLE

Source of Variation	Sum of Squares	Df	Variance	F
Between	SSB = 2	= n – 1 = 3 – 1 = 2	= 2 / 2 = 1	= 1 / 8 = 0.125
Within	SSW = 48	= N – n = 9 – 3 = 6	= 48 / 6 = 8	

Calculated F = 0.125

Table F (df 2, 6 at 5% significant level) = 5.14

Calculated F < Table F

Accept Null hypothesis

There is no significant difference in the sales of three different cities.

16.3. Two-way ANOVA

In two-way ANOVA a study variable is compared over three or more groups, controlling for another variable. The grouping is taken as one factor and the control is taken as another factor. The grouping factor is usually known as Treatment. The control factor is usually called Block.

The accuracy of the test in two-way ANOVA is considerably higher than that of the one way ANOVA, as the additional factor, block is used to reduce the error variance.

Basis of comparison	ANOVA	
	One-way	Two-way
Independent variable	One	Two
Compares	Three or more levels of one factor	Three or more levels of two factors, simultaneously
Number of observations	Need not be same in each treatment group	Need to be equal in each treatment group

Since the test statistic is much larger than the critical value, we reject the null hypothesis of equal population means and conclude that there is a (statistically) significant difference among the population means.

Let us sum up

In this unit, you have learned about the steps involved in one way ANOVA and two way ANOVA and also the applications of ANOVA and the difference between “t” test for double mean and ANOVA.

Check your progress

1. ANOVA technique originated in the field of

   (a) Industry
   (b) Agriculture
   (c) Medicine
   (d) Genetics

2. F-test is also called as

   (a) Mean ratio test
   (b) variance ratio test
   (c) variance test
   (d) standard deviation ratio test

3. The Analysis of Variance procedure is appropriate for testing the equivalence of three or more population

   (a) Variances
   (b) proportions
   (c) means
   (d) observations

Glossary

Analysis of Variance - null hypothesis – alternative hypothesis – level of significance – test statistic – calculated value – table value –degrees of freedom – conclusion.

Model Questions

1. Explain one way ANOVA with example.
2. Explain two way ANOVA with example.
3. Distinguish between one way and two way ANOVA.
4. The following table gives the yield of three sample varieties of rice. Using one way ANOVA find whether mean yield of rice differ significantly among the varieties.

Variety I	Variety II	Variety III
6	5	5
7	5	4
3	3	3
8	7	4

5. Analyze using one way ANOVA to find the difference in sales performance of 3 salesman.

A	B	C
6	4	8
7	6	6
8	5	10

Answer to check your progress

1. Agriculture
2. Variance ratio test
3. Mean

Suggested Readings

1. Groebner D.F., Shannon P.W., Fry P.C. and Smith K.D. Business Statistics, Pearson Education(2008).
2. R.S.N. Pillai, V. Bagavathi,” Statistics”, S.Chand Limited, 7thEd,2008

Block-5: Introduction

Block-5:Index Numbers and Time Series Analysis has been divided into four Units(Unit-17 to Unit-20).

Unit-17 : Introduction to Index Numbers explains about Introduction, Definitions of Index numbers and the various Applications of Index Numbers.

Unit-18: Construction of Price Index Numbers deals with Introduction, Laspeyre's price Index number, Paasche's price Index number and Fisher's price Index number.

Unit-19: Introduction to Time Series describes about the Introduction, Applications of Time Series, Components of Time Series and Mathematical model of Time series.

Unit-20 : Methods for Finding the Trend Values presents about the Introduction, the Graphical method, the Semi-average method, the Moving average method and the Method of least squares.

In all the units of Block -5: **Index Numbers and Time Series Analysis**, the Check your progress, Glossary, Answers to Check your progress and Suggested Reading has been provided and the Learners are expected to attempt all the Check your progress as part of study.

Unit-17
Introduction to Index Numbers

STRUCTURE

Overview

In this unit the Introduction of Index Numbers, definition and the various applications of Index Numbers has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the definitions of Index numbers
- To know the applications of Index numbers
- To describe the limitations of index numbers to avoid errors in interpretation.

17.1. Introduction

Index number is a technique of measuring changes in a variable or a group of variables with respect to time, location or other characteristics. It is one of the most widely used statistical methods. Index number is a specialized average designed to measure the change in a group of related variables over a period of time. For example, the price of cotton in 2010 is studied with reference to its price in 2000. It is used to feel the pulse of the economy and it reveals the inflationary or deflationary tendencies. In reality, it is viewed as barometers of economic activity because if one wants to have an idea as to what is happening in an economy, he should check the important indicators like the index

number of agricultural production, index number of industrial production, and the index number business activity etc.,

17.2. Definitions of Index numbers

An Index Number is defined as a relative measure to compare and describe the average change in price, quantity value of an item or a group of related items with respect to time, geographic location or other characteristics accordingly.

In the words of Maslow “An index number is a numerical value characterizing the change in complex economic phenomenon over a period of time or space”

Spiegal defines, “An index number is a statistical measure designed to show changes in a variable on a group of related variables with respect to time, geographical location or other characteristics”.

According to Croxton and Cowden “Index numbers are devices for measuring differences in the magnitude of a group of related variables”.

Bowley describes “Index Numbers as a series which reflects in its trend and fluctuations the movements of some quantity”.

17.3. Applications of Index numbers

Economic Parameters

The Index Numbers are one of the most useful devices to know the pulse of the economy. It is used as an indicator of inflanationary or deflanationary tendencies.

Measures Trends

Index numbers are widely used for measuring relative changes over successive periods of time. This enables us to determine the general tendency. For example, changes in levels of prices, population, production etc. over a period of time are analysed.

