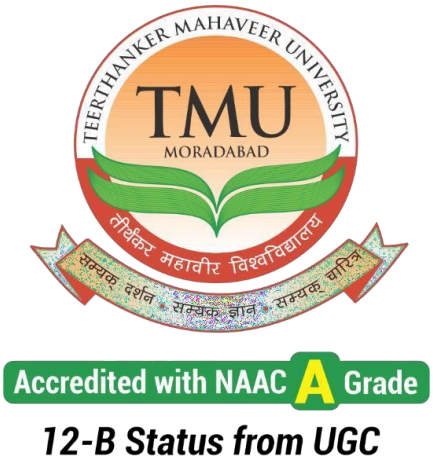
TEERTHANKER MAHAVEER UNIVERSITY
MORADABAD, INDIA

CENTRE FOR ONLINE & DISTANCE LEARNING



Programme: Bachelor of Commerce

Course: Business Statistics

Course Code: BCPCC301

Semester-III

Business statistics

Objective: the course will enable the students to understand statistics, how and when to apply statistical techniques to decision making situations and how to interpret the results.

Unit-I

Statistics : Definition, Importance & Limitation, Collection of data, classification and presentation of frequency distribution

Measures of Central Tendency and Dispersion : Meaning and objectives of measure of central tendency- arithmetic mean, median, mode, geometric mean and harmonic mean, characteristics, applications and limitations of these measures; Measure of variation, range, quartile deviation, mean deviation and standard deviation, coefficient of variation.

Unit II

Correlation and Regression : Meaning of correlation , types of correlation positive and methods of studying correlation, Lines of regression, co-efficient of regression, standard error of estimate.

Unit III

Index numbers and Time series : Index number and their uses in business; construction of simple and weighed price, quantity and value index numbers, test for an ideal index number, Components of Time series - secular trend, cyclical, seasonal and irregular variations. Methods of estimating secular trend and seasonal indices : use of time series in business

Unit IV

Probability & Probability Distributions : Definition, Basic concepts, Events and experiments, random variables, expected value Types of probability: Classical approach, Relative frequency approach or empirical probability, Subjective approach to probability**,** Theorems of probability: Addition theorem, Multiplication theorem, Conditional probability, Bays Theorem,

Suggested Readings:

1. Sancheti and Kapoor V.K,Statistics Theory, Methods & Application, Sultan Chand & Sons,
2. R.P.Hooda, Introduction to Statistics, Macmillan,
3. S.C. Aggarwal & R.K Rana, Basic Statistics for Economists: V.K. India.
4. Lewin and Rubin, Statistics for Management, Prentice-Hall of India, New Delhi
5. S P Gupta Statistical Methods Sultan Chand
6 Beri, Business Statistics Tata Mc Graw Hill
7 Chandan J S,Statistics for Business and Economics Vikas Publications

Table of Contents

Chapter 1: Statistics and Data Collection

In the epic Ramayana, king Rama when decided to go to war with Lanka king Ravana to free his wife Sita, he asked his ardent follower Hanuman to visit state of Lanka and get information about the whereabouts of his wife and also the strength of Ravana's army. Everybody including Ram had heard that Ravana along with his brothers and a big army was very powerful and almost insurmountable. The purpose of Hanuman's visit was to gather primary information on which Ram's decision was to be based. His decision was not based on the available data of what he and everybody has heard about power of Ravana.

In another incident, when king Ram's brother Lakshman in the war got critically injured, then Hanuman was sent to procure the medicine (Sanjivni) from a far off mountain. He didn't have the accurate data about which medicine from the specified mountain is required. So, he made indecision by bringing the entire mountain. In this case, the decision was made on incomplete information obtained from incomplete data. This indicates that proper decision by a manager can be made by having only proper information which is obtained by gathering reliable data.

1.0 Objectives

The readers of this chapter should be able to understand

- Meaning and significance of statistics
- Application of statistics
- Types of data and its usage
- The primary and secondary sources of data and some fundamental ways of collecting data
- Difference between descriptive and inferential statistics

1.1 Introduction

Statistics is concerned with understanding the real world through the information that we derive from classification and measurement of data. Its distinctive characteristic is that it deals with variability and uncertainty which is everywhere. This gives it its fundamental role and is the reason why its tentacles reach into every corner of the scientific enterprise. Variability and uncertainty are two sides of the same coin and, together, are the hallmark of a statistical problem. It is this which gives statistics right of entry to almost every sphere of human endeavor.

For instance, if somebody states to a garment manufacturer that 'December was colder than usual this year' then such statement leads to lot of variability and uncertainty. This does not tell the meaning of usual. Was this years' December winter was compare with last year or previous ten years. Because variability in winter data would change with change in time period. Also, can manufacturer reach to a conclusion from such a statement that next year winters would be harsh in December and the company should manufacture more of sweaters? To derive correct information and make a decision accurately manufacturer has to rely on proper and reliable collection of data and apply relevant statistical tools to reach a conclusion. Thus, statistics involves extraction of information from collected data.

Similarly, consider a statement 'the store is facing stiff competition from mom and pop stores'. Was the store a small sized store or a hyper mart? What was its location: residential, commercial or suburbs? What was the targeted population: high income, low income, youngsters or families? Should increased competition lead the store to increase promotion, lower its prices or shut shop? This indicates that every decision is bereft with variability and uncertainty. To deal with this aspect a manager relies on information which in turn is derived from data. Now, a manager confronts with huge amount of data on daily basis. How data is differentiated from information or relevant information which helps in decision making is extracted from huge amount of available data? Data is a combination of noise and information and statistics facilitates to segregate noise and information and obtain information for manager or decision maker. Thus, most of the businesses are affected by uncertainty and variability. So statistics is important to almost all business decisions.

1.2 Scope of statistics

Statistics is conveniently defined as that branch of mathematics which is concerned with the measurement and description of the variability to be found amongst the members of large populations. Elementary courses based on this definition deal with frequency distributions, measures of average and dispersion, correlation and regression analysis, and the use of samples to obtain information about populations. Advanced courses could use different labels and deal with all these topics in more depth. They could include stochastic models, statistical inference and multivariate analysis. If statistics is to be taught as an application of mathematics, it will be essential to emphasize the generality of statistical methods by using examples from more than one field of application.

Statistics is an abstract science as it is predominantly used to reach to a conclusion regarding population by generalizing the sampled data. To make a generalized viewpoint from an abstract observation in social sciences, scientists use mathematical models to describe relationships between the variables. The interpretation of the model and its parameters depends on the nature of the relationships being considered as well as the mathematical form of the model. Any problem concerning a specific set of population is studied by researcher by using the observed data. The social sciences make more use of statistical regression analysis than functional relationships and the parameters of their models generally vary across time and place. Validity of the model is a problem in social sciences as studying behavioral problems at work or of consumer is a difficult task. So, social sciences model work under lot of assumptions and are valid over narrow range of variables as these variables change over time and place. This makes the applicability of certain model limited and different models can be used to study the same phenomenon. Thus, importance of statistics gets amplified to make relevance of such inexact problems.

Taking this aspect of social sciences into consideration all formulated and advised models should be regarded as approximate descriptions of the phenomena they are being used to describe.

Some examples from different field of management to understand significance of statistics:

Electric Bikes: Business Statistics an assist companies in their marketing efforts. With continuous increase in fuel prices consumers were looking for alternative fuels to power their vehicles. Capturing this opportunity many auto companies ventured into the manufacturing of electronic bikes. Demographic data of urban consumers like family size, user of two wheeler, income group and age etc. could be obtained and then by using statistical methods like cross-tabulation company can reach to a conclusion that which income group people with small family size would like to buy such bikes. Further surveys can be conducted to understand the consumers' preferences regarding features and price of electronic bikes. It was observed that people have not adopted these bikes wholeheartedly. A company which wanted to be a major player in this sector found that lack of power in the bikes; intermittent electricity supply and inadequate carrying space were major reasons in non-acceptance of such bikes. There could be numerous reasons but statistics helped to find major reasons in quantitative terms which helped mangers to make informed decisions.

Shampoo sachets: FMCG companies selling shampoos, soaps and toiletries were very keen to enter into rural market to give impetus to their flagging sales from urban market. Data were collected through both secondary and primary sources from major villages of the country from each state. Data regarding people's occupation, their family size, intermittent source of income, role of women in household, disposable income in the hand of women etc. were obtained. From such huge data information could be extracted by using statistical tools like frequency charts to understand the product that women buy first and most if they have disposable income. As income is dependent on crop yield so by using regression analysis it was revealed that small packages like sachets and promotion in vernacular languages were key to increase awareness among rural consumers.

Stock Market Investments: Investing in various financial instruments for the clients and luring these clients to their investment products is an important decision that finance managers of a finance company makes. To understand

the nature of behavior that consumers present while investing for saving and earning purposes a finance manager collected data regarding investment done in various instruments available for last ten years (2002-2012). From such huge data manager was able to extract information by using statistics. It was found that in earlier part of the selected time period people were investing heavily in stock market as Indian economy was booming. People even took loans to invest in the market. But this period of hype remained for 2-3 years. Then people started investing in real estate and prices of real estate reached astronomical levels. After 3-4 years because of increase in inflation and economy going through a slowdown people became wary and started investing in gold and silver as hedge. By using statistics various reasons for this change in behavior can also be studied.

Business statistics has numerous applications in various fields of management. These would be discussed in subsequent sections and chapters.

1.3 Types of Data

Information in raw or unorganized form that refers to conditions or ideas is called as data. Data does not directly affect the concerned subject/party or person. The meaningful aspect of the data extracted by using certain statistical technique is called information. Information is accurate and presented within a context in an organized format. Information affects decision making. For example, a good monsoon leading to higher yield in crops and thereby increased income for farmers might be data for an urbanized dweller, but it can be information for a FMCG manager for whom a farmer is a prospective customer.

Data can be broadly categorized as qualitative and quantitative. Higher the rainfall better would be crop yield and higher would be the income of farmers. But how much rainfall would lead to how much increase in crop yield and it would result in how much increase in income that would allow farmers to spend on disposables. The former data is qualitative in nature as it approximates the data whereas latter data is quantitative in nature as it is definite and can be quantified and verified.

It is been emphasized in previous sections that informed and accurate decision depends on how statistics is used to segregate information from data by minimizing the impact of noise. The first step in the entire process is collection of data regarding concerned problem. So, one of the prominent aspects in statistical analysis is understanding of type of data to be collected. Different types of data can be used and interpreted in different manner. For instance, data regarding a students' roll number in the class, his grade in subject of statistics and his marks in the subject are different from each other and are analyzed differently. All such data should not be analyzed the same way statistically because the entities represented by numbers are different. For this reason, researchers need to know the level of data measurement represented by numbers being analyzed. Four common levels of data measurement have been discussed:

1.3.1 Nominal data

Numbers that are used only to classify or categorize represent nominal data. For example, jersey number of players, car registration numbers, roll number of students etc. Such data does not provide a value statement.

Student with roll number one does not indicate his/her superiority over roll number ten student in any way. Many demographic questions in surveys result in data that are nominal because the questions are used for classification purpose only. For instance, educational qualification asked by a researcher in a survey assigned '1' to graduate, '2' to post-graduate and '3' to above post graduate. The assignment of 3 does not in any way is used to imply that such a respondent is better or more qualified than a respondent with assignment '1'. This is done just for classification purposes and is done to differentiate a respondent from another. Few statistical techniques are applicable for such data such as chi-square and some other non-parametric tests. PAN card numbers, bank account numbers, ZIP codes, etc. are some other examples of nominal data. Nominal data is sometimes also used to represent qualitative data.

1.3.2 Ordinal data

Ordinal data involve ranking data in a particular order. The order can be either ascending or descending. For instance, if in a class of statistics course a student gets A grade and other B and another student C grade, then it can be interpreted that A grade student has got more marks then the students who have got B and C grade. Similarly student with B grade has scored more than student with C grade but less than student with A grade. But ordinal data does not indicate the interval between data. Student who has got A grade is better than student with B grade but he has scored how much higher than other is not told by ordinal data. In various surveys five point Likert scale is used to measure a variable. For instance, consider a restaurant which wants to find out the customers' response towards taste of the meal measured on a five point Likert scale designed from Very Good, Good, Average, Bad and Very Bad. If scale is arranged in such a way then the data to be measured is arranged in an order where Very Good indicates higher score than Very Bad. But the scale does not indicate the difference between the ordered data i.e. the quantitative difference between very good and good is not indicated in the ordinal data. Various statistical tools are applicable for ordinal data such as mean, median to describe the data. But mode cannot be applied as mode indicates only highest frequency occurring data which is applicable mostly for nominal data.

1.3.3 Interval data

Interval data involves assigning quantitative differences between two consecutive numbers. The differences represented between consecutive numbers are equal i.e. interval data have equal intervals. For instance in the above example of restaurant if Very Good is assigned '1', Good as '2', Average as '3', Bad as '4' and Very Bad as '5' then the difference between two consecutive value is one which is equal between every interval. Interval data can also be applied in the case of students' grade illustration. If corresponding to each students' grade is given their marks then the difference between marks of students under study can be calculated showing the difference between two data values. Thus, roll numbers indicate nominal data, grades ordinal data and marks interval data. This kind of data is most appropriate for various statistical techniques such as regression, t test, Analysis of Variance etc. as it is easy to study quantitative values than qualitative values.

1.3.4 Ratio data

Ratio data have all the properties of interval data but it has an absolute zero and ratio of two numbers is calculated with certain meaning. The notion of zero means that base value is fixed and it cannot be arbitrarily assigned. For example if a person has weight of 50 kgs and another has 100 kgs, then it can be interpreted by using ratio that second persons' weight is two times more than the first person. In this example, calculated ratio has a meaning whereas if in case of interval data ratio is calculated it would result in meaningless results. For example, in the restaurant example if one respondent gives taste of the meal as '1' indicating very good and another respondent gives '2' indicating good and then ratio of their responses is found i.e. 1 / 2 or 2/1. The ratio values does not have any meaning and cannot be interpreted. Other examples of ratio data involve, production cycle time, race time by athletes, number of shoes sold by two different stores etc. As ratio data is a metric data so this data can be analyzed by any of the statistical techniques.

Exercise 1

1. Classify each of the following as nominal, ordinal, interval or ratio data:
   (i) Time required to produce a tyre
   (ii) Liters of milk a family drinks per month
   (iii) Ranking of four courses in the university designated as excellent, good, satisfactory and poor.
   (iv) Age of employees
   (v) An employees' identification number
   (vi) Response time of an emergency unit.

1.4 Data Collection

This chapter in the beginning discussed the importance of collection of data for accurate decision making by applying statistics. The collected data should pertain to the research problem at hand. Before collecting data through primary sources like surveys it is important to get data through secondary sources like published articles to understand the research problem.

1.4.1 Secondary data

Secondary data is the data which got its origin for some other purpose. For instance, if research problem at hand was to find reasons for decrease in sales of washing machine which requires demographic data of consumers in a particular geographic area. But another study was already conducted involving determining sources of finance consumers use to buy washing machines from bank records. For this study the demographic data had been collected. So, this demographic data becomes a source of secondary data for research problem at hand.

Advantages of Secondary Data

The most significant advantages of secondary data are the cost and time economies they offer the researcher. If the required information is available as secondary data then collection of it should take no more than a few days and would involve little cost. On the other hand, a field survey would involve time and cost for creating, testing, delivering, collecting data; data coded, punched, and tabulated. With secondary data expenses have been incurred by the original source of the information and do not need to be borne by the user. Expenses are shared by the users of commercial sources of secondary data, but even here the user's costs will be much less than they would be if the firm collected the same information itself. These time and cost economies prompt the general caution: Do not bypass secondary data. Begin with secondary data, and only when the secondary data are exhausted or show diminishing returns, proceed to primary data. Thus, secondary data would typically

- help to better state the problem under investigation
- suggest improved methods or data for better coming to grips with the problem
- provide comparative data by which primary data can be more insightfully interpreted.

Disadvantages of Secondary Data

When confronted by a new problem, the researcher's first attempts at data collection should logically focus on secondary data. Secondary data are statistics gathered for some other purpose, in contrast to primary data, which are collected for the purpose at hand. Certain disadvantages associated with secondary data are discussed below.

- Secondary data possess significant cost and time advantages and it is only when their pursuit shows diminishing returns and the problem is not yet resolved that the researcher should proceed to primary data.
- Since secondary data are collected for other purposes, it will be rare when they fit perfectly the problem as defined. In some cases, the fit will be so poor as to render them completely inappropriate.
- It is not uncommon for secondary data to be expressed in units different from those deemed most appropriate for the project. Size of retail establishment, for instance, can be expressed in terms of gross sales, profits, square feet, and number of employees. Consumer income can be expressed by individual, family, households, and spending unit.
- Finally, secondary data quite often lack publication currency. The time from data collection to data publication is often long, sometimes as much as three years, for example, as with much government census data. While census data have great value while current, this value diminishes rapidly with time, as many marketing decisions require current, rather than historical, information.

1.4.2 Types of Secondary Data

There are a number of ways by which secondary data can be classified. One of the most useful is by source, which immediately suggests the classification of internal and external data.

Internal data are those found within the organization for whom the research is being done, while external data are those obtained from outside sources. The sales and cost data compiled in the normal accounting cycle represent promising internal secondary data for many research problems. This is particularly true when the problem is one of evaluating past marketing strategy or of assessing the firm's competitive position in the industry. It is less helpful in future directed decisions, such as evaluating a new product or a new advertising campaign. Two of the most significant advantages associated with internal secondary data are their ready availability and low cost.

The ***external*** sources can be further split into those that regularly publish statistics and make them available to the user at no charge, and those organizations that sell their services to various users. The many standardized marketing information services that are available are another important source of secondary data for the marketing researcher. These services are available at some cost to the user and in this respect are a more expensive source of secondary data than published information. This section reviews some of the main types and some of the main sources of standardized marketing information service data.

Industry Services are services available to the consumer goods manufacturer than to the industrial goods supplier. The consumer goods services are also much older than the industrial goods services. For instance, whereas the Nielsen Retail Index dates from 1934, the industry information services were born in the 1960s. This means that the industrial goods services are still evolving in terms of the type of information being collected and how it is made available to users.

Consumer Services A number of standardized marketing information services directly involve consumers and their behavior. Some are concerned with purchase or consumption behavior, some with viewing and reading habits while still others are used for a variety of purposes

1.4.3 Types of Primary Data

Demographic/Socioeconomic Characteristics: One type of primary data of great interest to marketers is the subject's demographic and socioeconomic characteristics, such as age, education, occupation, marital status, sex, income, or social class. These variables are used to cross classify the collected data and in some way make sense of it. We might be interested, for instance, in determining whether people's attitudes toward ecology and pollution are related to their level of formal education. Alternatively, a common question asked by marketers is whether the consumption of a particular product is related in any way to a person's or family's age, education, income, and so on.

Attitudes/Opinions: Attitude is one of the more important notions in the marketing literature, since it is generally felt that attitudes are related to behavior. Obviously, when an individual likes a product he will be more inclined to buy it than when he does not like it; when he likes one brand more than another, he will tend to buy the preferred brand. Attitudes may be said to be the forerunners of behavior. Thus, marketers are often interested in people's attitudes toward the product itself, their overall attitudes with respect to specific brands, and their attitudes toward specific aspects or features possessed by several brands.

Awareness/Knowledge: Awareness/knowledge as used in marketing research refers to what respondents do and do not know about some object or phenomenon. For instance, a problem of considerable importance is the effectiveness of magazine ads. One measure of effectiveness is the product awareness generated by the advertisement. Awareness and knowledge are also used interchangeably when marketers speak of product awareness. Marketing researchers are often interested in determining whether the respondent is aware of the product, its features, where it is available.

Behavior: Behavior concerns what subjects have done or are doing. Most typically in marketing this means purchase and use behavior. It takes place under specific circumstances, at a particular time, and involves one or more actors or participants. The focus on behavior then involves a description of the activity with respect to the various components.

1.4.4 Basic Means of Obtaining Primary Data

The researcher attempting to collect primary data has a number of choices to make among the means that will be used. The primary decision is whether to employ communication or observation. Communication involves questioning respondents to secure the desired information, using a data collection instrument called a questionnaire. The questions may be oral or in writing, and the responses may also be given in either form. Observation does not involve questioning. Rather, it means that the situation of interest is checked and the relevant facts, actions, or behaviors recorded. Choosing a primary method of data collection implies a number of supplementary decisions. For example, should we administer questionnaires by mail, over the telephone, or in person? Should the purpose of the study be disguised or remain undisguised? Should the answers be open ended or should the respondent be asked to choose from a limited set of alternatives? A decision with respect to method of administration, say, has serious implications regarding the degree of structure that must be imposed on the questionnaire. Depending on structure and purpose of questionnaire following communication methods have been discussed.

Survey method involving a set of questions is a structured and undisguised form of data collection. Such method does not involve asking ambiguous questions as purpose is not to study hidden or latent behaviour. Questions pertaining to number of magazines being read in a month, amount of money spend on eating out, quality of services used are some of the examples that fall in this category of data collection. The method is used extensively as it is both time and cost effective. Surveys are widely used by market researchers to determine the preferences and attitudes of consumers. The results can be used for a variety of purposes from helping to determine the target market for an advertising campaign to modifying a candidates' platform in an election campaign. For example, a television network might conduct a survey to profile characteristics of owners of luxury automobiles including what they watch on television and at what times. This information can be very useful in formulating an advertisement theme for an automobile company. Some basic points to consider regarding questionnaire design are as follows:

- Questionnaire should be kept as short as possible to encourage respondents to complete it.
- Questions should be simple and clearly worded to enable respondents to answer quickly, correctly and without ambiguity.
- Open ended questions are useful in providing free expression but are time consuming and difficult to analyze.
- Leading questions should be avoided as they lead respondent to answer in a particular way.
- It is useful to pre-test a questionnaire.

Personal interviews allow the researcher to handle complex issues more effectively. Cooperation from respondents is more as talking to someone one-on-one helps in rapport and confidence building. The method helps in finding out additional details that might not emerge from initial responses. Issues which are sensitive such as incidents of drug abuse or issues relating to personality traits are better dealt by using personal interview method. Unfortunately, individual interviewing can be quite expensive and may be intimidating to some who are not comfortable sharing details with a researcher.

Focus Groups overcome the drawbacks associated with personal interview. Under this research format, a group of respondents (generally numbering 8-12) are guided through discussion by a moderator. The power of focus groups as a research tool rests with the environment created by the interaction of the participants. In well-run sessions, members of the group are stimulated to respond by the comments and the support of others in the group. In this way, the depth of information offered by a respondent may be much greater than that obtained through individual interviews. However, focus groups can be costly to conduct especially if participants must be paid. Also, a respondent may get influenced by opinions of others. In case of sensitive subjects participants may be hesitant to share with others.

Projective techniques are unstructured and disguised forms of collecting behavioural data. This method involves presentation of ambiguous and unstructured object or activity that a respondent is asked to respond. Word Association, Sentence completion and picture interpretation test are certain projective techniques to reveal hidden feelings and opinions with which respondents might be unaware of.

The communication method of data collection has the general advantages of versatility, speed, and cost, while observational data are typically more objective and accurate. *Versatility* is the ability of a technique to collect information on the many types of primary data of interest to marketers. A respondent's demographic and socioeconomic characteristics, the individual's attitudes and opinions, awareness and knowledge, intentions, the motivation underlying the individual's actions, and even the person's behavior may all be ascertained by the communication method. Observation is limited in scope to information about behavior and certain demographic and socioeconomic characteristics. But there are certain limitations to these observations. Observations are limited to present behavior, for example. A person's past behavior cannot be observed. Nor person's intentions as to future behavior can be observed.

The *speed and cost* advantages of the communication method are closely intertwined. Communication is a faster means of data collection than observation because it provides a greater degree of control over data gathering activities. The researcher is not forced to wait for events to occur with the communication method as she or he is with the observation method. In some cases, it is impossible to predict the occurrence of the event precisely enough to observe it. For still other behavior, the time interval can be substantial. For instance, an observer checking for brand purchased most frequently in one of several appliance categories might have to wait a long time to make any observations at all. Much of the time the observer would be idle.

Observation method might suffer from disadvantages of limited time, scope and cost but this method has advantage of being more objective and accurate. This is because the observational method is independent of the respondent's unwillingness or inability to provide the information desired. For example, respondents are often reluctant to cooperate whenever their replies would be embarrassing, humiliating, or would in some way place them in an unfavorable light. Observation typical produces more objective data than does communication. The interview represents a social interaction situation. Thus the replies of the person being questioned are conditioned by the individual's perceptions of the interviewer. The same is true of the interviewer, although the interviewer's selection and training affords the researcher a greater degree of control over these perceptions than those of the interviewee. With observation, though, the subject's perceptions play less of a role. Sometimes people are not even aware that they are being observed. This removes the opportunity for them to tell the interviewer what they think the interviewer wants to hear, or to give socially acceptable responses.

Exercise 2

True or false

1. Begin with primary data and check the collected data with secondary references.
2. Secondary data provides comparative data by which primary data can be more insightfully interpreted.
3. Observation method of collecting primary data is more cost effective than communication method.
4. Consumers' behavior towards fast food products can be better evaluated by using survey method than observation method.

1.5 Types of statistics

Broadly, with respect to application statistics can be categorized into two branches: descriptive and inferential statistics. When the objective is to study limited data and reach decision only about those respondents for whom data is collected then ***descriptive statistics*** shall be applied. Descriptive statistics is used to describe the characteristics or other variables under study of a group. For instance, if a dean of university wants to deduce that from the current batch of management students how many students who have performed fairly well in their entrance exams have also performed correspondingly in their first year of business course. So, a decision has to be

taken regarding a set of students by extracting quantitative information from data of same set of students with the use of descriptive statistics techniques.

Descriptive statistics involves arranging, summarizing and presenting a set of data in such a way that useful information is produced. Its method uses both graphical and numerical techniques (such as mean, standard deviation etc.) to present data on the basis of which informed decision can be made. The statistical methods used in descriptive statistics are quite elementary but important. For instance, for the application of advanced hypothesis testing tools such as t, F or other tests it is mandatory to check that selected data follows normal distribution. A normally distributed data follows certain characteristics which can be deduced by using elementary descriptive statistics techniques. If the selected data is not normally distributed and shows skewness then application of advanced statistical techniques is modified. Thus, descriptive statistics in addition to the role of describing the data also plays an important role of diagnosing the data.

In most of business problems a manager has to make a decision about large set of customers. For instance, to predict sales of air conditioners in coming summer season the manager might conduct a survey to observe the willingness and capability of acquiring new consumers. In such a scenario it is almost impossible to study the entire population in a specific geographic area. This entails taking a sample from the concerned population. This branch of statistics which gathers data from the sample and makes decision about the population from which the sample was taken by applying statistics is called ***inferential statistics***. The advantage of using inferential statistics is that it enables to take study wide range of phenomenon without conducting a census. For example, a bank manager wants to determine the average withdrawal from ATMs during week end. Because if withdrawals are more, then ATMs can go short of cash upsetting the customers. To study this bank manager might take a sample of the customers using those ATMs during week end and find out the probability of the population by studying the sample that customers will withdraw more than the average money.

Thus, the important and widely used concept of inferential statistics requires making inference about population parameters which involve uncertainty. To reduce uncertainty in the conclusion inferential statistics use the concept of probability. As discussed in above examples a sample was taken, the most important aspect in inferential statistics is the selection of sample which is true representative of the population under study. It derives from the discovery that much can be learnt from rather little data if it has been collected carefully. It is the statistics of small random samples drawn from well-defined populations, of survey and experimental design of formal inference. Adequate power and precision often did not need large samples. By enabling more care to be taken with collection and measurement a sample could be better than a full enumeration. Inferential statistics uses power of probability theory to bridge the gap between population and sample.

But this branch of statistics comes with its own failings. The trouble is that it is ill equipped to deal with many challenging and urgent problems as among the presence of huge amount of data concerning every problem it becomes difficult to judge which and how much data should be collected. For instance, the question, for example, of whether television and video violence is a cause of violence in society is amenable to statistical

treatment but the data fall far short of what is necessary and, in any case, a great variety of other economic, social and political issues must be taken into account. In such scenarios some of the argument will be recognizably statistical but much will be in the form of personal judgments, non- numerical information of uncertain quality, prejudice and self-interest.

Understanding the involvement of the concept of probability in making more precise decisions is illustrated by taking a imagined example. Suppose there is an infection among a population of animals. The incidence is partial. We wish experimentally to determine what that incidence is. How do we proceed? First, we may conceivably test all the animals. If we do this we come to a definite statement of the percentage affected (errors and omissions of observation accepted). The result is a particular fact, it is not a generalization. As we have used up, by hypothesis, our total universe of animals there is no need of generalization. But if the universe is large it would be impracticable to test all the animals. We must resort to sampling. Suppose we select at random a dozen animals and find eight infected and four free of infection. Shall we pass to the conclusion that two thirds of the animals are infected or one third are not infected? Obviously not. We could not expect all samples to run exactly alike. Suppose that we make the tentative assumption that the incidence is exactly two thirds in the total population. What would be the results of repeated selection and observation of random samples of twelve?

This type of statement of the inference assumes that the experience realized in the sample is not too anomalous. The inference might better be put thus: If the proportion of infection in the universe lie between .52 and .78 (this interval is calculated by using probability which is not shown here) then the experience 2/3=.667 in a random sample of twelve is not so unusual, or if the incidence of infection in the universe is less than 52 per cent. or more than 78 per cent, then realization of a random sample of twelve will be rarer than one chance in three. The enigmatical little word random deserves attention. We mean that the sample of twelve is merely picked out by chance.

Thus, inferential statistics is bereft with the problem of how to take a sample and is the selected sample is true representative of the population. This, randomness could be the source of inaccurate and unreliable results. So, statistical techniques used in inferential statistics despite of their widespread and effective usage should be clearly and thoroughly understood before their application.

Exercise 3

1. Summarizing of data involves.....statistics
2. Inferential statistics is based on mathematical concept of ……

1.6 Summary

Statistics is an important decision-making tool in business and is used in virtually all areas of business. The main objective of this field is to provide a decision maker relevant information from available huge amount of data. The understanding of type of data important in application of statistics. This chapter has discussed nominal,

ordinal, interval and ratio data by illustrating few examples. Also, two branches of statistics viz. descriptive and inferential statistics have been discussed. Both types of statistics should be applied before reaching a conclusion. Descriptive statistics only make decisions about the group for which data has been collected whereas inferential statistics helps to make decisions about larger set of population if sample of subjects belong to selected population.

1.7 Glossary

- **Statistics:** is an abstract science to extract information from data for effective decision making.
- **Nominal data:** is the lowest level of data used only to classify or categorize.
- **Ordinal data:** is used to rank or put data in either ascending or descending order.
- **Interval data:** is used to put meaning to equal intervals between data
- **Secondary data:** is already published data which was collected for a purpose, but it can be used for current research problem. The ways of collecting secondary data can be both internal and external.
- **Primary data:** is the data which needs to be collected regarding attitudes, behavior and intentions of customers or respondents regarding research problem on hand. The ways of collecting primary data involves communication and observation method.
- **Inferential statistics:** represents the branch of statistics where data is collected from sample to make a probable decision about population

1.8 Answers to check your progress/ Self assessment exercise

Exercise 1

(i) Ratio

(ii) Ratio

(iii) Ordinal

(iv) Interval

(v) Nominal

(vi) Ratio

Exercise 2

1. False
2. True
3. False
4. True

Exercise 3

1. Descriptive
2. Probability

1.9 References/ Suggested Readings:

- Black, K., *Business Statistics For Contemporary Decision Making,* Fifth Edition, Wiley India,
- Keller, G., *Statistics for Management,* First India Reprint 2009, Cengage Learning India Private Limited.
- Donald R. Cooper & Pamela S. Schindler, *Business Research Methods,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 9th Edition.
- S.P. Gupta, *Business Statistics*, Sultan Chand, New Delhi, 2006.

1.10 Terminal and Model Questions

1. Give an example of descriptive research in the college of your graduation. Give an example of inferential statistics from the same college. Compare the two examples. What makes them different?
2. What is the difference between primary and secondary data?
3. What are the advantages and disadvantages of secondary data?
4. What distinguishes internal secondary data from external secondary data?
5. What are the cautions that you should take before using secondary data?

Chapter 2: Analysing the Data

Structure

2.0 Objectives

- Understand the meaning, objectives and criteria of classification and tabulation.
- Describe various terms relating to frequency distribution.
- Distinguish between discrete and continuous frequency distributions.
- Construct a frequency distribution from a set of data.
- Construct different types of quantitative data graphs, including histograms, frequency polygons, ogives, in order to interpret the data being graphed.
- Construct different types of qualitative data graphs like pie charts and bar graphs in order to interpret the data being graphed.

2.1 Introduction

In every business a manager deals with various types of data. This raw data do not help in arriving at any decision. Thus, raw data is also known as ungrouped data. Statistics help in grouping the data in meaningful and presentable form. Data presented in the form of frequency distribution is known as grouped data. Grouped data is more meaningful, thus helps in interpretation and decision making. Classification and Tabulation of data help in organization of data. Classification is the first step in tabulation and prepares the ground for proper presentation of statistical data. After proper tabulation, statistical analysis and its interpretation is possible.

2.2 Classification of Data

Classification is the process of arranging data in groups or classes according to some common characteristics possessed by the facts constituting the data. In other words, it is the arrangement of data into sequences or groups according to their common characteristics, or separating them into different but related parts. For example, students of a class may be grouped together with respect to their marks obtained in an examination, their area of specialization or their age etc. Thus, by way of classification scattered heterogeneous data is converted into homogeneous groups.

2.2.1 Types of Classification

Data can be classified into the following types:

(i) Geographical Classification: When the classification of data is done on the basis of area or region such as cities, states etc. it is known as geographical classification. For example, production of rice in different states of the country.

(ii) Chronological Classification: When the data are classified on the basis of time, it is called chronological classification. Statistical data arranged chronologically are termed as *time series data.* For example, sales of a company for various years, exports of the company for various years, etc.

(iii) Qualitative Classification: Data classified according to some non-measurable characteristics such as, occupation, gender, honesty, literacy, etc. is known as qualitative classification.

(iv) Quantitative Classification: Data classified on the basis of some measurable characteristics such as, age, weight, income, sales, etc. is known as quantitative classification.

2.3 Tabulation of Data

After classification, tabulation is done to condense the data in a compact form which can be easily understood. Tabulation is defined as an orderly and systematic presentation of numerical data in rows and columns according to some characteristics. A statistical table is a classification of related numerical facts in vertical columns and horizontal rows. Tabulation helps in drawing useful interpretations about the data.