Useful for comparison

The index numbers are given in percentages. So it is useful for comparison and easy to understand the changes between two points of time.

Help in framing suitable policies

Index numbers are more useful to frame economic and business policies. For example, consumer price index numbers are useful in fixing dearness allowance to the employees.

Useful in deflating

Price index numbers are used for connecting the original data for changes in prices. The price index are used to determine the purchasing power of monetary unit.

Compares standard of living

Cost of living index of different periods and of different places will help us to compare the standard of living of the people. This enables the government to take suitable welfare measures.

Special type of average

All the basic ideas of averages are employed for the construction of index numbers. In averages, the data are homogeneous (in the same units) but in index number, we average the variables which have different units of measurements. Hence, it is a special type of average.

Let us sum up

In this unit, you have learned about the following:

- Index numbers are barometers of an Economy.
- It is a specialized average designed to measure the changes in a group of variables over time.
- Widely used index numbers are wholesale price index, consumer price index, index of industrial production, agricultural production index and Sensex.
- The index numbers are indispensable in economic policy making.

Check your progress

1. In simple aggregate method, the aggregate price of all items in the given year is expressed as percentage of the same in the

   (a) current year
   (b) base year
   (c) Quarterly
   (d) half yearly

2. If the index for 1990 to the base 1980 is 250, the index number for 1980 to the base 1990 is

   (a) 4
   (b) 400
   (c) 40
   (d) 4000

3. If Laspeyre's price index is 324 and Paasche's price index is 144, then Fisher's ideal index is:

   (a) 234
   (b) 243
   (c) 261
   (d) 216

Glossary

Index Numbers :	Index numbers are a specialized type of average.
Sensex:	Normal time period in a particular situation.
Price Index Numbers:	Measure and permit comparison of the prices.

Model Questions

1. Discuss the different definitions of index numbers.
2. Explain the applications of index numbers.
3. List out the importance of index numbers.

Answers to check your progress

1. Base year
2. 40
3. 216

Suggested readings

1. Beri, "Business Statistics" Tata McGraw Hill,2ndEd,2009
2. Keller. G, "Statistics for Management", Cengage Learning, 1st Ed, 2009.

Unit-18
Construction of Price Index Numbers

STRUCTURE

Overview

In this unit the construction of different price index numbers like Laspyres, Paasche and Fisher index numbers and also the application problems has been clearly explained.

Objectives

After completion of this unit, you will be able :

- To know the different types of index numbers.
- To classify the different types of price index numbers.
- To discuss the problems related to various price index numbers.

18.1. Introduction

Types of index numbers: The different types of Index Numbers is presented below:

(i) Price Index Numbers

Price index is a 'Special type' of average which studies net relative change in the prices of commodities, expressed in different units. Here comparison is made in respect of prices. Price index numbers are wholesale price index numbers and retail price index numbers.

(ii) Quantity Index Numbers

This number measures changes in volume of goods produced, purchased or consumed. Here, the comparison is made in respect of quantity or volume. For example, the volume of agricultural goods produced, consumed, import, export etc.

(iii) Value Index

Value index numbers study the changes in the total value of a certain period with the total value of the base period. For example, the indices of stock-in-made, purchase, sales profit etc., are analysed here.

18.2. Laspeyre's price Index number

Laspyer's price index number = $\frac{\sum p1q0}{\sum p0q0}$ X 100

P=Price

Q=Quantity

0=Base Year

1=Current Year

P0 = Base Year Price

Q0 = Base Year Quantity

P1 = Current Year Price

Q1 = Current Year Quantity

18.3. Paasche's price Index number

Paasche's price index number = $\frac{\sum p1q1}{\sum p0q1}$ X 100

P=Price

Q=Quantity

0=Base Year

1=Current Year

P0 = Base Year Price

Q0 = Base Year Quantity

P1 = Current Year Price

Q1 = Current Year Quantity

18.4. Fisher's price Index number

Fisher's price index number = $\sqrt{}$ ($\frac{\sum p1q0}{\sum p0q0}$ X $\frac{\sum p1q1}{\sum p0q1}$) X 100 (GM of Laspyer and Paasche)

P=Price

Q=Quantity

0=Base Year

1=Current Year

P0 = Base Year Price

Q0 = Base Year Quantity

P1 = Current Year Price

Q1 = Current Year Quantity

Sample problem

Calculate Laspyre, Paasche, Fisher and Bowleys price index numbers for the following details.

Commodity	2016		2017	
	Price	Quantity	Price	Quantity
A	2	8	4	6
B	5	10	6	5
C	4	14	5	10
D	2	19	2	13

Solution

2016 = base year. 2017 = current year

P0	Q0	P1	Q1	P0Q0	P1Q1	P0Q1	P1Q0
2	8	4	6	16	24	12	32
5	10	6	5	50	30	25	60
4	14	5	10	56	50	40	70
2	19	2	13	38	26	26	38
-	-	-	-	160	130	103	200

Commodity	2016		2017		p_1q_0	p_0q_0	p_1q_1	p_0q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
A	2	8	4	6	32	16	24	12
B	5	10	6	5	60	50	30	25
C	4	14	5	10	70	56	50	40
D	2	19	2	13	38	38	26	26
					Σp_1q_0=200	Σp_0q_0=160	Σp_1q_1=130	Σp_0q_1=103

Laspyer's price index number = $\frac{\sum p1q0}{\sum p0q0}$ X 100 = $\frac{200}{160}$ X 100 = 125 %

Paasche's price index number = $\frac{\sum p1q1}{\sum p0q1}$ X 100 = $\frac{130}{103}$ X 100 = 126.2%

Fisher's price index number = $\sqrt{\left(\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}\right)} \times 100$

$= \sqrt{(1.25 \times 1.262)} \times 100$

$= \sqrt{1.5775} \times 100$

= 1.2559 X 100

= 125.59%

(OR)

Fisher's price index number = $\sqrt{\left(\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}\right)} \times 100$

$= \sqrt{(125 \times 126.2)} = 125.59\%$

Bowley's price index number = $\frac{\frac{\sum p1q0}{\sum p0q0} + \frac{\sum p1q1}{\sum p0q1}}{2}$ X 100

$$\frac{125 + 126.2}{2} = \frac{251.2}{2} = 125.6\%$$

Thus, an index number is a specialized average which helps in comparison of the level of magnitude of a group of related variables with respect to time, geographical location or other characteristics such as production, income, employment, etc. It combines two or more time series variables related to non-comparable units. Index numbers can be used in several ways, such as study trends and tendencies of business

activities, provide guidelines in framing suitable policies, measure real purchasing power of money, help in transforming nominal wage into real wage and so on. The researcher may face various problems in the construction of different types of indices. They may be selection of the base period, collection of data, selection of commodities, choice of averages and weights, selection of an appropriate index.

Let us sum up

In this unit, you have learned about the construction of different price index numbers like Laspyres, Paasche and Fisher index numbers and also the application problems.

Check your progress

1. The index that satisfies factor reversal test is

   (a) Paasche's Index
   (b) Laspeyre's Index
   (c) Fisher's Ideal Index
   (d) Walsh price index

2. The Dorbish - Bowley's price index is the

   (a) Geometric mean of Laspeyre's and Paasche's Price indices
   (b) Arithmetic mean of Laspeyre's and Paasche's Price indices
   (c) Weighted mean of Laspeyre's and Paasche's Price indices
   (d) Weighted mean of Laspeyre's and Paasche's quantity indices

3. The geometric mean of Laspeyre's and Paasche's price indices is also known as

   (a) Dorbish – Bowley's price index
   (b) Kelly's price index
   (c) Fisher's price index
   (d) Walsh price index

Glossary

- Laspyer's price index number = $\frac{\sum p1q0}{\sum p0q0}$ X 100
- Paasche's price index number = $\frac{\sum p1q1}{\sum p0q1}$ X 100
- Fisher's price index number = $\sqrt{}$ ($\frac{\sum p1q0}{\sum p0q0}$ X $\frac{\sum p1q1}{\sum p0q1}$) X 100
- Bowley's price index number = $\frac{\frac{\sum p1q0}{\sum p0q0}+\frac{\sum p1q1}{\sum p0q1}}{2}$ X 100

Model Questions

1. Calculate index number by using Laspeyres method, Paasche method and Fisher's ideal method

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	5	10	11	8
B	8	15	15	10
C	7	6	9	6

2. Evaluate by calculating values of Laspeyres method, Paasche method and Fisher's ideal method

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	2	40	6	50
B	4	50	8	40
C	6	20	9	30
D	8	10	6	20
E	10	10	5	20

3. From the following data calculate Laspeyres method, Paasche method and Fisher's ideal method

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
P	5	10	7	15
Q	12	4	15	6
R	10	8	9	4

Answers to check your progress

1. Fisher's Ideal Index
2. Arithmetic mean of Laspeyre's and Paasche's Price indices
3. Fisher's price index

Suggested readings

1. J. K Sharma, "Business Statistics", Pearson, 2nd Ed,2010.
2. Arora PN &others," Complete Statistical Methods", S. Chand, 3rd Ed,2010
3. Groebner D.F., Shannon P.W., Fry P.C. and Smith K.D. Business Statistics, Pearson Education(2008).

Unit-19
Introduction to Time Series

STRUCTURE

Overview

In this unit applications of time series, definition of time series and components of time series like trend, seasonal variation, cyclical variation and random variation along with mathematical model of time series has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the definitions of time series
- To know the applications time series
- To write the additive and multiplicative model of time series

19.1. Introduction

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

According to Mooris Hamburg “A time series is a set of statistical observations arranged in chronological order”.

Ya-Lun-Chou : “A time series may be defined as a collection of readings belonging to different time periods of some economic variable or composite of variables”.

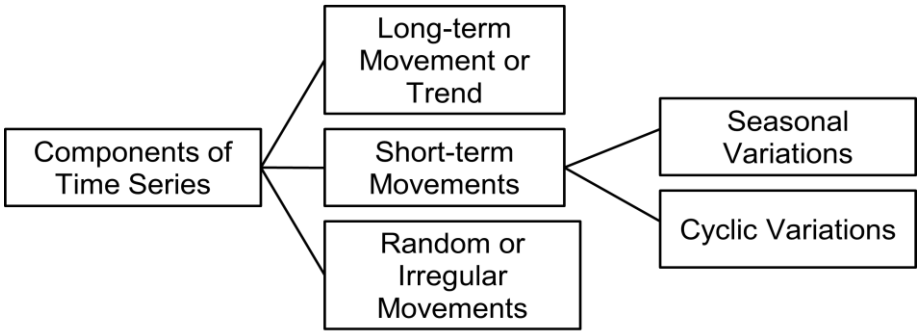
W.Z. Hirsch says “The main objective in analyzing time series is to understand, interpret and evaluate change in economic phenomena in the hope of more correctly anticipating the course of future events”.

19.2. Applications of Time Series

- Time series is used to predict future values based on previously observed values.
- Time series analysis is used to identify the fluctuation in economics and business.
- It helps in the evaluation of current achievements.
- Time series is used in pattern recognition; signal processing, weather forecasting and earthquake prediction.

It can be said that time series analysis is a big tool in the hands of business executives to plan their sales, prices, policies and production.