2.3.1 Frequency Distribution

A **frequency distribution** is a convenient way of presenting a large mass of data in tabular for by grouping the data. There are two types of frequency distributions, i.e. discrete and continuous.

2.3.1.1 Discrete Frequency Distribution (or ungrouped data)

A frequency distribution represented by a discrete variable is called a *discrete frequency distribution*. A discrete frequency distribution presents the data of a discrete variable in tabular form by listing all possible values that a discrete variable can take on, along with the corresponding frequencies.

If X I a discrete variable that can take on values $x_1, x_2, \ldots x_n$ with the corresponding frequencies $f_1, f_2, \ldots f_n$ then the frequency distribution of X is given by

X :	x_1	x_2	. . .	x_n
Frequency :	f_1	f_2	. . .	f_n

Constructing a Discrete Frequency Distribution

A discrete frequency distribution can be constructed by preparing a table having three columns. In the first column, all possible values of the variable are placed starting with the lowest and going up to the highest. In the second column, tally marks for each value are shown i.e. a mark is placed for each value,

and in order to facilitate counting, after drawing four marks fifth mark is put across them. The third column i.e. the frequency column shows the total number of tally marks corresponding to each value of the variable.

2.3.1.2 Continuous Frequency Distribution

A frequency distribution represented by a continuous variable is called a *continuous frequency distribution*. A continuous frequency distribution presents the data of a continuous variable in tabular form by grouping the data into different classes and then recording the number of items that appear in each class.

Constructing a Continuous Frequency Distribution

Following are the steps for constructing a continuous frequency distribution:

- **To determine the number of classes:** The decision on the number of class groupings depends largely on the judgment of the individual investigator and/or the range that will be used to group the data, although there are certain guidelines that can be used. As a general, a frequency distribution should have at least five class intervals, but not more than fifteen. In fact the number of classes depends on the number of data points and the range of the data collected. Some statisticians use a formula known as **'Sturges'** rule to determine the necessary number of classes:

  $k = 1+3.3 \log_{10} n$

  Where *n* is the number of items in the given data and *k* is the number of required classes.

  It is worth mentioning here that the classes must be selected in such a way so as to conform to following rules:

  I. All data items from the smallest to the largest must be included, and
  II. Each item must be assigned to one and only one class.

- **Determine the width of class intervals:** When constructing the frequency distribution it is desirable that the width of each class interval should be equal in size. The size (or width) of each class interval can be determined by first taking the difference between the largest and smallest numerical values in the data set and then dividing it by the number of class intervals desired.

  Width of class interval (*h*) = (Largest numerical value – Smallest numeric value)/Number of classes desired

  = Range / Number of classes desired

- **To determine the frequency of each class:** Finally, we count the number of observations falling in each class and record the appropriate number in the frequency column. It is advisable to use the method of *Tally Marks* to find the frequency of each class.

Types of classes

The classes can be formed in the following two ways for the grouping of the data:

(i) ***Inclusive Method*:** In this method grouping of the data is done in a way that both the lower and the upper class limits of each class interval are included in that class. The upper class limit of one class does not coincide with the lower class limit of the next class. For example, the scores obtained by the students in a class may be grouped as 0-9, 10-19, 20-29, . . . , where the class interval 0-9 includes all values from 0 to 9 (both inclusive).

Example: Form a frequency distribution from the following data by Inclusive method taking 4 as the magnitude of class intervals:

10 17 15 22 11 16 19 24 29 18 25 26 32 14 17 20 23 27 30 12

15 18 24 36 18 15 21 28 33 38 34 13 10 16 20 22 29 29 23 31

Solution: Since the minimum value of the variable is 10 we take the lower limit of the first class and the magnitude of the class intervals is given to be 4, the classes for preparing frequency distribution by the 'Inclusive Method' will be 10 – 13, 14 – 17, 18 – 21, 22 – 25, …, 34 – 37, 38 – 41, the last class being 38 – 41, because the maximum value in the distribution is 38. The final frequency distribution along with the tally marks is as under:

Class – Intervals	Tally bars	Frequency
10 – 13	𝍸	5
14 – 17	𝍸 III	8
18 -21	𝍸 III	8
22 – 25	𝍸 II	7
26 – 29	𝍸	5
30 – 33	IIII	4
34 – 37	II	2

38 – 41	I	1

(ii) ***Exclusive Method*:** When the class intervals are fixed in such a way that the upper limit of one class is the lower limit of the next class, it is the case of the exclusive method. The upper class limit of any class is excluded from that class. For example, the scores obtained by the students in a class are grouped from 0-10, 10-20, 20-30, . . . , then the class interval 0-10 includes all values which are greater than or equal to zero, but less than 10. Thus, a student scoring 10 marks will be included in the class 10-20.

Example: Prepare a suitable frequency distribution of the weights (in kgs) of 33 students of a class given below:

42, 74, 40, 60, 82, 115, 41, 61, 75, 83, 63, 53, 110, 76, 84, 50, 67, 65, 78, 77, 56,95, 68, 69, 104, 80, 79, 54, 73, 59, 81, 100.

Solution: For the present data, let us take the class – intervals of width 10 and put tally marks as follows:

Class – Interval	Tally Marks	Frequency
40 – 50	III	3
50 – 60	𝍸	5
60 – 70	𝍸 II	7
70 – 80	𝍸 III	8
80 – 90	𝍸	5
90 – 100	I	1
100 – 110	II	2
110 – 120	II	2
Total		33

2.3.1.3 Types of Frequency Distributions

1. **Cumulative Frequency Distribution:** Sometimes it is preferable to present data in a cumulative frequency distribution or simply a distribution which shows the cumulative number of observations below the upper boundary of each class in the given frequency distribution. A cumulative frequency distribution is of two types: (i) *more than* type, and (ii) *less than* type.

In the less than cumulative frequency distribution, the frequencies of each class interval are added successively from top to bottom and represent the cumulative number of observations less than or equal to the class frequency to which it relates. But in the more than cumulative frequency distribution, the frequencies of each class interval are added successively from bottom to top and represent the cumulative number of observations greater than or equal to the class frequency to which it relates.

2. **Percentage Frequency Distribution:** A percentage frequency distribution is one in which the number of observations for each class interval is converted into a percentage frequency by dividing it by the total number of observations in the entire distribution.

Example: The heights in cm of 30 persons are given below:

133 125 137 129 130 130 131 125 137 147 128 127 147 141 148 149 145 148 139 125 145 134 129 145 127 147 132 128 130 131

Prepare a frequency distribution for the above data.

Solution: We first decide on the number of classes into which data are to be grouped. Let us arbitrarily choose 9 classes. To determine the width of each class, we divide the range by the number of classes. The largest value is 149 and smallest is 125.

Therefore, range = largest value – smallest value = 149 – 125 = 24

Class width = range/no. of classes = 24/9 = 2.67 = 3 when rounded

Thus, the class width for each class is 3. Taking the first class interval as 125 – 127 and using tally marks, we get the required frequency distribution.

Frequency Distribution of Heights

Class Interval	Tally Marks	Frequency
125 – 127	𝍸	5
128 – 130	𝍸 II	7
131 – 133	IIII	4
134 – 136	I	1
137 – 139	III	3

140 – 142	I	1
143 – 145	III	3
146 – 148	IIII	4
149 – 151	II	2
		30

2.4 Graphical Presentation of Data

Most marketing managers prepare presentations by including a large number of suitable graphs. Graphical Presentation of data seems to be more appealing when we simply want to convey the trend of data. It is an effective visual impression and makes viewers aware of the important features of the data. Graphical presentation in statistics helps a researcher to understand the shape of the distribution. Diagrams and graphs are nothing but pictorial representation of statistical data. They are geometrical figures like points, bars, squares, rectangles, circles, cubes, pictures, maps or charts etc. Diagrams are useful because facts can be known at a glance. Following are some of the methods of graphical presentation of data:

2.4.1 Bar Chart /Bar Diagram

A bar chart is a graphical device used in depicting data that have been summarized as frequency, relative frequency, or percentage frequency. The class intervals are classified on the horizontal axis (X-axis). The frequencies are specified on the vertical axis (Y-axis).

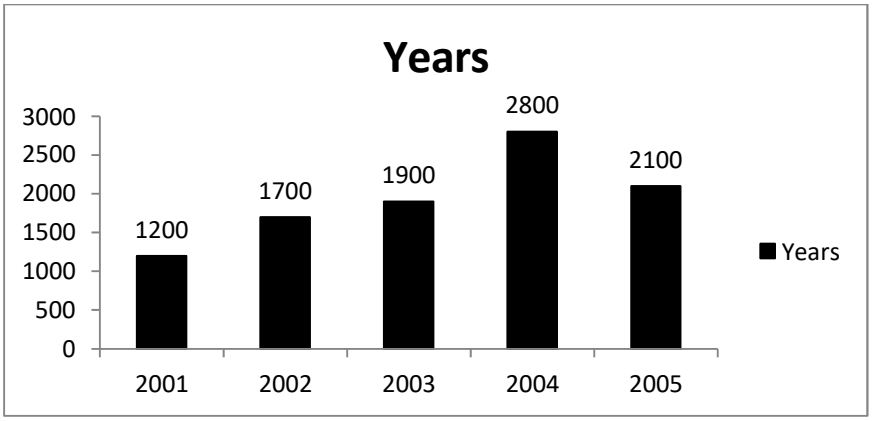
There are various kinds of bars.

- **Simple Bar Diagram**: In simple bar diagram, equal width is given to all the lines, but length of the bar (equal to frequency) differs. It is used to represent one variable.

Example: Draw a simple bar diagram to represent the following figures relating to manufacturing of number of products for various years.

Years	2001	2002	2003	2004	2005
No. of Products	1200	1700	1900	2800	2100

Solution: We can construct a bar diagram from the following data:



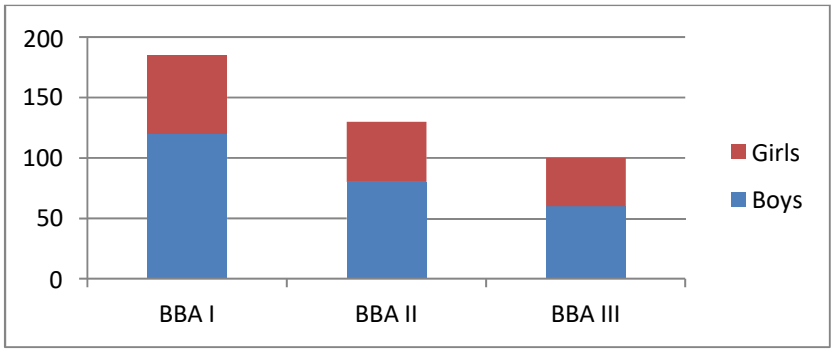
- **Sub-divided Bar Diagrams:** Sub-divided bar diagrams are useful for presenting several items of a variable and helps to make comparative study. To distinguish between different components a separate colour is used for each component**.**

  Example: The following table shows the No. of students (gender-wise) studying in AB College:

CLASS	BOYS (NO.)	GIRLS (NO.)	TOTAL (NO.)
B.A.I	120	65	185
B.A.II	80	50	130
B.A.III	60	40	100

  Represent the above data by a suitable diagram.

  Solution:



- **Percentage Bar Diagrams:** Sub-divided bar diagrams presented on percentage basis are known as percentage bar diagrams. The total of each bar is taken as 100 and the value of each component is expressed as percentage of the respective totals. In this method, all the

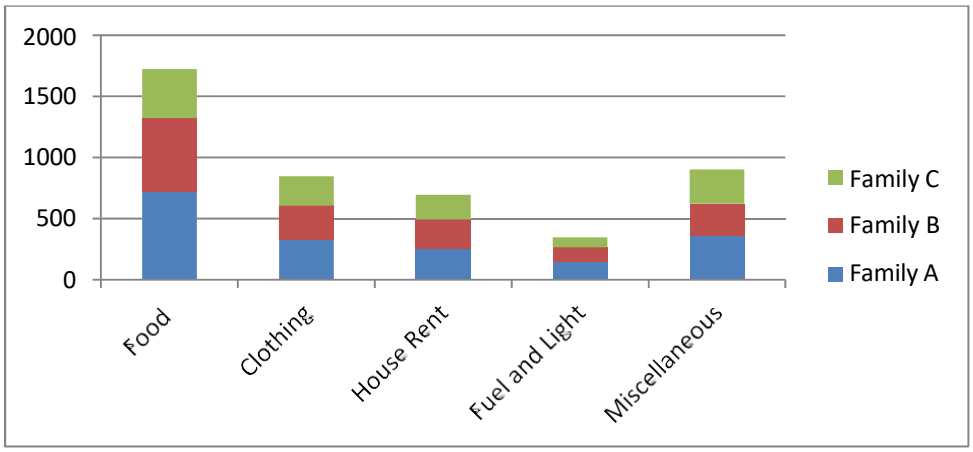
bars are of equal height and the various segments of the bar representing the different components will vary in height depending on their percentage value to the total. There bars are used to compare different data sets.

Example: The following table shows the monthly expenditure of three families:

Items	Family 1	Family 2	Family 3
Food	720	600	400
Clothing	324	280	240
House Rent	252	240	200
Fuel	144	120	80
Misc.	360	260	280

Represent the above data by a suitable diagram.

Solution:



- **Multiple Bar Diagrams:** Multiple bar diagrams are used when two or more sets of interrelated variables are to be presented graphically for comparison. A set of simple adjacent bars(one for each variable) is drawn. Different bars are distinguished by different colours.

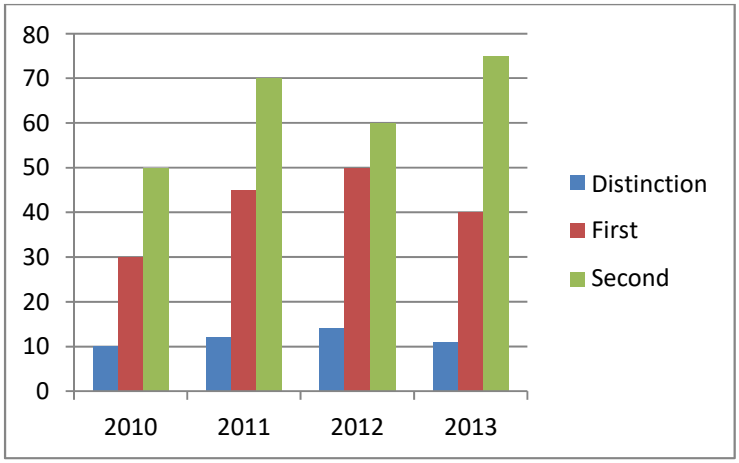
Example: The following table shows the result of M.B.A. students of a University for four consecutive years:

		No. of Students	
Year	**Distinction**	**First**	**Second**

2010	10	30	50
2011	12	45	70
2012	14	50	60
2013	11	40	75

Represent the above data by a suitable diagram.

Solution:



2.4.2 Pie Chart

A pie chart is a circular representation of data in which a circle is divided into sectors, with areas equal to corresponding component. These sectors are known as slices and represent the percentage breakdown of the corresponding component. The pie chart is the most common way of data presentation in today's business scenario. The construction of pie charts begins with determining the proportion of the component to the whole. The circle measures 360°each component proportion is multiplied by 360 to get the number of degrees to represent each component. Pie charts are useful in representing market share, budget categories and resource allocation, etc.

Example: The data shows market share (in percent) by revenue of the following companies in a particular year:

Companies	Market Share (%)
A	30
B	26

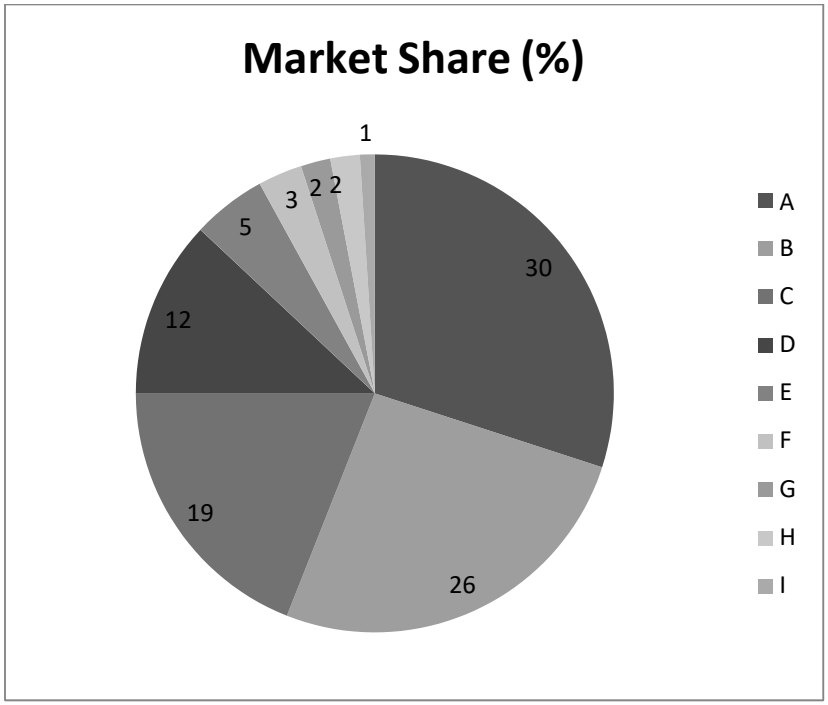
C	19
D	12
E	5
F	3
G	2
H	2
I	1

Draw a pie diagram for the above data.

Solution: Converting percentage figures into angle outlay by multiplying each of them by 3.6 as shown in the following table:

Company	Market Share (%)	Angle Outlay (Degree)
A	30	108.0
B	26	93.6
C	19	68.4
D	12	43.2
E	5	18.0
F	3	10.8
G	2	7.2
H	2	7.2
I	1	3.6
Total	**100**	**360.0**

Using the given data, construct the pie chart by dividing the circle into 9 parts (being 9 companies) according to the degrees of angle at the center.



2.4.3 Histogram

One of the most widely used types of graphs for quantitative data is a histogram. A histogram is a series of rectangles, each proportional in width to the range of the values with in the class and proportional in height to the class frequencies of the respective class interval. In order to plot histogram, the variable is displayed on the X-axis and the number of percentage of observations per class intervals is represented by Y-axis. In case the class intervals are equal, the heights of the rectangle represent the frequency of values in the given class interval. In case the class intervals are unequal, then the areas of the rectangles are used for relative comparisons.

Example: Draw a histogram to represent the following frequency distribution:

Marks:	0 – 10	10 – 20	20 – 40	40 – 50	50 – 60	60 – 70	70 – 90	90 – 100
No. of students:	4	6	14	16	14	10	16	5

Solution: Since the class intervals are of unequal width, therefore, the height of the rectangle for each class is proportional to the frequency density of that class. In fact, the height of the rectangle for each class corresponds to the product of the corresponding frequency density and the width of the class having the smallest size. That is,

Height of rectangle = frequency density * width of the smallest class

Computation of Frequency Density

Class Interval	Class Width	Frequency	Frequency Density	Height of the Rectangle
(1)	(2)	(3)	(4) = (3)/(2)	(4) * 10
0 – 10	10	4	0.4	4
10 – 20	10	6	0.6	6
20 – 40	20	14	0.7	7
40 – 50	10	16	1.6	16
50 – 60	10	14	1.4	14
60 – 70	10	10	1.0	10
70 – 90	20	16	0.8	8
90 - 100	10	5	0.5	5

The histogram representing the given frequency distribution is shown below:

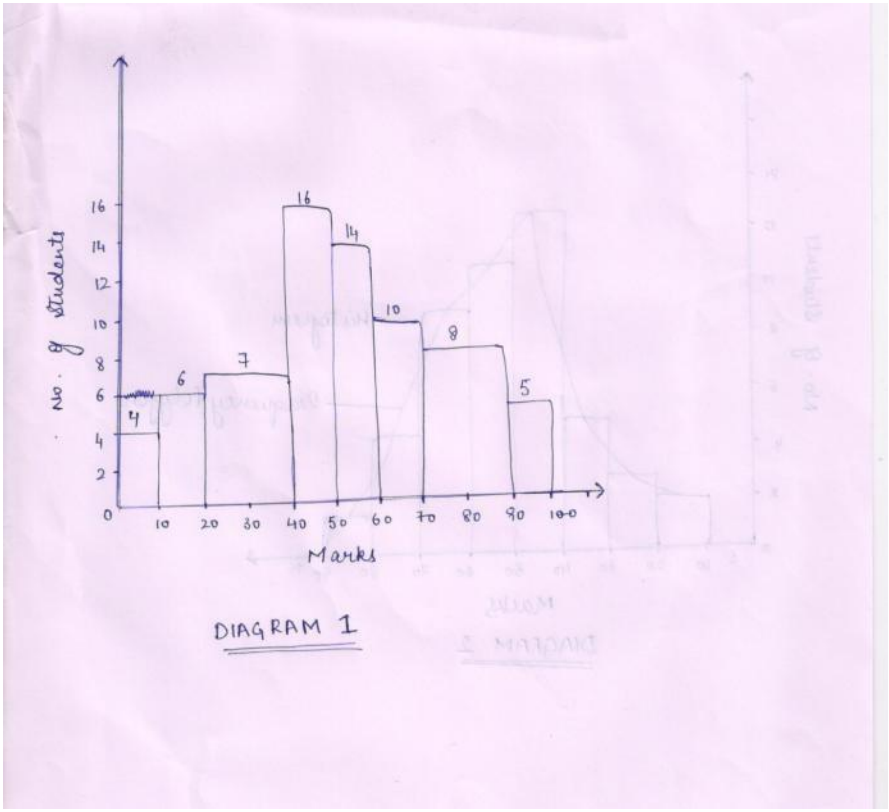


DIAGRAM 1

2.4.4 Frequency Polygon

A frequency polygon is a graphical display of class frequencies. However, instead of using rectangles, in a frequency polygon, each class frequency is plotted as a dot at the class mid-point and the dots are connected by a series of line segments. To construct a frequency polygon, frequencies are taken on Y-axis and the value of the variable (mid-point) is taken on the horizontal axis. A frequency polygon can also be constructed by connecting mid-points of individual bars.

Example: Draw a frequency polygon for the following distribution of marks obtained by 50 students in an examination:

Marks	**:** 10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70	70 – 80	80 – 90
No. of students:	2	3	7	13	11	9	4	1

To construct the frequency polygon, we mark the frequency (i.e., no. of students) on the vertical axis and the values of the variable (i.e., marks) on the horizontal axis. Then, we construct the histogram representing the frequency distribution and then join the mid points of the tops of the adjacent rectangles with the straight lines to form a polygon.

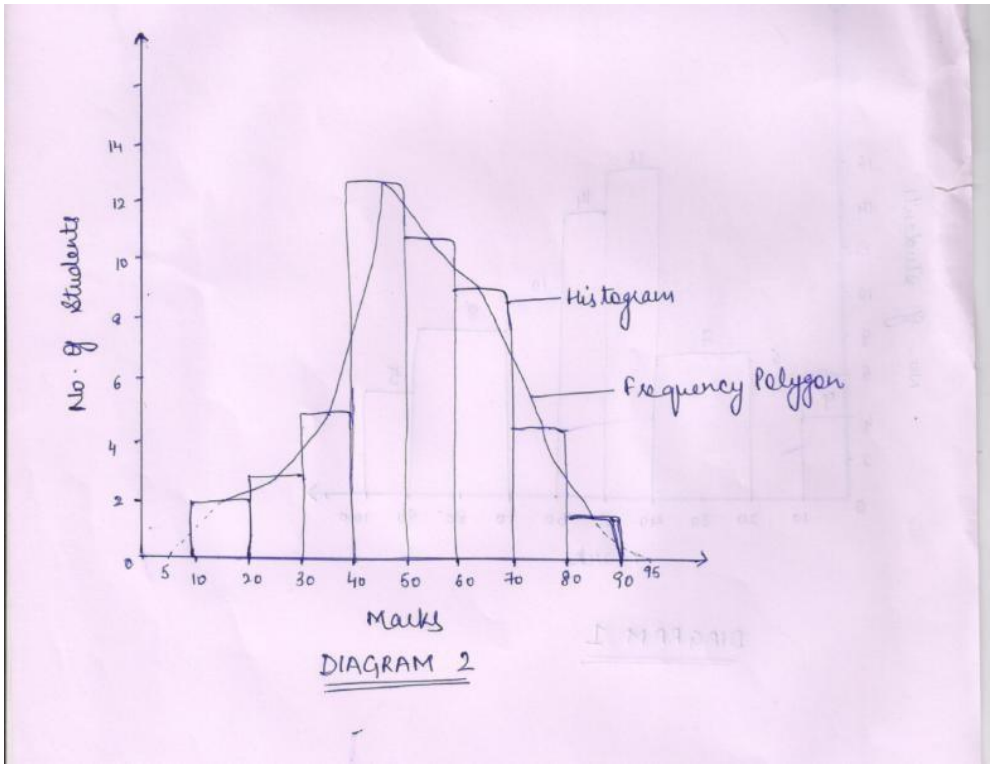


DIAGRAM 2

2.4.5 Ogive

An ogive (o-jive) is a cumulative frequency polygon. The data values are taken on horizontal axis and the cumulative frequencies on vertical axis. A dot of zero frequency is plotted at the beginning of the first class and the construction proceeds by marking a dot at the end of each class interval for the cumulative value. All the dots are connected to get an ogive.

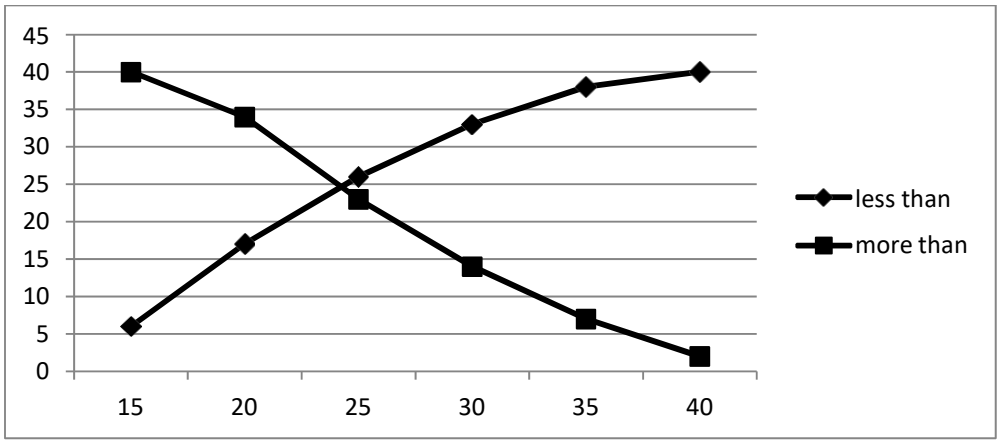
Example: Draw a 'less than' and 'more than' ogives from the following data:

Share Prices (Rs.)	**Number of Shares (f)**
10-15	6
15-20	11
20-25	9
25-30	7
30-35	5
35-40	2

Solution: First of all calculate the cumulative frequency:

Share Prices (Rs.)	**Upper Class Boundary**	**Number of Shares (f)**	**Cumulative Frequency**	
			Less than	**More than**
10-15	15	6	6	40
15-20	20	11	6+11=17	40-6=34
20-25	25	9	17+9=26	34-11=23
25-30	30	7	26+7=33	23-9=14
30-35	35	5	33+5=38	14-7=7
35-40	40	2	38+2=40	7-5=2

After calculating the cumulative frequency, draw the less than and more than ogive curves:



Self-Assessment Review:

1. The owner of a fast-food restaurant ascertains the ages of a sample of costumers. From these data, the owner constructs the frequency distribution as shown below. For each class interval of the frequency distribution, determine the class midpoint, the relative frequency, and the cumulative frequency and construct a frequency distribution table.

Class Interval	Frequency
0 - 5	6
5 - 10	8
10 - 15	17
15 - 20	23
20 - 25	18
25 - 30	10
30 - 35	4

2. Construct an Ogive for the following data:

Class Interval	Frequency
3 - 6	2
6 - 9	5
9 - 12	10
12 - 15	11
15 - 18	17
18 - 21	5

3. Assembly times for components must be understood in order to level the stages of a production process. Construct both histogram and a frequency polygon for the following assembly time data.

Class Interval	Frequency
30 - 32	5
32 - 34	7
34 - 36	15
36 - 38	21
38 - 40	34
40 - 42	24
42 - 44	17
44 - 46	8

4. Construct a pie chart from the following data:

Label	Value
A	55
B	121
C	83
D	46

5. Construct a bar graph from the following data.

Category	Frequency
A	7
B	12
C	14
D	5
E	19

2.5 Advantages of Diagrammatic Presentations

- Diagrams give a very clear picture of data. Data with the help of diagrams can be understood and grasped even by a layman in a very short time.
- This technique can be used universally at any place and at any time.
- Diagrams have impressive value also. A common man remains more interested in reading good diagrams than to go through the tabulated data.
- These give us much more information as compared to tabulation.

- Data can be condensed with diagrams. Small pictures sometimes represent the data which even thousands of words cannot express so clearly.
- Besides being informative, diagrammatic presentation is an entertaining means of data presentation.

2.6 Limitations of Diagrammatical Presentations

- Diagrammatic presentation of data is just an approximation of the actual behavior of the variables, i.e., diagrams show only an aggregated behavior of the variables.
- Only a limited set of data can be presented in the form of a diagram.
- Diagrammatic presentation of data is a time-consuming process.
- It is not very easy to arrive at final conclusions after seeing the diagrams. A diagram offers only preliminary conclusions.

2.7 Summary

A decision maker has to arrange a data in proper order in order to arrive at any conclusion. Data is of two types – ungrouped and grouped. Raw or scattered data is knows as ungrouped data while, data organized in the form of frequency distribution is called grouped data. Constructing a frequency distribution involves several steps. The first step is to determine the range of the data i.e. the difference between the largest and the smallest value. Then the number of classes is determined. Next step is to determine the width of the class by dividing the range of the values by the number of classes. In order to convey the trend of the data, graphical presentation is used. These are – bar charts, pie charts, histogram, frequency polygon and ogive etc. a bar chart is a graphical device for depicting data that have been summarized in frequency, whether relative or percentage. A pie chart is a circular depiction of data in which a circle is divided into sectors with areas equal to the corresponding component. A histogram can be defined as a set of rectangles, each proportional in width to the range of the values within a class and proportional in height to the class frequencies of the respective class interval. A frequency polygon is a graphical representation of the frequencies in which line segments connecting the dots depict the frequency distribution. An ogive is a cumulative frequency curve or a cumulative frequency polygon. Points on the ogive are plotted on the class end points.

2.8 Glossary

- **Frequency distribution:** A tabular summary of data showing the number (frequency) of observations in each of several non-overlapping class intervals.

- **Class mid-point:** The point in each class that is halfway between the lower and upper class limits.
- **Cumulative frequency distribution:** The cumulative number of observations less than or equal to upper class limit of each class.
- **Cumulative relative frequency distribution:** The cumulative number of observations less than or equal to upper class limit of each class.
- **Cumulative percentage frequency distribution:** The cumulative percentage of observations less than or equal to upper class limit of each class.
- **Bar graph:** A graphical device for depicting data that has been summarized in a frequency distribution, relative frequency distribution, or percent frequency distribution.
- **Class Interval:** An interval defining a class is called a class interval (or simply class).
- **Class limits:** The smallest and largest values that define a given class interval are referred to as its class limits. The smaller number is called the lower class limit and the larger number is called the upper class limit.
- **Class Boundaries:** The class boundaries are the values halfway between the upper class limit of one class and the lower class limit of the next class.
- **Class Width:** The difference between the upper and lower class boundaries of a class interval is called the class width.
- **Class Frequency:** The number of observations that fall in a particular class is called the class frequency and is denoted by the letter 'f'.

True or False:

1. Frequency distribution of continuous data can be presented graphically as histograms or frequency polygons.
2. A frequency polygon is appropriate for graphing continuously distributed variables.
3. Simple bar diagram is used only for one - dimensional comparisons.
4. Pie diagram is a circle divided into sections with areas equal to the corresponding component.
5. The height of a bar represents the frequency rather than the value of the variable.
6. The wider the class interval, the more specific information is lost about the actual data.
7. The frequency distribution represents data in a compressed form.
8. A frequency polygon can always be used to construct a histogram.

9. Diagrammatic presentation is a tabular form of classified data.
10. Only the length of the bar is considered in a bar diagram.
11. A multiple bar diagram is used when a comparison is made between values on two or more variables.
12. Deviation bar-diagrams are two-dimensional in nature.

2.9 Suggested Answers to SAQ's

[Answers: T, F, T, T, F, F, T, T, F, T, T, F]

2.10 Suggested Readings:

- Bajpai, N., Business Statistics, Pearson Education
- Srivastava, T.N. and Rego,S., Statistics for Management, Fourth Reprint, Tata McGraw Hill Companies.
- Thukral, J.K. Business Statistics, Second Edition, TAXMANN'S.
- Aggarwal, B.M., Business Statistics, Second Edition, Ane Books Pvt. Ltd.

2.11 Model Questions

1. Distinguish between continuous and discrete frequency distribution.
2. Distinguish between exclusive and inclusive class intervals.
3. Point out the role of diagrammatic presentation of data. Explain briefly the different types of bar diagrams.
4. Discuss the utility and limitations (if any) of diagrammatic presentation of statistical data.
5. Following are the number of items of similar type produced in a factory during the last 50 days.

   21 22 17 23 27 15 16 22 15 23 24 25 36 19 14 21 24 25
   14 18 20 31 22 19 18 20 21 20 36 18 21 20 31 22 19 18
   20 20 24 35 25 26 19 32 22 26 27 22

   Arrange these observations into a frequency distribution with both inclusive and exclusive class intervals choosing a suitable number of classes.
6. The distribution of ages of 500 readers of a nationally distributed magazine is given below:

Age (in years)	Number of Readers

Below 14	20
15 – 19	125
20 – 24	25
25 – 29	35
30 – 34	80
35 – 39	140
40 – 44	30
45 and above	45

Find the relative and cumulative frequency distributions for this distribution.

7. Make a diagrammatic representation of the following data:

Firms	Output in thousand tons
A	32
B	30
C	20
D	5
E	1
F	1

8. A shoe manufacturing company has collected data on its sales in different shoe size categories. The following table shows the sales of different shoe size categories. Construct a pie-chart using this data.