19.3. Components of Time Series



The components are:

i. Secular trend
ii. Seasonal variation
iii. Cyclical variation
iv. Irregular (random) variation

(i) Secular trend

It refers to the long term tendency of the data to move in an upward or downward direction. For example, changes in productivity, increase in the rate of capital formation, growth of population, etc., follow secular trend which has upward direction, while deaths due to improved medical

facilities and sanitations show downward trend. All these forces occur in slow process and influence the time series variable in a gradual manner.

(ii) Seasonal variation

Seasonal variations are fluctuations within a year over different seasons. Estimation of seasonal variations requires that the time series data are recorded at even intervals such as quarterly, monthly, weekly or daily, depending on the nature of the time series. Changes due to seasons, weather conditions and social customs are the primary causes of seasonal variations. The main objective of the measurement of seasonal variation is to study their effect and isolate them from the trend.

(iii) Cyclical variation

Cyclical variations refer to periodic movements in the time series about the trend line, described by upswings and downswings. They occur in a cyclical fashion over an extended period of time (more than a year).

(iv) Irregular variation

In practice, the changes in a time series that cannot be attributed to the influence of cyclic fluctuations or seasonal variations or those of the secular trend are classified as irregular variations. In the words of Patterson, “Irregular variation in a time series is composed of non-recurring sporadic (rare) form which is not attributed to trend, cyclical or seasonal factors”. Nothing can be predicted about the occurrence of irregular influences and the magnitude of such effects. Hence, no standard method has been evolved to estimate the same. It is taken as the residual left in the time series, after accounting for the trend, seasonal and cyclic variations.

19.4. Mathematical model of Time series

The additive approach

The additive approach is used when the four components of a time series are visualized as independent of one another. Independence implies that the magnitude and pattern of movement of the components do not affect one another. Under this assumption the magnitudes of the time series are regarded as the sum of separate influences of its four components.

$$Y = T + C + S + R$$

Where Y = magnitude of a time series

T = Trend,

C =Cyclical component,

S =Seasonal component, and

R = Random component

In additive approach, the unit of measurements remains the same for all the four components.

The Multiplicative approach

The multiplicative approach is used where the forces giving rise to the four types of variations are visualized as interdependent. Under this assumption, the magnitude of the time series is the product of its four components.

$$Y = T \times C \times S \times R$$

This unit has introduced you to the concept of time series and its analysis with a view to making more accurate and reliable forecasts for the future. A set of quantitative data arranged on the basis of TIME are referred to as 'Time Series'. The analysis of time series is done to understand the dynamic conditions for achieving the short-term and long-term goals of institution(s). With the help of the techniques of time series analysis the future pattern can be predicted on the basis of past trends. The quantitative values of the variable under study are denoted by y1, y2, y3 and the corresponding time units are denoted as x1, x2, x3..... The variable 'y' shall have variations; you will see ups and downs in the values. There are a number of causes during a given time period which affect the variable. Therefore, time becomes the basis of analysis. Time is not the cause and the changes in the values of the variable are not the effect.

Let us sum up

In this unit, you have learned about the applications of time series, definition of time series and components of time series like trend, seasonal variation, cyclical variation and random variation and mathematical model of time series'

Check your progress

1. The component having primary use for long term forecasting is
   (a) cyclical variation
   (b) irregular variation
   (c) seasonal variation
   (d) trend
2. A time series is a set of data recorded
   (a) periodically
   (b) at equal time intervals

(c) at successive points of time
(d) all the above

3. A time series consists of
   (a) two components
   (b) three components
   (c) four components
   (d) five components

Glossary

- **Time Series:** is the data on any variable accumulated at regular time intervals.
- **Secular Trend:** A type of variation in a time series, the long-term tendency of a time series to grow or decline over a period of time.
- **Seasonal variations**: Seasonal variations are fluctuations within a year over different seasons.
- **Cyclical variations:** Cyclical variations refer to periodic movements in the time series about the trend line, described by upswings and downswings. They occur in a cyclical fashion over an extended period of time (more than a year).
- **Irregular variation**: Irregular variation in a time series is composed of non-recurring sporadic (rare) form which is not attributed to trend, cyclical or seasonal factors.

Model Questions

1. Explain Time Series with an example.
2. Explain the components of Time Series.
3. What are the uses of time series analysis?
4. Write the mathematical model of time series.

Answer to check your progress

1. Trend
2. All the above
3. Four components

Suggested readings

1. Groebner D.F., Shannon P.W., Fry P.C. and Smith K.D. Business Statistics, Pearson Education(2008).
2. R.S.N. Pillai, V. Bagavathi," Statistics", S.Chand Limited, 7thEd,2008.

Unit-20

Methods for finding the trend values

STRUCTURE

Overview

In this unit the finding trend values by the different method like graphical method, semi average method, moving average method and method of least squares and also the method to predict the future values using method of least squares has been clearly explained.

Objectives

After completion of this unit, you will be able:

- To understand the concept of time series
- To draw the trend line using graphical method
- To know the upward and downward trends
- To calculate the trend values using semi - average and moving average methods
- To estimate the trend values using method of least squares

20.1. Introduction

In modern times we see data all around. The urge to evaluate the past and to peep into the future has made the need for forecasting. There are many factors which change with the passage of time. Sometimes sets of observations which vary with the passage of time and whose

measurements made at equidistant points may be regarded as time series data. Statistical data which are collected, observed or recorded at successive intervals of time constitute time series data. In the study of time series, comparison of the past and the present data is made. It also compares two or more series at a time. The purpose of time series is to measure chronological variations in the observed data.

In an ever changing business and economic environment, it is necessary to have an idea about the probable future course of events. Analysis of relevant time series helps to achieve this, especially by facilitating future business forecasts. Such forecasts may serve as crucial inputs in deciding competitive strategies and planning growth initiatives.

20.2. Graphical method

Under this method the values of a time series are plotted on a graph paper by taking time variable on the X-axis and the values variable on the Y-axis. After this, a smooth curve is drawn with free hand through the plotted points. The trend line drawn above can be extended to forecast the values. The following points must be kept in mind in drawing the freehand smooth curve.

(i) The curve should be smooth

(ii) The number of points above the line or curve should be approximately equal to the points below it

(iii) The sum of the squares of the vertical deviation of the points above the smoothed line is equal to the sum of the squares of the vertical deviation of the points below the line.

Merits

- It is simple method of estimating trend.
- It requires no mathematical calculations.
- This method can be used even if trend is not linear.

Demerits

- It is a subjective method
- The values of trend obtained by different statisticians would be different and hence not reliable.

20.3. Semi-average method

In this method, the series is divided into two equal parts and the average of each part is plotted at the mid-point of their time duration.

(i) In case the series consists of an even number of years, the series is divisible into two halves. Find the average of the two parts of the series and place these values in the mid-year of each of the respective durations.

(ii) In case the series consists of odd number of years, it is not possible to divide the series into two equal halves. The middle year will be omitted. After dividing the data into two parts, find the arithmetic mean of each part. Thus we get semi-averages.

(iii) The trend values for other years can be computed by successive addition or subtraction for each year ahead or behind any year.

Merits

- This method is very simple and easy to understand
- It does not require many calculations.

Demerits

- This method is used only when the trend is linear.
- It is used for calculation of averages and they are affected by extreme values.

20.4. Moving average method

Moving averages is a series of arithmetic means of variate values of a sequence. This is another way of drawing a smooth curve for a time series data.

Moving averages is more frequently used for eliminating the seasonal variations. Even when applied for estimating trend values, the moving average method helps to establish a trend line by eliminating the cyclical, seasonal and random variations present in the time series. The period of the moving average depends upon the length of the time series data.

The choice of the length of a moving average is an important decision in using this method. For a moving average, appropriate length plays a significant role in smoothening the variations. In general, if the number of years for the moving average is more than the curve becomes smooth.

Merits

- It can be easily applied
- It is useful in case of series with periodic fluctuations.
- It does not show different results when used by different persons

- It can be used to find the figures on either extremes; that is, for the past and future years.

Demerits

- In non-periodic data this method is less effective.
- Selection of proper 'period' or 'time interval' for computing moving average is difficult.
- Values for the first few years and as well as for the last few years cannot be found.

Moving averages odd number of years (3 years)

To find the trend values by the method of three yearly moving averages, the following steps have to be considered.

1. Add up the values of the first 3 years and place the yearly sum against the median year. [This sum is called moving total]
2. Leave the first year value, add up the values of the next three years and place it against its median year.
3. This process must be continued till all the values of the data are taken for calculation.
4. Each 3-yearly moving total must be divided by 3 to get the 3-year moving averages, which is our required trend value.

Moving averages - even number of years (4 years)

1. Add up the values of the first 4 years and place the sum against the middle of 2nd and 3rd year. (This sum is called 4 year moving total)
2. Leave the first year value and add next 4 values from the 2nd year onward and write the sum against its middle position.
3. This process must be continued till the value of the last item is taken into account.
4. Add the first two 4-years moving total and write the sum against 3rd year.
5. Leave the first 4-year moving total and add the next two 4-year moving total and place it against 4th year.
6. This process must be continued till all the 4-yearly moving totals are summed up and centered.
7. Divide the 4-years moving total by 8 to get the moving averages which are our required trend values.

Check Your Progress-1

True/False

a. Moving averages is less frequently used for eliminating the seasonal variations.

b. In modern times we see data all around.

c. Analysis of relevant time series helps to achieve this, especially by facilitating future business forecasts.

d. In general, if the number of years for the moving average is more than the curve becomes smooth

e. Among the four components of the time series, secular trend represents the long term direction of the series.

20.5. Method of least squares

Among the four components of the time series, secular trend represents the long term direction of the series. One way of finding the trend values with the help of mathematical technique is the method of least squares.

This method is most widely used in practice and in this method the sum of squares of deviations of the actual and computed values is least and hence the line obtained by this method is known as the line of best fit. It helps for forecasting the future values.

It plays an important role in finding the trend values of economic and business time series data.

- The line y=a+bx found out using the method of least squares is called 'line of best fit'.
- Normal equations involved in the method of least squares are

$$\sum_{i=1}^{n} Y_i = na + b \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} X_i Y_i = a \sum_{i=1}^{n} x + b \sum_{i=1}^{n} x_i^2$$

Merits

- The method of least squares completely eliminates personal bias.
- Trend values for all the given time periods can be obtained.
- This method enables us to forecast future values.