Category-wise sales of shoes by a shoe manufacturing company

Category Size	Sales ($ million)
3	110
4	120
5	115
6	95
7	155
8	140

9	80
Total	815

9. The following table lists the number of cars produced in a country from January 2013 to July 2013. On the basis of the data given in the table, construct a histogram.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul
Production (in numbers)	41999	41055	35942	37346	38003	39137	41151

10. Construct an Ogive from the data given in the table below.

Class Interval	**Frequency**
5 under 10	5
10 under 15	7
15 under 20	10
20 under 25	15
25 under 30	13
30 under 35	11

11. In a medium-sized city, 86 houses are for sale, each having about 2000 square feet of floor space. The prizes vary. The frequency distribution shown contains the prize categories for the 86 houses. Construct a histogram, a frequency polygon, and an ogive form these data.

Price	**Frequency**
$ 80,000 – under $ 100,000	21
100,000 – under 120,000	27
120,000 – under 140,000	18
140,000 – under 160,000	11
160,000 – under 180,000	6
180,000 – under 200,000	3

CHAPTER 3: MEASURES OF CENTRAL TENDENCY

Structure

3.0. Objectives

- To understand the meaning of the term central tendency.
- To learn various measures of central tendency like Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, and Mode for various types of data.

3.1. Introduction

An average is a single value within the range of the data that is used to represent all of the values in the series. Since an average is somewhere within the range of the data, it is known as the measure of central value. Thus, the tendency of the observations to concentrate around a central point is known as central tendency.

3.2. Measures of Central Tendency

A measure of central tendency is a typical value around which other figures congregate or which device their number in half. Measures of central tendency permit us to compare individual items in the group with it and also permit us to compare different series of figures with regard to their central tendencies. Measures used to locate the position of central value of the entire data are called the measures of central tendency. They have been classified as below.

3.2.1. The Arithmetic Mean

Arithmetic Mean (A.M.) is the most popular and commonly used measure of central tendency. It is the only measure in which all values play an equal role. The A.M. of a set of observations is computed by adding together all the values in the data set and then dividing that sum by the number of values, i.e. the observations in data set. It is denoted by $\bar{x}$ (x bar) or A.M.

So, A.M. = (Sum of all observations)/(No. of observations)

Arithmetic Mean is of two types:

- Simple Arithmetic Mean.
- Combined Arithmetic Mean
- Weighted Arithmetic Mean.

3.2.1.1 Simple Arithmetic Mean: Simple arithmetic mean assumes all the values of the seriesto the values of the series to be of same weightage. It can be calculated for different types of series.

Individual Observations

In Individual observations, the A.M. is calculated with the help of the following formula:

$$\text{Arithmetic Mean} = \frac{Total\ of\ all\ observations}{No.of\ observations}$$

If a sample contains n observations $x_1, x_2, x_3, \ldots, x_n$.

The Arithmetic Mean is calculated as follows:-

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N} = \frac{1}{N} \sum_{i=1}^{n} x_i$$

Example: The following data relates to walk time in minutes of an individual for ten days. What is the average time he walks daily?

Day	1	2	3	4	5	6	7	8	9	10
Time (Minutes)	40	30	42	53	40	45	38	30	44	35

By using the formula;

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N} = \frac{1}{N} \sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{40 + 30 + 42 + 53 + 40 + 45 + 38 + 30 + 44 + 35}{10} = 397/10$$

$\bar{x}$ = 39.7

Discrete Frequency Distribution

Arithmetic Mean of discrete frequency distribution is calculated by multiplying each term by its corresponding frequency and then dividing the sum of these products by the sum of the frequencies.

For a given series x_1, x_2, x_3….x_n with corresponding frequencies f_1, f2, f_3, . . ., f_n the A.M. is calculated as below:

$$A.M. = \frac{f_1 x_1 + f_2 x_2 + f_3 x_3 + \ldots + f_n x_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

Example: From the following data, calculate A.M.:

Wages (INR) (x)	110	112	113	117	120	125	128	130

No. of Workers (f)	25	17	13	15	14	8	6	2

Given, Mean wage = 115.86

Solution:

Wages (x)	No. of Workers (f)	fx
110	25	2750
112	17	1904
113	13	1469
117	15	1755
120	14	1680
125	8	1000
128	6	768
130	2	260
	N=100	11586

Using the formula

$$A.M. = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_nx_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

A.M. = 11586/100 =115.86

Continuous Frequency Distribution

In continuous frequency distribution, class intervals are used. Midpoint of each class interval is taken as value x and the remaining procedure is same as of discrete distribution.

$$A.M. = (\sum_{i=1}^{n} f_i m_i) / (\sum_{i=1}^{n} f_i)$$

Where, m_i = mid value of i^{th} class interval

f_i = frequency of i^{th} class interval

n = sum of all frequencies

Continuous series (*Short cut method*)

Sometimes it may be little difficult deal with decimals. Hence short cut may be useful for manual calculations.

$$\bar{x} = A + \frac{\Sigma f_i d_i'}{n} * h$$

Where

A = Assumed value of A.M.

n = sum of all frequencies

h = width of the class intervals

m_i = mid value of i^{th} class interval

d_i' = $\frac{M_i - A}{h}$ deviation from the assumed mean

Mid value can be calculated as an average of highest and lowest value of respective class intervals. i.e. (H+L)/2

Example:

City traffic police is planning to improve the safety of the general public. For this road accident data for the last 50 weeks was compiled and grouped into the frequency distribution as given in the following table. Calculate the average number of road accidents per week.

No. of road accidents	0-4	5-9	10-14	15-19	20-24
No. of weeks	5	22	13	8	2

Solution:

No. of road accidents	Mid-value (m_i) = (H+L)/2	No. of weeks (f_i)	$f_i m_i$
0-4	(0+4)/2 = 2	5	10
5-9	(5+9)/2 = 7	22	154

10-14	(10+14)/2 = 12	13	156
15-19	(15+19)/2 = 7	8	136
20-24	(20+24)/2 = 22	2	44
Total		50	500

A.M. of the no. of road accidents per week:

$$A.M. = (\sum_{i=1}^{n} f_i m_i) / (\sum_{i=1}^{n} f_i)$$

$$= 500/50 = 10$$

Properties of Arithmetic Mean:

- The algebric sum of the deviations of the given set of observations taken from the arithmetic mean is zero.
- The sum of the squares of deviations of a given set of observations is minimum when taken from arithmetic mean.
- If each observation of a data is increased or decreased by a constant k, then the arithmetic mean of the new data also gets increased or decreased by k.
- If each observation of a data is multiplied or divided by a constant k, then the arithmetic mean of the data so obtained also gets multiplied or divided by k.

3.2.1.2 Combined Arithmetic Mean: If the arithmetic means of two or more sets of data are known, then the arithmetic mean of the combined data can be calculated. If $n_1, n_2, \ldots, n_k$ are the number of observations with $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ as respective means of the data sets then the arithmetic mean of the combined series can be calculated as follows:

$$\bar{X}_{12} = (n_1 \bar{X}_1 + n_2 \bar{X}_2) / (n_1 + n_2)$$

Example:

The average turnover of 200 small and medium enterprises (SME_S) in textile industry financed by 'X' bank in a state is Rs 50 crores, and the average turnover of 300 small and medium enterprises(SME_S) in textile industry financed by 'Y' bank in the same state is Rs 60 crores. Find

the combined average turnover of the small and medium enterprises (SME_S) financed by both the banks.

Solution:

Combined Mean of 500 SME_s financed = Total turnover of all SME_S / Number of SME_s

= (Turnover of 200 SME_s financed by 'X' Bank +

Turnover of 300 SME_s financed by 'Y' Bank) / 500

= [(200*50) + (300*60)] / 500

= (10,000 + 18,000) / 500

= 28,000 / 500

= 56.

Thus, the combined average turnover of 500 SME_s financed by both the banks is 56 crores.

Merits and Demerits of A.M.:

Merits

- It is originally defined.
- It is easy to calculate and simple to understand
- It is based on all the observations
- It is suitable for further mathematical treatment.

Demerits

- It is very much affected by extreme values.
- It can not be computed for qualitative data like honesty, beauty, etc.

3.2.1.3 Weighted Arithmetic Mean

The above calculations of A.M. give equal importance to each observation in the data. However there are cases where all the items are not of equal importance. Thus in such cases, it becomes important to assign different weights to different items.

eg. $w_1, w_2, w_3, \ldots, w_n$

$$\bar{x}_w = \frac{\sum x_i w_i}{\sum w_i}$$ Where i=1,2,3, . . ., n.

Example: A candidate obtained the following percentage of marks in different subjects in the half yearly examination: English 46%, Statistics 67%, Cost Accountancy 72%, Economics 58%, and Income Tax 53%. It is agreed to give double weights to marks in English and statistics as compared to other subjects. Calculate the simple and weighted arithmetic mean.

Solution: Let marks as 'x' variable and 'w' as weights

Subject	Marks (%) (x)	Weight (w)	wx
English	46	2	92
Statistics	67	2	134
Cost accountancy	72	1	72
Economics	58	1	58
Income Tax	53	1	53
	∑x = 296	∑w = 7	∑wx = 409

Simple Arithmetic Mean = ∑x/n = 296/5 = 59.2%

Weighted Arithmetic Mean = ∑wX/∑w = 409/7 = 58.43%

Self-Assessment Exercise 1:

1. Compute the arithmetic mean of the following frequency distribution:

Marks:	20-30	30-40	40-50	50-60	60-70	70-80
No. of students:	5	11	18	22	16	8

2. The mean height of 25 male students in a class is 161cm and the mean height of 35 female students in the same class is 158 cm. Find the combined mean of 60 students in the class.

3. Suppose a person buys 5 kg. of vegetable 1 at the rate of Rs.20 per kg.,2 kg. of vegetable 2 at the rate of Rs.30 per kg. and 6 kg. of vegetable 3 at the rate 35 per kg. Find the average rate of purchase per kg.
4. Calculate the arithmetic mean from the following frequency distribution:

Marks:	0-10	10-20	20-30	30-40	40-50	50-60
No. of Pupils:	10	9	25	30	16	10

3.2.2. Geometric Mean

While calculating Arithmetic Mean, equal weight is attached to all the observations. Thus, A.M is not always a true representative of the observations e.g the Arithmetic Mean of 4 observations i.e 1, 2, 3 and 10 is 4 which is not a true representative of the data. In such cases the measure of location which can be used is the Geometric Mean.

The Geometric Mean (G.M.) of a series of observations $x_1, x_2, x_3, \ldots, x_n$ is defined as the n^{th} root of the product of these values e.g. if there are 3 items in a series i.e 5, 10, 20, the geometric mean would be the cube root of the product of all the 3 items.

G.M. = $(5*10*20)^{1/3}$

= $(1000)^{1/3} = 10$

Mathematically, G.M. = $\{x_1*x_2*x_3*\ldots*x_n\}^{1/n}$

It is important to mention that the G. M. cannot be defined if any value of x is zero or negative. G.M. can also be calculated with the help of logarithm values.

log G.M = $1/n\{\log x_1 + \log x_2 + \log x_3 + \ldots + \log x_n\}$

$$= 1/n \sum_{i=1}^{n} \log x_i$$

$$\text{G.M.} = \text{antilog}\{1/n \sum_{i=1}^{n} \log x_i\}$$

Example: The annual rate of growth for a factory for 5 years is 7%, 8%, 4%, 6% and 10% respectively. Find the average rate of growth per annum for this period.

Solution: If the growth in the beginning is 100, then for 5 years, the growth is given the following table.

Year	X	log x
1	100+7= 107	2.0293
2	100+8= 108	2.0334
3	100+4= 104	2.0170
4	100+6= 106	2.0253
5	100+10= 110	2.0413
Total		10.1463

$$G.M. = \text{antilog}\{1/n \sum_{i=1}^{n} \log x_i\}$$

$$= \text{antilog}\{10.1463/5\}$$

$$= \text{antilog}\{2.0293\}$$

$$= 106.98$$

The required average growth is 106.98 – 100 = 6.98. thus, the average rate of growth per annum is 6.98%.

Average rate of Growth:

Geometric Mean is commonly used for the calculation of the average rate of growth. The following formula is used for this purpose.

$$P_n = P_0\{1 + r\}^n$$

Where P_n is the amount at the end of the period n. P_0 is the amount at the beginning of the period, r is the average rate of change and n is the length of the period.

Example: If Rs. 1000 deposited at a certain rate of interest becomes Rs 2000 in 10 years, what is the compound annual rate of interest?

Solution: Applying the formula $P_n = P_0\{1 + r\}^n$

$2000= 1000(1 + r)^{10}$

$2= (1 + r)^{10}$

i.e. $(1 + r)^{10}=2$

Taking logrithms on both sides

$10 \log (1 + r) = \log 2 = 0.3010$

$\log (1 + r) = 0.3010/10 = 0.03010$

$1 + r = \text{antilog } (0.03010) = 1.0718$

Therefore, $r = 0.0718$

Thus, the rate of interest in percentage terms is $100*0.0718 = 7.18\%$.

Geometric Mean is useful for calculating average percentage increase or decrease. It is considered as the best average in construction of index numbers. Extreme values in the data have lesser impact on the value of this mean as compared to A.M. However, it is very difficult to manually calculate the value of G.M.

Self-Assessment Exercise 2:

1. The population of a country has increased from 100 million in 1975 to 144 million in 1985. Find the annual rate of growth of population.

2. Find the Geometric mean, if 4,5,6 and 2 occur with frequencies 1,2,3 and 4 respectively.

3.2.3. Harmonic Mean:

The Harmonic Mean is defined as the reciprocal of the Arithmetic Mean of the reciprocal of the observations. So firstly we calculate reciprocal of all values and calculate simple mean from these values. Thereafter reciprocal of this arithmetic mean becomes harmonic mean.

$1/H = \text{Arithmetic Mean of } 1/x_1, 1/x_2, \ldots, 1/x_n$

$1/H = 1/n\ (1/x_1 + 1/x_2 + \ldots + 1/x_n)$

$$1/H = 1/n\left(\sum_{i=1}^{n} 1/x_i\right)$$

Thus, $H = n / \sum_{i=1}^{n}(1/x_i)$

Harmonic Mean is used in averaging rates when the time factor is variable and the act being performed is same i.e average speed, average price, average profit etc. If the distance travelled from A to B and B to A is same but time taken to travel is different because of different speeds i.e 40 km/hr and 60 km/hr respectively. Harmonic Mean or average speed can be calculated as follows:

H.M = 2/(1/40+1/60) = 48 km/hr.

H.M. cannot be calculated if any value in the series is zero.

Relationship between A.M., G.M. and H.M.

- A.M ≥ G.M. ≥ H.M
- G.M. = $(A.M.* H.M.)^{1/2}$

Self-Assessment Exercise 3:

1. Find the harmonic mean for the following frequency distribution:

X:	22	30	40	50	60	70
f:	8	12	20	10	6	4

2. A man travelled by car for 3 days. He covered 480 kms each day. On the first day he drove for 10 hours at 48 kms an hour, on the second day he drove for 12 hours at 40 kms an hour and on the last day he drove for 15 hours at 32 kms per hour. What was his average speed?

3.2.4. Median:

Median is one of the important measures of central tendency. It is known as positional average because it is based on position or location of center when the series is in ascending or descending order. The median is that value of the variable which divides the group into two equal parts. One group shows values greater than and the other lesser than that value. The median is not affected by extreme values, so it is an important measure to use when extreme values are present in the data.

Calculation of Median:

Like Arithmetic mean,median can also be calculated for different types of series.

Individual Observations

Data is arranged either in ascending or descending order of magnitude. Middle value of this arranged data shows the median. If N is odd, then $\frac{(N+1)^{th}}{2}$ item is median value. If N is even, then average of $\frac{N^{th}}{2}$ and$[\frac{N}{2}+1]^{th}$ item gives the median value. For example, N = 8, the median would be the average of 4^{th} and 5^{th} item.

Discrete Series

In discrete series variables take values x_1, x_2, $x_{3, \ldots,}$ x_n with repetitive frequencies f_1, f_2, $f_{3, \ldots,}$ f_n with $\sum f$ = N, Median is the size of $\left(\frac{N+1}{2}\right)^{th}$ item.

Steps

- Calculate cummulative frequency (c.f.) distribution.
- Find $\frac{N+1}{2}$
- Check the c.*f* just greater than or equal to $\frac{N+1}{2}$
- The value of the variable corresponding to the c.*f*. obtained in the previous step is median value.

Example : Find the median from the following data:

Wages (Rs.)(x)	100	150	80	200	250	180
No. of wage earners (f)	24	26	16	20	6	30

Solution: Arranging the data in ascending order, we have;

Wages (Rs.)(x)	No. of Wage earners (f)	Cumulative Frequency
80	16	16

100	24	40
150	26	66
180	30	96
200	20	116
250	6	122
Total	N = 122	

Since N (=122) is even and N/2=61

C.F. ≥ 61 is 66. Hence Median is value corresponding to selected C.F. which is 150.

Continuous Series

In this case, median is the size of observations, N = $\sum f$ is the total frequency.

Steps

- Prepare a cumulative frequency distribution.
- Find $\frac{\mathrm{N}}{2}$
- Check the c.f just greater than $\frac{\mathrm{N}}{2}$
- Then find the class corresponding to the c.f obtained. This becomes the median class.
- Apply the formula:

Median = $l + \frac{\frac{\mathrm{N}}{2} - c.f}{\mathrm{f}} * \mathrm{h}$

Here

l = lower limit of the median class

f = frequency of the median class

c.f. = cumulative frequency of the class preceding the median class.

h = size or width of median class

Example: Star Ltd., employed 159 employees for a factory located at Mumbai. The company's management is worried about the high absenteeism rate in the organization. Before taking any corrective action, the management has decided to calculate the median leaves availed by the employees. The following table shows vacations availed in a year and the number of employees who availed vacations. Compute the median from the data.