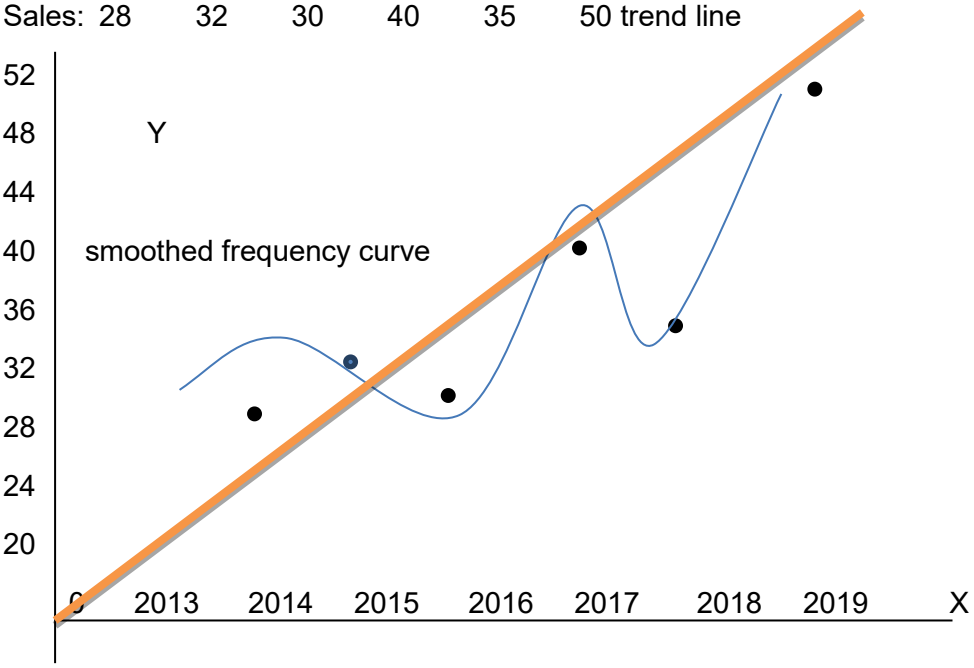
Demerits

- The calculations for this method are difficult compared to the other methods.
- Addition of new observations requires recalculations.
- It ignores cyclical, seasonal and irregular fluctuations.
- The trend can be estimated only for immediate future and not for distant future.

Problem: 1

Calculate trend values by graphical method.

Year:	2013	2014	2015	2016	2017	2018
Sales:	28	32	30	40	35	50



Year:	2013	2014	2015	2016	2017	2018	2019
Trend values:	28	32	36	40	44	50	52

Problem: 2

Calculate trend values by semi average method.

Year:	2013	2014	2015	2016	2017	2018
Sales:	28	32	30	40	35	50

Year	Sales	Semi-total	Semi-average(Trend values)
2013	28		30-3.9 =26.1
2014	32	90 divided by 3	30
2015	30		30+3.9 = 33.9
2016	40		33.9+3.9 = 37.8
2017	35	125 divided by 3	41.66
2018	50		41.66+3.9 =45.56

Difference between the year 2017 and 2014 = 3 years

3 yearly increment = 41.66- 30 = 11.66

Annual increment = $\frac{\text{3 yearly increment}}{3} = \frac{11.66}{3} = 3.88 = 3.9$

Trend for the year 2015 = trend for 2014+ annual increment

= 30+3.9 =33.9

Year:	2013	2014	2015	2016	2017	2018
Sales:	28	32	30	40	35	50
Trend values:	26.1	30	33.9	37.8	41.7	45.6

Problem: 3

Find trend value by 3 yearly moving average method.

Year:	2013	2014	2015	2016	2017	2018	2019
Sales:	28	32	30	40	35	50	55

Year	Sales(Y)	3 yearly moving total	3 yearly moving average(Trend values)
2013	28	-	-
2014	32	90 divided by 3	30
2015	30	102	34
2016	40	105	35
2017	35	125	41.7
2018	50	140	46.7
2019	55	-	-

Problem: 4

Find trend values by 4 yearly moving average method.

Year:	2013	2014	2015	2016	2017	2018	2019
Sales:	28	32	30	40	35	50	55

Year	Sales(Y)	4 yearly moving total	Centered 4 yearly moving total	Centered 4 yearly moving average(Trend values)
2013	28			-
2014	32			-
		130		
2015	30		267 divided by 8	33.37
		137		
2016	40		292	36.50
		155		
2017	35		335	41.87
		180		
2018	50			-
2019	55			-

Problem: 5

Calculate trend values by the method of least squares.

Year:	1993	94	95	96	97	98
Sales:	12	13	15	10	16	18

Also find the sales for the year 2000.

Solution:

Straight line trend equation

Y = a + b X

Here a = middle year trend = $\frac{\sum Y}{n}$

: b= annual increment = $\frac{\sum XY}{\sum X^2}$

Year	Y (values)	X=year-mid year X=year-1995.5	X^2	XY	Trend values
1993	12	-2.5	6.25	-30.00	11.57
1994	13	-1.5	2.25	-19.50	12.54
1995	15	-0.5	0.25	-07.50	13.51
1996	10	0.5	0.25	05.00	14.48
1997	16	1.5	2.25	24.00	15.45
1998	18	2.5	6.25	45.00	16.42
TOTAL	**84**	**0**	**17.50**	**17.00**	**84.00**

a = $\frac{\sum Y}{n}$ = $\frac{84}{6}$ = 14

: b= $\frac{\sum XY}{\sum X^2}$ = $\frac{17}{17.5}$ = 0.97

Y = a + b X

Y = 14 + 0.97(year – 1995.5)

When year=1993

Y=14+0.97(1993-1995.5)

Y= 14+ 0.97(-2.5)

Y = 14-2.43

Y=11.57

When year 2000

Y=14+0.97(2000-1995.5)

Y= 14 +4.37

Y = 18.37

Let us sum up

In this unit, you have learned about the following:

The secular trend is the main component of a time series which results from long term effects of socio-economic and political factors. This trend may show the growth or decline in a time series over a long period. This is the type of tendency which continues to persist for a very long period. Prices and export and import data, for example, reflect obviously increasing tendencies over time.

Check your progress

1. Business forecasts are made on the basis of

   (a) future data

   (b) past data

   (c) tax regulations

   (d) Government policies

2. The four components of time series in a multiplicative model are

   (a) independent

   (b) interdependent

   (c) constant

   (d) additive

3. In the least square theory the sum of squares of residuals is

   (a) zero

   (b) minimum

   (c) constant

   (d) maximum

Glossary

- **Time series :** Time series is a time oriented sequence of observations.
- **Components of time series :** Components of time series are secular trend, seasonal variations, cyclical variations and irregular (erratic) variations
- **Methods of calculating trend values:** Methods of calculating trend values are graphical method, semi - averages method, moving averages method, and method of least squares.
- '**Line of best fit':** The line y = a + b x found out using the method of least squares is called 'line of best fit'.

Model Questions

1. Find the trend values by 3 yearly moving average method.

Year:	1993	1994	1995	1996	1997	1998	1999
Values:	20	24	35	36	40	45	50

2. Find the trend values by the method of 4 yearly moving averages.

Year:	1993	1994	1995	1996	1997	1998	1999
Values:	20	24	35	36	40	45	50

3. Calculate the trend values by the method of least squares.

Year:	2013	2014	2015	2016	2017
Y:	50	65	75	100	120

Answers to check your progress-1

a-False

b-True

c-True

d-True

e-True

Answers to check your progress-2

1. Past data

2. Independent

3. Zero

Annexure - Case Studies

1. Annual wage distribution of workers in a factory is given. Find out Arithmetic Mean, Median, Range, Standard deviation and mean deviation of the following.

(Annual wages) Class interval	(No. of workers) Frequency (f)
50 – 80	5
80 – 110	12
110 – 140	26
140 – 170	30
170 - 200	15
200– 230	8
230- 260	4

2. Calculate the Karl Pearson's correlation coefficient between the marks (out of 10) in statistics and mathematics of 6 students.

Student	1	2	3	4	5	6
Statistics	7	4	6	9	3	8
Mathematics	8	5	4	8	3	6

3. Suppose we have ranks of 5 students in three subjects Computer, Physics and Statistics and we want to test which two subjects have the same trend.

Rank in Computer	2	4	5	1	3
Rank in Physics	5	1	2	3	4
Rank in Statistics	2	3	5	4	1

4. In a correlation analysis, between production (X) and price of a commodity (Y) we get the following details. Variance of X = 36. The regression equations are:

   $12X – 15Y + 99 = 0$ and $60 X – 27 Y = 321$.

   Calculate (a) The average value of X and Y. (b) Coefficient of correlation between X and Y.

5. Construct the cost of living index number of the year 2014 using family budget method.

Expenses	%	base year (2000)	current year (2004)
Food	40	150	174
Rent	15	50	60
Clothing	15	100	125
Fuel	10	20	25
Misc	20	60	90

6. The following data describes the export quantity of a company.

Year	1995	1996	1997	1998	1999	2000	2001
Export (in millions)	12	13	13	16	16	19	23

Fit a straight line trend and estimate the export for the year 2005.

7. A test was conducted with 6 students before and after the training programme. Their marks were recorded and tabulated as shown below. Test whether the training was helpful in improving their scores.

Before training	100	160	113	122	120	105
After training	120	155	120	128	115	100

8. In a sample of 200 households in a colony, two brands of hair oils A and B are applied by 90 females. Further, 60 females and 70 males are using brand A. To test whether there is any association between the gender and brand of hair oil used, by constructing a contingency table.

9. Assess the following table with given number of subscribers added to four major telecom players in India. Use one way ANOVA to find the difference in number of subscribers between the telecom players (in Lakhs)

Companies			
Airtel	**BSNL**	**Reliance**	**Vodafone**
6	6	2	5
7	6	2	3
7	6	6	4
7	8	7	4

Model End Semester Examination Question Paper

Master of Business Administration (MBA)

Course Code- **DCMBA-12 :** Course Title- **Business Statistics**

Max. Marks: 70 Time: 3 hours

PART – A (10x2 =20 Marks)

Answer any **TEN** questions out of TWELVE questions

[All questions carry equal marks]

1. Detail about Coefficient of Variation.
2. What are Primary and Secondary data?
3. Differentiate Independent and Dependent events with an example.
4. Discuss about multiplication theorem of probability
5. Brief the utility of regression analysis.
6. Write the properties of correlation coefficient.
7. Explain rank correlation with formula.
8. Explain Sample and Population with examples.
9. Write a short note on parameter and statistic.
10. Write a brief note on Chi-Square test.
11. Describe any five uses of index numbers.
12. Write note on Fisher's price index number.

PART – B (5X8=40 Marks)

Answer any **FIVE** questions out of SEVEN questions

[All questions carry equal marks]

13. Compute Arithmetic mean & Median.

Wages (Rs)	20-30	30-40	40-50	50-60	60-70
No. of labourers (f)	3	5	20	10	5

14. One card is drawn from a standard pack of 52. What is the probability of getting either a king or queen? Use addition law of probability for independent events.
15. If a player plays a game of chance where he can win Rs.1000 with probability 0.5, win Rs.500 with probability 0.3 and lose Rs.3000 with probability 0.2, what is his expected gain in one play of the game?
16. Determine the Karl Pearson correlation coefficient between the following observations to find the relationship.

X	3	5	6	7	9	12
Y	20	14	12	10	9	7

17. Utilize Chi-square test find out whether the opinion towards autonomous status depends upon level of graduation [UG and PG]. You are given Chi-square table value for Degrees of Freedom 1 at 5% Level of Significance as 3.841.