Vacations availed in a year	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Number of employees	2	18	30	45	35	20	6	3

Solution:

Vacations availed	Number of employees(f)	Cumulative frequency(c.f.)
0-10	2	2
10-20	18	20
20-30	30	50
30-40	45	95
40-50	35	130
50-60	20	150
60-70	6	156
70-80	3	159

Total of the observations i.e., N=159

N/2 = 159/2 = 79.5, which falls in the class 30-40.

Hence the median class is 30-40,

l = lower limit of median class = 30

f = frequency of median class = 45

c. f. = total of all frequencies preceding median class = 50

h = width of class interval of median class = 10

Median = $l + \frac{\frac{\mathrm{N}}{2} - c.f}{\mathrm{f}} * \mathrm{h}$

$= 30 + \frac{\frac{159}{2} - 50}{45} * 10$

$= 30 + \frac{79.5 - 50}{45} * 10$

$= 30 + \frac{29.5}{45} * 10$

$= 30 + \frac{295}{45} = 36.55$

Merits and Demerits (Median):

Merits

- It is rigidly defined.
- It is easy to calculate and simple to understand.
- It is most appropriate average to be used while dealing with qualitative data.

Demerits

- It is a positional average so it is not based on each and every item of the distribution.
- In comparison to arithmetic mean, it is much affected by sampling fluctuations.
- It is necessary to average data in order of the magnitude.

Self-assessment Exercise 4:

1. The marks obtained by 200 students in a certain examination are given below:

Marks	0-25	25-50	50-75	75-100

No. of students:	30	50	80	40

2. Calculate the median of the following data:

X	10	20	30	40	50	60	70
f	1	5	12	20	19	9	4

3.2.5 Mode

Mode is the most common value. It is the value having the maximum frequency in the series. There are various situations in which A.M. does not always provide accurate characteristics of the data due to the presence of extreme values e.g. Most preferred brand of product. Here average represents majority.

Calculation of Mode

Individuals Observations

In case of individual observations, the value occurring most frequently is the mode.

Example: Given below is the data of 20 initial public offerings by Indian companies during July 2013. Find out the price mostly offered by the companies during this month.

14.25	19.00	11.00	28.00
24.00	23.00	43.25	19.00
27.00	25.00	15.00	7.00
34.22	15.50	15.00	22.00
19.00	19.00	27.00	21.00

Solution: For the data in the above table, the mode is INR 19.00 because the offer price that recurred the most times (four) was INR 19.00. Organizing the data into an ordered array (an ordering of the numbers from smallest to largest) helps to locate the mode. The following is an ordered array of the values given in the problem:

7.00 11.00 14.25 15.00 15.00 15.50 19.00 19.00 19.00 19.00 21.00 22.00 23.00 24.00 25.00 27.00 27.00 28.00 34.22 43.25

This grouping makes it easier to see that 19.00 is the most frequently occurring number.

If series has one clear mode ,it is said to be unimodal. In the case of a tie for the most frequently occurring value, two modes are listed. Then the data are said to be **bimodal**. If a set of data is not exactly bimodal but contains two values that are more dominant than others, some researchers take the liberty of referring to the data set as binomial even without an exact tie for the mode. Data sets with more than two modes are referred to as **multimodal**.

Discrete Series

Mode can be associated just by inspection of data. The value having the maximum frequency represents the mode. However, this method is applicable only if the distribution is **unimodal.** While determining mode by inspection in the case of discrete frequency distribution, an error of judgement is possible when the difference between the greatest frequency and the frequency preceding it or succeeding it is very small and the values are heavily concentrated on either side. In such cases, it is desirable to locate the mode by 'Grouping Method'.

Grouping Method:

The method of grouping involves preparing a group table. A grouping table has six columns. In column (1), we write down the original frequencies. The greatest frequency in this column is put in a circle or marked by bold type. In column (2), frequencies are grouped in two's. In column (3), we leave the first frequency and then group the remaining frequencies in two's. In column (4), frequencies are grouped in three's. In column (5), we leave the first frequency and then group the remaining frequencies in three's. In column (6), we leave the first two frequencies and then group the remaining frequencies in three's. In each of the columns, the highest total is put in a circle or marked by bold type.

After completing the group table, we prepare an Analysis table. In the Analysis table, column numbers are put in the left hand side and the various probable values of the mode are put on the right hand side. The values against which frequencies are highest are entered by means of a bar in the relevant box corresponding to the values they represent. The value which is repeated the maximum number of times represents the mode.

Example: Calculate mode from the following data:

Weight in Kg.	56	58	59	60	61	62	63	64	66	68
No. of girls	3	7	6	9	20	22	24	5	3	1

Solution: By inspection one is likely to say that the mode is 63 because it occurs the maximum number of times, i.e., 24.

Grouping Table

Height in inches	Col.1	Col.2	Col.3	Col.4	Col.5	Col.6
56	3					
		10				
58	7			16		
			13			
59	6				22	
		15				
60	9					35
			29			
61	20			51		
		42				
62	22				66	
			46			
63	24					
		29		32		
64	5					51
			8			
66	3				9	
		4				
68	1					

Analysis Table

Col. No.	56	58	59	60	61	62	63	64	66	68
1							1			
2					1	1				
3						1	1			
4				1	1	1				

5 6					1	1 1	1 1	1		
Total				1	3	5	4	1		

Since the value 62 has occurred the maximum number of times, i.e., 5, the mode is 62.

Continuous Series

In continuous frequency distribution, frequencies are given for various classes and the class having maximum frequency is called the modal class. Thus, first step is to ascertain the modal class. This can be done either by inspection or by preparing the grouping table and analysis table.

$$\text{Mode} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} * h$$

Here l = lower limit of modal interval

f_m = frequency of the modal class

f_{m-1} = frequency of the class preceding the mode class interval.

f_{m+1} = frequency of class following the mode class interval.

h = Width of the mode class interval.

Example: Calculate mode from the following data relating to production of a factory of 60 days.

Production (in Tons per day)	21-22	23-24	25-26	27-28	29-30
Number of Days	7	13	22	10	8

Solution:

Class Interval	Frequency (f_i)	Midpoint of class interval (m_i)	$f_i m_i$	Cumulative Frequency
20.5-22.5	7	21.5	150.5	7
22.5-24.5	13	23.5	305.5	20
24.5-26.5	22	25.5	561.0	42

26.5-28.5	10	27.5	275.0	52
28.5-30.5	8	29.5	236.0	60
Total	60		1528	

Using the formula:

$$\text{Mode} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} * h$$

Here Modal class is 24.5-26.5

Thus, l = 24.5

f_m = 22

f_{m-1} = 13

f_{m+1} = 10

h = 2

Therefore,

$$\text{Mode} = 24.5 + \frac{22 - 13}{2 * 22 - 13 - 10} * 2$$

$$\text{Mode} = 24.5 + \frac{9}{21} * 2$$

$$\text{Mode} = 24.5 + 0.86$$

$$\text{Mode} = 25.36$$

Merits and Demerits (Mode):

Merits

- It is simple to understand and easy to calculate.
- It is not at all affected by extreme observations and can be calculated even if extreme values are not known.

Demerits

- It is not rigidly defined.
- It is not based on all the observations.
- As compared to mean, mode is affected to a greater extent by the fluctuation of sampling.

3.3. Empirical Relationship Mean, Median and Mode

For a symmetrical Distribution

Mean = Median = Mode

$$\overline{X} = M_d = M_o$$

For a Moderately skewed distribution

3 (Mean-Median) = Mean - Mode

Mode = 3 Median - 2 Mean

$$M_0 = M_d - 2\overline{X}$$

Self-Review Exercise 5:

1. Compute the mode for the following distribution:

Class Interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	7	16	24	15	7

2. Find out the modal weight from the following data of the weights of 122 persons:

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

3. State True or False:

(i) Harmonic mean is the reciprocal of arithmetic mean.

(ii) Sum of absolute deviations from median is minimum.

(iii) Weighted mean is useful in problems relating to the construction of index numbers and standardized birth and death rates.

(iv) If the number of observations is even, the median is in the middle of the distribution.

(v)The mode is always found at the highest point of a graph of a frequency distribution.

(v) For grouped data, it is possible to calculate an approximate mean by assuming that each value in a given class is equal to its mid-point.

3.4. Summary

A measure of central tendency is a single value which represents an entire set of data. An average is known as a measure of central tendency as individual observations usually cluster around it. There are various types of measures of central tendency. Arithmetic mean, Geometric Mean and Harmonic Mean are the mathematical averages while the Median, Mode are positional averages. The Arithmetic mean of a set of observations is their sum divided by the number of observations. The weighted mean enables us to calculate an average value that takes into account the importance of each value with respect to the overall total. Geometric mean is the n^{th} root of the product of n items of a series. Geometric mean is useful in calculating the average percentage increase or decrease. Harmonic mean of any series is the reciprocal of the Arithmetic mean of the reciprocal of the variate. The Median of a distribution is the value of the variable which divides it into two equal parts. Mode is the value that is repeated most often in the data set.

3.5. Glossary:

- **Arithmetic Mean**: Sum of observations divided by number of observations
- **Geometric Mean**: The n^{th} root of the product of n observations.
- **Grouped Data:** Data available in class intervals as summarized by a frequency distribution.
- **Harmonic Mean:** The reciprocal of the arithmetic mean of the reciprocals of the observations.
- **Median:** The middle most value of a series of values arranged in ascending/ descending order.
- **Mode:** The value, which maximum number of observations has or tends to have as compared to any other value.

- **Weighted Mean:** The mean obtained by assigning each observation a weight that reflects its importance.

3.6 Answers to SAQ's:

Exercise 1:

1. 51.6 2. 159.25 cm 3. 31.3

Exercise 2:

1. 0.037 or 3.7% 2. 8.837

Exercise 3:

1. 36.21 2. 38.92 km/h

Exercise 4:

1. 40 2. 40

Exercise 5:

1. 27.76 2. 140.05
3. [(i) F, (ii) T, (iii) T, (iv) F, (v) T, (vi) T]

3.7. Suggested Readings:

- Black, K. Business Statistics for Contempory Decision Making, Fifth Edition, Wiley India.
- Srivastava, T.N. and Rego,S., Statistics for Management, Fourth Reprint, Tata McGraw Hill Companies.
- Thukral, J.K. Business Statistics, Second Edition, TAXMANN'
- Bajpai, N., Business Statistics, Pearson Education
- Aggarwal, B.M.., Business Statistics, Ane Books Pvt. Ltd.

3.8. Model Questions:

1. What do you mean by Arithmetic mean? Differentiate between Simple and Weighted arithmetic mean.
2. When is the usage of median considered more appropriate than mean?

3. Which measure of Central Tendency is usually preferred if the distribution is Considered to be single peaked and skewed? Why?
4. Distinguish between geometric and harmonic mean
5. On a certain day, the average closing force of a graph of stocks on the stock on the exchange (in '00 Rs.) is given below:

21 21 21 22 23 25 28 29 33 35 38 56 61

Calculate Mean, Median and Mode for the above data.

6. Find the missing frequencies from the following data if median is 33.5 and mode is 34.

X	0-10	10-20	20-30	30-40	40-50	50-60	60-70	Total
F	4	16	?	?	?	6	4	230

7. In a moderately skewed distribution, the mode is 20 and median is 24. Find the value of Mean.
8. The weighted geometric mean of the four numbers 20, 18, 12 and 4 is 11.75. If the weights of the first three numbers are 1,3, and 4 respectively. Find the weight of the fourth number.
9. Calculate the A.M., G.M., and H.M. of the following observations and show that A.M>G.M.>H.M.

32 35 36 37 39 41 43

10. The mean monthly salary paid to all employees in a company is Rs. 16,000. The mean monthly salaries paid to technical and non-technical employees are Rs. 18,000 and 12,000 respectively. Determine the percentage of technical and non-technical employees in the company. The unemployment rate in a city for the 12 months of 2013 is given in the table below:

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
8.7	8.8	8.7	7.8	7.3	7.8	6.6	6.5	6.5	6.8	7.3	7.6

(a) What is the arithmetic mean of the city's unemployment rate?

(b) Find the median and the mode for the unemployment rates.

(c) Compute the arithmetic mean and median for just the winter (Dec-Mar) months. Is it much different?

11. Determine the modal value in the following series:

Value	10	12	14	16	18	20	22	24	26	28	30	32
Frequency	7	15	21	38	34	34	11	19	10	38	5	2

Chapter 4. Measures of Variability and Measures of Shapes

Structure

4.7. Glossary

4.8 Answers to SAQ's

4.9 Suggested Readings

4.10. Model Questions

4.0. Objective:

- To know the importance of variability concept.
- To understand the concept of range, quartile deviation, mean deviation, standard deviation and variance.
- To understand various measures of shape.
- To understand the concept of skewness and understand the methods to measure skewness.
- To understand the meaning of Kurtosis and its measurement method.

4.1. Introduction

Measures of central Tendency help to know the average or middle values of the data. Measures of variability help to know the amount of variation among the values in the data set.

Thus, the degree to which numerical data tend to spread about an average value is known as variation or dispersion of the data.

4.2. Properties of a Good measure of Dispersion

- It should be simple to understand
- It should be rigidly defined
- It should be easy to calculate
- It should be based on all the observations

- It should not affected by fluctuations of sampling
- It should not be affected by extreme values

4.3 Methods of Measuring Dispersion

Following are the methods of measuring dispersion:

4.3.1 Range

The Range of a set of data is defined as the difference between the largest and smallest value in the set.

Range = L-S

Where L is the largest value and S is the smallest value. This is generally used in quality control charts. It is an absolute measure of dispersion so cannot be used for comparison purpose. Thus, for that purpose, a relative measure i.e. coefficient of Range is used.

$$\text{Coefficient of Range} = \frac{\text{Range}}{\text{Sum of largest and the lowest values}}$$

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

Example: Find out the range and coefficient in the following series.

x:	5-45	45-85	85-125	125-165	165-205
f:	3	5	6	3	2

Solution: The following table exhibits the class interval and frequencies.

x	F
5-45	3
45-85	5

85-125	6
125-165	3
165-205	2
∑f	19

As discussed above, the range can be determined by subtracting the lower limit of the lower class interval from the upper limit of the higher class interval

Upper limit of the highest class interval, $L = 205$

Lower limit of the lowest class interval, $S = 5$

Range, $(R) = L - S = 205 - 5 = 200$

Coefficient of Range $= \frac{L-S}{L+S} = \frac{200}{210} = 0.95$

$So\ Range = 200\ and\ coefficient\ of\ Range = 0.95$

4.3.2 Quartile Deviation or Semi Inter Quartile Range

It is an absolute measure of dispersion. It is the Range of values between the first and third quartile.

$$\text{Inter Quartile Range} = Q_3 - Q_1$$

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

It determines the average amount by which the two quartiles differ from the median.

$$\text{Coefficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It is a relative measure of dispersion.

Example: Find the quartile deviation of the daily wages (in Rs.) of 7 persons given below:

120 70 150 100 190 170 250

Solution: Arranging the data in ascending order of magnitude, we get

70 100 120 150 170 190 250

Here n = 7

Q1 = size of $\left(\frac{n+1}{4}\right)^{th}$ item = size of second item = 100

Q3 = size of $\frac{3(n+1)}{4}^{th}$ item = size of sixth item = 190

Hence Quartile Deviation = $\frac{Q_3 - Q_1}{2} = \frac{190-100}{2} = 45$

4.3.3 Mean or Average Deviation

Average absolute deviation is the average scatter of the items in a distribution, from either the mean or median or the mode, ignoring the signs of deviation.

Mean deviation in case of Individual observation

For a given set of an observations $x_1, x_2, \ldots, x_n$, the mean deviation about on average i.e. A is calculated as follows:

$$\text{Mean Deviation (about on average)} = \frac{\Sigma|x - A|}{n}$$

$$= \frac{\Sigma|D|}{n}$$

Average can be mean, median or mode as given in the data.

Mean deviation for discrete frequency distribution

In case of discrete series, where the variables x take the values x_1, $x_{2, \ldots,}$ x_n with respective frequencies f_1, f_2, f_n. The mean deviation about an average A is calculated as $\frac{\Sigma f \mid x - A \mid}{N} = \frac{\Sigma f \mid D \mid}{N}$

Where N = Σf

Example: Calculate mean deviation about the mean for the following data:

x:	10	11	12	13	14	Total
f:	3	12	18	12	3	48

Solution:

CALCULATIONS FOR MEAN DEVIATION

X	F	Fx	D = x- $\bar{x}$	|D|	f|D|
20	3	60	-2	2	6
21	12	252	-1	1	12
22	18	396	0	0	0
23	12	276	1	1	12
24	3	72	2	2	6
$N = \Sigma f = 48$		$\Sigma fx = 1056$			$\Sigma f \mid D \mid = 36$

Mean: $x = \frac{\Sigma fx}{\Sigma f} = \frac{1056}{48} = 22$

Mean Deviation about Mean = $\frac{\Sigma f \mid D \mid}{N} = \frac{36}{48} = 0.75$

Mean deviation for Continuous frequency Distribution

In this case mid values are used to take the deviation.