Particulars	Under Graduates	Post Graduates
Favoring autonomous	190	200
Unfavoring autonomous	110	100

18. Calculate index number by using Laspeyres method, Paasche method and Fisher's ideal method

Commodity	Base year		Current year	
	Price	Quantity	Price	Quantity
A	5	10	11	8
B	8	15	15	10
C	7	6	9	6

19. Fit a straight line trend equation by method of least square and estimate the sales values of 2019 to 2023.

Year	2014	2015	2016	2017	2018
Sales (Rs. In Cr)	69	60	87	95	102

PART - C (1x10=10 Marks)

CASE STUDY (Covering the Whole Course)

20. Assess the following table with given number of subscribers added to four major telecom players in India. Use one way ANOVA to find the difference in number of subscribers between the telecom players (in Lakhs)

Companies			
Airtel	BSNL	Reliance	Vodafone
6	6	2	5
7	6	2	3
7	6	6	4
7	8	7	4

Ouriginal

Document Information

Analyzed document	ODL-BUSINESS STATISTICS-SLM(new).docx (D162486595)
Submitted	3/29/2023 7:22:00 AM
Submitted by	Thirumagal
Submitter email	tmagal.sms@velsuniv.ac.in
Similarity	10%
Analysis address	tmagal.sms.vels@analysis.urkund.com

Sources included in the report

W	URL: https://icmai.in/upload/Students/Syllabus2016/Foundation/Paper-4-16092021.pdf Fetched: 11/12/2021 12:47:32 AM	37
W	URL: https://mis.alagappauniversity.ac.in/siteAdmin/dde-admin/uploads/3/UG_B.B.A_English_10432-BUSI... Fetched: 12/29/2022 6:56:26 PM	38
W	URL: https://www.slideshare.net/neerazz/measure-of-dispersion-by-neeraj-bhandari-surkhetnepal Fetched: 3/12/2020 5:03:05 PM	2
W	URL: https://mu.ac.in/wp-content/uploads/2021/03/Statistics-and-Probability_Full-File.pdf Fetched: 4/13/2022 8:25:13 PM	16
W	URL: https://www2.karlin.mff.cuni.cz/~nagy/NMSA202/dice1.pdf Fetched: 9/18/2020 4:58:04 PM	1

Entire Document

Block-1 Introduction Unit -1 Introduction to Statistics
Structure Overview 1.0 Objectives 1.1 Origin and growth of statistics 1.2 Definition of Statistics 1.3 Characteristics of statistics 1.4 Importance and Scope of Statistics 1.5 Limitations of Statistics 1.6 Descriptive and Inferential statistics Let us sum up Check your Progress Glossaries Model Questions Answers to check your progress Suggested Readings
Overview In this unit contains the origin and growth of statistics, definition of statistics and limitations statistics. Also contains the details about descriptive statistics and inferential statistics.
1.0 Objectives • Highlights the origin and growth of statistics • Introduces the meaning and definitions of statistics • Presents the scope and functions of statistics • Explains the applications of statistics in different fields
1.1. Origin and growth of statistics The origin of statistics can be traced back to the primitive man, who put notches on trees to keep an account of his belongings. During 5000 BCE, kings used to carry out census of populations and resources of the state. Kings of olden days made their crucial decisions on wars, based on statistics of infantry, cavalry and elephantary units of their own and that of their enemies. Later it enhanced its scope in their kingdoms' tax management and administrative domains. Thus, the word 'Statistics' has its root either to
Latin word 'Status' or

Marching Beyond 30 Years Successfully

43 Institutions, 36,000 Students 6000 Staff Members

VELS

INSTITUTE OF SCIENCE, TECHNOLOGY & ADVANCED STUDIES (VISTAS)
(Deemed to be University Estd. u/s 3 of the UGC Act, 1956)
PALLAVARAM - CHENNAI
www.vistas.ac.in

VELS
MEDICAL COLLEGE & HOSPITAL
Under VELS INSTITUTE OF SCIENCE, TECHNOLOGY & ADVANCED STUDIES (VISTAS)
(Deemed to be University Estd. u/s 3 of the UGC Act. 1956)
Uthukottai Taluk, Tiruvallur District - 601 102
www.velsmedicalcollege.com

MAHAVIR INSTITUTE OF MEDICAL SCIENCES (MIMS)
VIKARABAD, HYDERABAD- TELANGANA
www.mahavirmedicalcollege.org

SRI VENKATESWARA
DENTAL COLLEGE & HOSPITAL
Thalambur, Chennai
www.svdentalcollege.com

MEGHNA INSTITUTE OF DENTAL SCIENCES (MIDS)
NIZAMABAD, HYDERABAD - TELANGANA
www.meghnadentalcollege.com

VEL
NURSING COLLEGE
Under VELS INSTITUTE OF SCIENCE, TECHNOLOGY & ADVANCED STUDIES (VISTAS)
(Deemed to be University Estd. u/s 3 of the UGC Act. 1956)
www.velnursingcollege.com

VENKATESWARA
NURSING COLLEGE
Thalambur, Chennai
www.venkateswaranursing.ac.in

VELS
VIDYASHRAM
CBSE
Pallavaram - Thalambur - Cantonment
www.velsvidyashram.com

VELS GLOBAL SCHOOL
TAMIL NADU | KARNATAKA | DELHI - NCR | WEST BENGAL
www.velsglobalschool.com

VAELS
INTERNATIONAL SCHOOL
KKIC / CISCE / CAMBRIDGE INTERNATIONAL
NEELANKARAI - INJAMBAKKAM
www.vaelsinternationalschool.com

KINDLE KIDS
INTERNATIONAL SCHOOL
SLF Building, 510 Thomson Road #B2-00, Singapore
www.kindlekids.sg

VELS
KINDER KIDS
No. 9, Dr. Ranga Road,
Mylapore, Chennai
www.velskinderkids.com

CTC
Cambridge Tutors College
LONDON
South Croydon, Greater London
www.ctc.ac.uk

SHRI ISARI VELAN MISSION HOSPITAL
A Comfort Care Centre
150 Bedded Multi Speciality Hospital
www.velanhospital.com