Coefficient of Mean Absolute Deviation = $\frac{Mean\,Deviation}{Mean} = \frac{\Sigma f\,|x-\bar{x}|}{\Sigma f}$

Example: Calculate the mean deviation from the mean for the following distribution:

Class	0-5	5-10	10-15	15-20	20-25
Frequencies	5	8	15	16	6

Solution

Weight	Mid-value (*x*)	$\frac{x-12.5}{5}=d'$	*f*	*fd'*	$\|x-\bar{x}=\|x-13.5\|$	$f\,\|x-\bar{x}\|$
0-5	2.5	-2	5	-10	11	55
5-10	7.5	-1	8	-8	6	48
10-15	12.5	0	15	0	1	15
15-20	17.5	1	16	16	4	64
20-25	22.5	2	6	12	9	54
			Σf=50	$\Sigma fd'$=10		$\Sigma f\,\|x-\bar{x}\|$ = 236

Mean = A + $\frac{\Sigma fd'}{\Sigma f} * i = 12.5 + \frac{10}{50} * 5 = 12.5.$

Mean deviation about the mean = $\frac{\Sigma f\,|x-\bar{x}|}{\Sigma f} = \frac{236}{50} = 4.72.$

4.3.4 Standard Deviation

The concept of standard deviation was firstly used by Karl Pearson. This is the most accepted, and widely used measure of computing dispersion.

Standard Deviation (S.D.) is the positive square root of sum of the square deviation of various values from their arithmetic mean divided by sample size.

In a set of n observations i.e.,x_1, x_2, $x_{3,\ldots,}$ x_n, S.D. is calculated as follows:

Individual Observation:

$$\sigma = \sqrt{\frac{\Sigma\left(x-\bar{x}\right)^2}{n}} \; where \; \bar{x} = \frac{\Sigma x}{n}$$

Discrete series:

$$\sigma = \sqrt{\frac{\Sigma f\left(x-\bar{x}\right)^2}{N}} \; Where \; N = \Sigma f$$

$$\bar{x} = \frac{\Sigma fx}{N}$$

Example: Calculate the standard deviation for the following data:

X:	20	30	40	50	60	70
f:	8	12	20	10	6	4

Calculation of Standard Deviation

X	*f*	*Fx*	$x-\bar{x}$	$\left(x-\bar{x}\right)^2$	$f\left(x-\bar{x}\right)^2$
20	8	160	-21	441	3528
30	12	360	-11	121	1452
40	20	800	-1	1	20
50	10	500	9	81	810
60	6	360	19	361	2166
70	4	280	29	841	3364
	N=60	$\Sigma fx = 2460$			$\Sigma f\left(x-\bar{x}\right)^2 = 11{,}340$

$$x = \frac{\Sigma fx}{N} = \frac{2460}{60} = 41$$

$$\sigma = \sqrt{\frac{\Sigma f\left(X - \bar{X}\right)^2}{N}} = \sqrt{\frac{11340}{60}} = \sqrt{189} = 13.75.$$

Continuous Series:

$$\sigma = \sqrt{\frac{\Sigma f d'^2}{N} - \left[\frac{\Sigma f d'}{N}\right]^2} \times h$$

Where $d = \frac{x - A}{h}$

And 'A' is Assumed Mean and 'h' is class internal.

Example: Calculate the arithmetic mean and standard deviation from the following series:

Class Interval	5-15	15-25	25-35	35-45	45-55
Frequency	8	12	15	9	6

Calculation of A.M. and S.D.

Class Interval	**Mid-value (x)**	**Frequency f**	$d' = \frac{x - A}{h}$ $(A = 30, h = 10)$	***fd'***	***fd'²***
5-15	10	8	-2	-16	32
15-25	20	12	-1	-12	12
25-35	30	15	0	0	0
35-45	40	9	1	9	9
45-55	50	6	2	12	24
N = 50			$\Sigma fd' = 7$		$\Sigma fd'^2 = 77$

Calculation of Arithmetic Mean.

$$\bar{x} = A + \frac{\Sigma fd}{N} * h = 20 + \frac{-7}{50} * 10 = 30 - 1.4 = 28.6$$

Calculation of Standard Deviation:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} * h = \sqrt{\frac{77}{50} - \left(\frac{-7}{50}\right)^2} * 10$$

$= \sqrt{1.54 - 0.0196} * 10 = 12.33.$

Variance:

Variance is the square of standard deviation;

In case of Individual Series: $\sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n}$

In case of Discrete and Continuousl Series: $\sigma^2 = \frac{\Sigma f(x - \bar{x})^2}{N}$

Coefficient of Variation:

It is the relative measure of Standard deviation. It is denoted as C.V.

$$\text{Coefficient of Variation} = \frac{\text{Standard deviation}}{\text{Mean}} * 100$$

$$= \frac{\sigma}{\bar{x}} * 100$$

Where $\bar{x}$ is the Arithmetic Mean and σ is the standard deviation. If coefficient of variation is low the data sets are considered as consistent and uniform while in case coefficient of variation is larger, distribution is considered to have more variation.

Example: Share prices of two companies A Ltd., and B. Ltd., were recorded as follows:

A Ltd.	12	13	15	14	14	14	13	17
B Ltd.	113	114	113	115	117	114	112	114

Which company's share prices are more variable?

Solution: Taking assumed mean A=14 for x series and B=114 for y series

Share Prices of A (x)	d_x	d_x^2	***Share Price of B (y)***	d_y	d_y^2
12	-2	4	113	-1	1
13	-1	1	114	0	0
15	1	1	113	-1	1
14	0	0	115	1	1
14	0	0	117	3	9
14	0	0	114	0	0
13	-1	1	112	-2	4
17	3	9	114	0	0
n=18	$\sum d_x = 0$	$\sum d^2_x = 16$	n=8	$\sum d_y = 0$	$\sum d^2_y = 16$

Company A Ltd:

$$\bar{x} = A + \frac{\Sigma d_x}{n} = 14 + \frac{0}{8} = 14$$

$$\sigma_x = \sqrt{\frac{\Sigma d_x^{\,2}}{n} - \left(\frac{\Sigma d_x}{n}\right)^2}$$

$$= \sqrt{\frac{16}{8}} = \sqrt{2} = 1.414$$

Company B Ltd.

$$\bar{y} = B + \frac{\Sigma d_y}{n} = 114 + \frac{0}{8} = 114$$

$$\sigma_Y = \sqrt{\frac{\Sigma d_y^{\,2}}{n} - \left(\frac{\Sigma d_y}{n}\right)^2}$$

$$= \sqrt{\frac{16}{8}} = \sqrt{2} = 1.414$$

C.V. (x) = $\frac{\sigma_x}{\bar{x}} * 100 = \frac{1.414}{14} * 100$

= 10.1%

C.V. (y) = $\frac{\sigma_Y}{\bar{y}} * 100 = \frac{1.414}{114} * 100$

= 1.24%

Since Coefficient of Variation (x) is more than Coefficient of Variation (y) so, the share prices of A Ltd. are more variable and the prices of B Ltd. are more consistent.

4.3.4.1 Mathematical Properties of Standard deviation

- S.D is independent of change of origin but not of scale.
- S.D. of a normal distribution is related to the area under normal curve as follows:

68.27% of the total area under the curve is covered from $\mu \pm \sigma$

95.45% of the total area under the curve is covered from $\mu \pm 2\sigma$

99.73% of the total area under the curve is covered from $\mu \pm 3\sigma$

- The sum of the squares of the deviation of a set of values is minimum when taken from the mean. That is why standard deviation is always calculated from the A.M.

4.3.5 Empirical relationship between measures of Dispersion

- $QD = \frac{2}{3}\sigma$
- $MAD = \frac{4}{5}\sigma$
- $QD = \frac{5}{6}MAD$
- $SD = \frac{5}{4}MAD = \frac{3}{2} or\ D.$

Or 4 SD =5 MAD =6 QD

4.3.6. Lorenz Curve

The Lorenz Curve is a graphic method of studying dispersion. It is a cumulative percentage curve in which the percentage of items(or frequencies) are shown with the corresponding percentage of factors like Income, Wages, Wealth, Profits, Turnover etc. The following are the steps to draw Lorenz curve.

- Find the cumulative values of the values of the items as well as the frequencies.
- Express the cumulated values as percentages by taking total of each as 100.
- Represent the percentages of cumulated frequencies on the X-axis and the percentages of cumulated values along the Y-axis. Scale of X- axis is generally taken from 100 to 0 and for Y- axis as 0 to 100.
- The percentage of cumulative values and the percentage of cumulative frequencies are plotted on the graph to get the Lorenz curve for the given data.

  If data is equally distributed, Lorenz curve is a a straight line which is obtained by joining '0' on X- axis with '100' on Y- axis. This line is known as 'Line of Equal Distribution'. The distance between the Line of Equal Distribution and Lorenz curve shows the extent of inequality. The *larger* the distance, the *greater* is the Variability. Lorenz curve has a limitation that it does not give a numerical value of the measure of distribution.

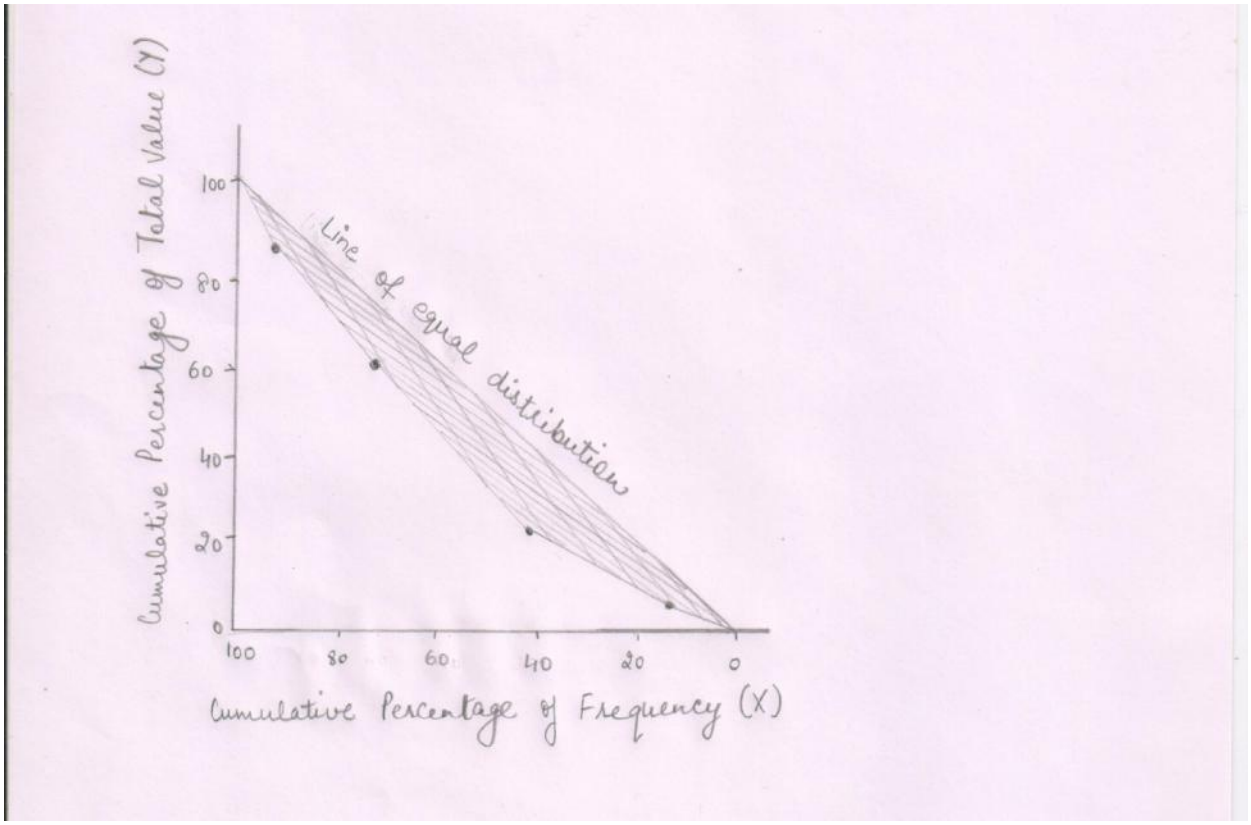
  Example: Draw the Lorenz Curve for the following data and show the inequality in income.

Income ('000 Rs.)	10-20	20-30	30-40	40-50	50-60
No. of employees	5	13	18	10	4

Solution:

Income ('000	Mid value	Frequency(f) No. of	Cum.Freq.	% Cum. Freq.(X)	Total value(m*f)	Cum.Total Values	% Cum. Total

Rs.)	(m)	employees					Values(Y)
10 - 20	15	5	5	10	75	75	44
20 – 30	25	13	18	36	325	400	23.5
30 – 40	35	18	36	72	630	1030	60.6
40 – 50	45	10	46	92	450	1480	87.1
50 - 60	55	4	50	100	220	1700	100.0



Self-Assessment Questions (Exercise 1):

(1) The weights of containers being shipped to Ireland are (in thousands of pounds):

95 103 105 110 104 105 112 90

(a) What is the range of the weights?

(b) Compute the arithmetic mean weight.

(c) Compute the mean deviation of the weights.

(2) Determine the mean and the standard deviation of the following frequency distribution.

Class	**Frequency**
20 – 30	7
30 – 40	12
40 – 50	21
50 – 60	18
60 – 70	12

(3) The number of employees, average wages per employee and variance of the wage per employee for two organisations are given below:

	Organisation 'X'	**Organisation 'Y'**
Number of Employees	100	200
Average Wage Per Employee (Rs.)	5000	8000
Variance of Wages per Employee	6000	10000

Determine the wages of which organisations are uniform?

(4) Draw Lorenz curves for the following;

Income(Rs.)	No. of employees(Firm A)	No. of employees(Firm B)

10000	60	150
12000	80	100
14000	120	90
16000	90	110
20000	100	30
22000	50	20

(5) State True or False:

1. The semi-interquartile range is inappropriate to use the skewed distribution.
2. Mean absolute deviation taken from median is least.
3. The standard deviation is measured in the same unit as the observations in the data set.
4. In a symmetrical distribution, semi-interquartile range is one-fourth of the range.
5. The coefficient of variation is a absolute measure of dispersion.
6. The inter-quartile range measures the average range of the lower fourth of a distribution.
7. For a symmetrical distribution, mean absolute deviation equals 4/5 of standard deviation.
8. Sample standard deviation provides an accurate estimate of the population standard deviation.
9. Variance is the square root of standard deviation.

10. Standard deviation can be calculated by taking deviation from any measure of central tendency.

4.4. Measures of Shape:

Measures of shape are the tools used for describing the shape of a distribution of the data.

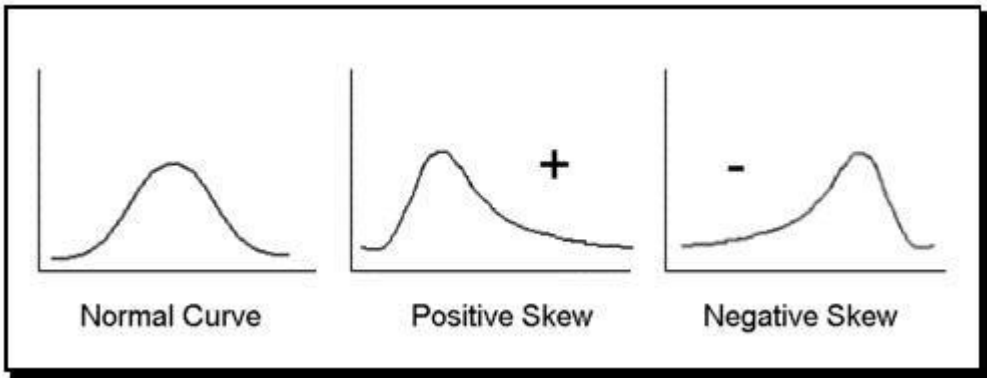
4.5. Methods to Measure Shape

There are two measures of shape.

(a) Skewness (b) Kurtosis

4.5.1 Skewness

It means lack of symmetry. If the frequency curve of a distribution is not symmetrical, the distribution is said to be skewed. The following figure shows the normal, positive and negatively skewed curves.



For a symmetrical distribution, the mean, the median and mode, all coincide.

$$\text{Mean} = \text{Median} = \text{Mode}$$

$$\bar{x} = M_d = M_o$$

4.5.1.1 Measures of Skewness

The degree of skewness in a distribution can be measured in absolute and relative manner.

Absolute measure of skewness

S_k = Mean – Mode

Or

S_k = Mean – Median

If mean is greater than mode or median, skewness is positive and if mean is less than mode of median skewness is negative.

If skewness is to be measured in quartiles.

$$S_k = Q_3 + Q_1 - 2 \text{ Median.}$$

Relative Measures of Skewness

These measures can be used to compare two distributions expressed in difference units. The following are the methods to measure skewness:

Karl Pearson's Coefficient of skewness

Median can also be used in place of mode as

Mean – Mode = 3 (Mean-Median)

$$S_{KP} = \frac{3\left(\bar{x} - M_d\right)}{\sigma}$$

$$S_{KP} = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}}$$

$$= \frac{\bar{x} - M_o}{\sigma}$$

Value of S_{Kp} varies between $\pm$ 3. But for a moderately skewed distribution. $S_{Kp} = \pm 1$

Example: From the following data on age of employees, calculate the coefficient of skewness and comment on the result.

Age below (years)	25	30	35	40	45	50	55
Number of employees	8	20	40	65	80	92	100

Solution: The data are given in a cumulative frequency distribution form. So, to calculate the coefficient of skewness, convert this data into a simple frequency distribution as shown in following table.

Table: Calculations for Coefficient of Skewness

Age (Years)	*Mid-value (m)*	*Number of Employees (f)*	*d' = (m-A)/h = (m-37.5)/h*	*fd'*	*fd'²*
20-25	22.5	8	-3	-24	72
25-30	27.5	12	-2	-24	48
30-35	32.5	20	-1	-20	20
35-40	37.5	25	0	0	0
40-45	42.5	1	1	15	15
45-50	47.5	12	2	24	48
50-55	52.5	8	3	24	72
		N = 100		-5	275

$$\text{Mean, } \bar{x} = A + \frac{\Sigma fd'}{N} * h = 37.5 - \frac{5}{100} * 5 = 37.25$$

Mode of value lies in the class interval 35-40. Thus

$$\text{Mo} = l + \frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} * h$$

$$= 35 + \frac{25 - 20}{2 * 25 - 20 - 15} * 5 = 35 + \frac{5}{15} * 5 = 36.67$$

Standard deviation, $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} * h$

$$= \sqrt{\frac{275}{100} - \left(\frac{-5}{100}\right)^2} * 5 = \sqrt{2.75 - 0.0025} * 5 = 8.29$$

Karl Pearson's coefficient of skewness:

$$S_{KP} = \frac{\text{Mean} - \text{Mode}}{\sigma} = (37.25 - 36.67) / 8.29 = 0.07$$

The positive value of S_{Kp} indicates that the distribution is slightly positively skewed.

Bowley's Coefficient of Skewness

This measure is based on the relative positions of the median and the quartiles in a distribution. If a distribution is symmetrical then Q_1 and Q_3 would be at equal distribution from the value of the median.

$$\text{Median} - Q_1 = Q_3 - \text{Median}$$

Or

$$\text{Median} = \frac{Q_3 + Q_1}{2}$$

When a distribution is asymmetrical, quartiles are not equal distance from the median.

If Q_1 – Median > Q_3 – Median, distribution is positively skewed:

$$\text{Relative } S_B = \frac{Q_3 + Q_1 - 2\,\text{Median}}{Q_3 - Q_1}$$

If $S_B > 0$, the distribution is positively skewed and if $S_B < 0$, the distribution is negatively skewed. Thus, for moderately skewed distribution, S_B lies in $\pm$ 1.

Example: Calculate Bowley's coefficient of skewness from the following data:

x:	10	5	7	11	8
f:	15	20	15	18	12

Solution: Firstly arrange data in ascending order.

X	***F***	***c.f.***
5	20	20
7	15	35
8	12	47
10	15	62
11	18	80

Computation of Q_1 . We have $\frac{N+1}{4}=\frac{81}{4}=20.25$; and the c.f. just greater than or equal to 20.25 is 35.

The corresponding value of x is 7.

$\therefore$ $Q_1 = 7$

Computation of Q_2: We have $\frac{2(N+1)}{2}=40.5$; and the c.f. just greater than or equal to 40.5 is 47. The

corresponding value of x is 8.

$\therefore$ $Q_2 = 8$

Computation of Q_3: We have $\frac{3(N+1)}{4}$ = 60.75; and the c.f. just greater than or equal to 60.75 is 62. The

corresponding value of x is 10.

$\therefore$ $Q_3 = 10$

$\therefore$ Bowley's coefficient of skewness is given by

$$S_B=\frac{Q_3+Q_1-2Q_2}{Q_3-Q_1}=\frac{10+7-2(8)}{10-7}=\frac{1}{3}=0.33.$$

Kelly's Coefficient of skewness

In this method, percentiles and deciles are used.

$$S_K = \frac{P_{10} + P_{90} - 2\text{Median}}{P_{90} - P_{10}}$$

Or

$$S_K = \frac{D_1 + D_9 - 2D_5}{D_9 - D_1}$$

If $S_{KB} > 0$, the distribution is positively skewed and If $S_K < 0$, distribution is negatively skewed.

Example: Calculate Kelly's coefficient of skewness from the following data:

Class	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	8	12	20	30	15	10	5

Solution:

Classs (x)	***Frequency (f)***	***Cumulative Frequency (c.f.)***
0-10	8	8
10-20	12	20
20-30	20	40
30-40	30	70
40-50	15	85
50-60	10	95
60-70	5	100
N = Σf = 100		

Computation of P_{10} : We have $\frac{10N}{100} = 10$. The cumulative frequency (c.f.), just greater than or equal to 10 is 20. Hence the corresponding class 10-20 contains P_{10} which is given by

$$P_{10} = l + \frac{\frac{10N}{100} - \text{c.f.}}{f} * h = 10 + \frac{10-8}{12} * 10 = 10 + 1.67 = 11.67$$

Computation of Median: We have $\frac{N}{2} = 50$. The c.f., just greater than or equal to 50 is 70. Hence the median lies in the class 30-40.

$$M_d = l + \frac{\frac{N}{2} - \text{c.f.}}{f} * h = 30 + \frac{50-40}{30} * 10 = 30 + 3.33 = 33.33$$

Computation of P_{90} : We have $\frac{90N}{100} = 90$. The c.f., just greater than or equal to 90 is 95. Hence the corresponding class 50-60 contains P_{90} which is given by

$$P_{90} = l + \frac{\frac{90N}{100} - \text{c.f.}}{f} * h = 50 + \frac{90-85}{10} * 10 = 50 + 5 = 55$$

Computation of Kelly's coefficient of skewness

$$S_K = \frac{P_{90} + P_{10} - 2\text{Median}}{P_{90} - P_{10}} = \frac{55 + 11.67 - 2(33.33)}{55 - 11.67}$$

$$= \frac{0.01}{43.33} = 0.0002 .$$

Example: Pearson's coefficient of skewness for a data distribution is 0.5 and coefficient of variation is 60%. Its mode is 100. Find the mean and median of the distribution

Solution:

Given; $S_{KP} = 0.5$, CV = 60%, and Mode = 100

Thus, $CV = \frac{\sigma}{\bar{x}} * 100$

$60 = \frac{\sigma}{\bar{x}} * 100$

$\sigma = 0.6\,\bar{x}$

Also S_{KP} = (Mean – Mode)/S.D.

$= 0.5 = (\bar{x} - 100)/0.6\,\bar{x}$

$0.5 * 0.6\,\bar{x} = \bar{x} - 100$

$\bar{x} = 142.86$

Now, Mode = 3Med – 2Mean

100 = 3Med – 2(142.86)

Med = 365.72/3 = 121.9

Self-Assessment Questions(Exercise 2):

1. The data of the marks obtained by 60students is as follows:

Marks :	>10	10-20	20-30	30-40	40-50	<50
No. of students:	5	12	20	16	5	2

   A. Obtained limits of marks of the central 50% students.

   B. Calculate Bowley's coefficient of skewness

2. In a frequency distribution, coefficient of skewness based on Quartiles is 0.6. If the sum of upper and lower Quartile is 100 and the median is 38, find the values of lower and upper Quartiles.

3. From the following data calculate Karl Pearson's coefficient of skewness:

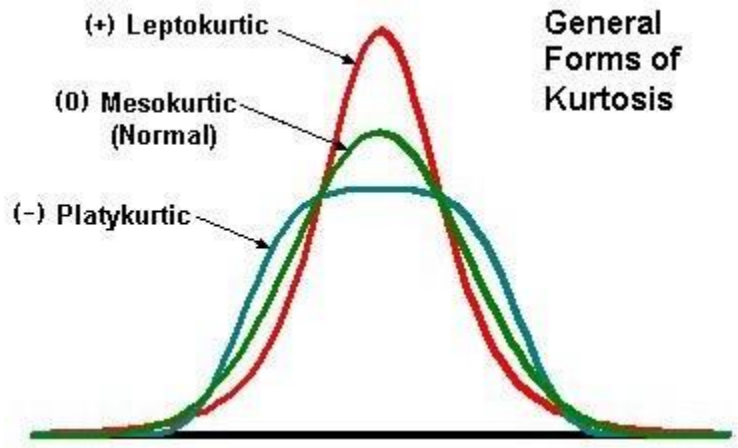
Marks (more than):	0	10	20	30	40	50	60	70	80

Number of students:	150	140	100	80	80	70	30	14	0

4. In a frequency distribution, the coefficient of skewness based on Quartiles is 0.6. If the sum of upper and lower Quartile is 100 and the median is 38, find the values of upper Quartiles.

4.5.2 Kurtosis

The word Kurtosis originated from a greek letter whose meaning is 'humped'. It is said that the two distributions with same mean, variance and skewness can be slightly different in shape. Kurtosis describes the degree of concentration of observations in a given distribution i.e. the degree of flatness or peakedness of the curve. A flatter distribution than normal distribution is known as 'Platokurtic'. A more peaked distribution than the normal distribution is known as 'Leptokurtic'. In between these two types of distribution, the distribution which is more normal in shape in known as 'Mesokurtic' distribution.



4.5.2.1 Moment Coefficient of Kurtosis

Karl Pearson gave a method known as Moment Coefficient of Kurtosis

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

Where μ_2 and μ_4 are second and fourth moments about the mean respectively.

β_2 measures the degree of peakedness. The greater the value of β_2, the more peaked is the distribution.

$\beta_2 = 3$ for the normal distribution.

$$\text{Kurtosis } (\gamma_2) = \beta_2 - 3$$

The value of Kurtosis (γ_2) is positive for a leptokurtic curve, negative for a platykurtic curve and Zero for a normal curve.

4.6. Summary:

Dispersion is 'scatteredness'. There are two types of measures of dispersion. Absolute measures of dispersion is used in case of the unit of distribution is given while the relative measure of dispersion is useful for comparing different sets of data with different units of measurement. Range is the simplest measure of dispersion and is calculated as the difference between the smallest and the greatest value of the items. Inter quartile range is the different between the third quartile and first quartile. Average deviation is the average amount of scatteredness of the items in a distribution from either the mean or median or mode ignoring the signs of deviations. Standard deviation is the square root of the sum of square deviation of various values from their arithmetic mean divided by sample size minus one variance is the square root of standard deviation. Standard deviation is an absolute measure of dispersion. Relative measure is coefficient of variation. Lorenz Curve is a graphical method of measuring Dispersion.

Shape of the distribution has a lot of importance in order to analyze data. There are two methods to measure shape; skewness and Kurtosis. Skewness shows the lack of symmetry. Skewness is measured with the help of various methods like Karl Pearson's method, Bowley's method and Kelley's method. Kurtosis describes the degrees of concentration of observations in a given distribution which is measured through flatness or peakedness of the curve. It is calculated through a method given by Karl Pearson and known as moment coefficient of Kurtosis.

4.7. Glossary

- **First Quartile (Q1):** The value which divides the data in two parts, 25% less than this value and 75% more than this value.
- **Third Quartile (Q3):** The value which divides the data in two parts, 75% less than this value and 25% more than this value.
- **Range:** The difference between the maximum and the minimum values of the observations.

- **Inter-Quartile Range (IQR):** The difference between the third Quartile (Q3) and the first Quartile (Q1).
- **Semi Inter-Quartile Range or Quartile Deviation:** The difference between the third Quartile (Q3) and the first Quartile (Q1) divided by 2.
- **Mean or Average Deviation:** Sum pf the absolute deviations of the values from their mean or median divided by the number of values.
- **Variance:** The sum of squares of deviations of observations from mean divided by the number of observations.
- **Standard Deviation:** The square root of variance.
- **Coefficient of Variation (C.V.):** The ratio of standard deviation to the mean, usually expressed in % form.
- **Lorenz Curve :** It is a graphical method of measuring Dispersion.
- **Skewness:** Indicator of lack of symmetry in data.
- **Kurtosis:** Indicator of the peakedness of a distribution.
- **Moments:** Represent a convenient and unifying method for summarizing certain descriptive statistical measures.
- **Leptokurtic:** A frequency curve that is more peaked than the normal curve.
- **Platykurtic:** A frequency curve that is flat-topped than the normal curve.
- **Mesokurtic:** A frequency curve that is normal (symmetrical) curve.

4.8 Suggested Answers for SAQ's:

Exercise 1

1. [Ans: (a) 22 thousand of pounds ; (b) 103 thousand of pounds ; (c) 5.25 thousand of pounds]
2. [Ans: mean=47.29 ; S.D.=12.18]
3. [Ans: org 'X' = 1.55% org 'Y' = 1.25% ; wages uniform in org 'Y']

5.[Ans: 1. T 2. T 3. T 4. T 5. F 6. F 7. T 8. F 9 F 10. F]

Exercise 2

1. [Ans: A, Marks of central 50% students lies between 10.83 and 35; B, 0.02]

2. [Ans: Q1 = 30; Q3 = 70]

3. [Ans:-0.754]

4. [Ans: 70]

4.9 Suggested Readings:

- Black, K. Business Statistics for Contempory Decision Making, Fifth Edition, Wiley India.
- Srivastava, T.N. and Rego,S., Statistics for Management, Fourth Reprint, Tata McGraw Hill Companies.
- Thukral, J.K. Business Statistics, Second Edition, TAXMANN'S.
- Bajpai, N., Business Statistics, Pearson Education.

4.10. Model Questions

Q1. What is the meaning of dispersion. Discuss various measures of dispersion.

Q2. What are the mathematical properties of standard deviation.

Q3. Distinguish between skewness and Kurtosis. Explain briefly the measures of skewness.

Q4. Calculate (i) Quartile Deviation (ii) Standard Deviation of marks from following data:

Marks	35-40	40-45	45-50	50-55	55-60	60-65	65-70
No. of Students	14	20	42	54	45	18	7

Q5. In two factories A and B, engaged in the same activity, the average weekly wages and standard deviation are as follows:

Factory	Average Weekly Wages (Rs.)	S.D. of Wages (Rs.)	No. of Wage Earners
A	460	50	100

B	490	40	80

(i) Which factory pays larger amount as weekly wages?

(ii) Which factory shows greater variability in the distribution of wages?

Q.6. +

Draw the Lorenz Graph for the two sets of five workers each whose weekly income figures are given below. Point out which set has greater inequalities in income.

Income of Group A ('00 ₹)	96	104	103	99	98
Income of Group B ('100 ₹)	100	270	580	620	430

Q.7. Calculate Bowley's Coefficient of skewness from the following data:

Income (in '000 Rs.): 1-5 6-10 11-15 16-20 21-25 26-30 31-35

No. of Families : 20 27 29 38 48 53 70

Q.8. Calculate Karl Pearson's coefficient of skewness from the following data:

Marks (More than):	0	10	20	30	40	50	60	70	80
No. of Students	150	140	100	80	80	70	30	14	0

Q .9. In a certain distribution, the first four moments about 5 are 2, 20, 40, 50. Calculate β_2 and state the type of distribution.

Chapter 5: Sampling

Structure

5.0. Objectives:

- To understand the need and importance of sampling
- To differentiate probability and non-probability sampling methods
- To learn types of probability sampling and how they differ from each other
- To learn types of non- probability sampling and how they differ from each other
- To differentiate sampling and non-sampling errors
- To understand the relation between sample size and errors

5.1. Introduction:

The objective of every research is to collect information about the characteristics of the population. Such information can be obtained by using census or a sample. When each and every unit of the population is studied, it is census method. However, in most of the cases it becomes difficult to study each and every element. Thus sample may be used for this purpose. A sample is a sub group of population presuming to have similar characteristics. It is used for making generalizations regarding the characteristics of the population from which it is selected. Sampling is a procedure, where in a fraction of the data is taken from a large set of data, and the inference drawn from the sample is extended to whole group. It is the important tool to obtain important information from the population.

5.2. Definitions of Related terms:

- **Population:** The entire set of individuals or other entities to which study findings are to be generalized.
- **Sample:** A subset of a population that is used to study the population as a whole.
- **Elements:** The individual members of the population whose characteristics are to be measured.
- **Sampling frame:** The list of all elements or other units from which sample is to be chosen. e.g. telephone directory
- **Target population:** The set of elements larger than or different from the population sampled and to which the researcher would like to generalize study findings containing the elements in a population.
- **Sampling Unit:** It is the basic unit containing the elements of the population to be sampled.
- **Sample Size:** The number of units in a sample is called sample size

- **Representative sample:** The sample that "looks like" the population from which it was selected in all respects which are potentially relevant to the study. The distribution of characteristics among the elements of a representative sample is the same as the distribution of those characteristics among the total population.
- **Probability sampling method:** The sampling method that relies on a random, or chance, selection method so that the probability of selection of population elements is known.
- **Nonprobability sampling method:** The Sampling method in which the probability of selection of population elements is unknown.

5.3. Sampling Techniques:

The most important distinction that needs to be made about the samples is whether they are based on a probability or a nonprobability sampling method. Sampling methods that allow us to know in advance how likely it is that any element of a population will be selected for the sample are termed **probability sampling methods.** Sampling methods that do not let us know in advance the likelihood of selecting each element are termed **nonprobability sampling methods**.

5.3.1 Probability sampling methods:

Probability sampling methods rely on a random, or chance, selection procedure, which is, in principle, the same as flipping a coin to decide which of two people "wins" and which one "loses." Heads and tails are equally likely to occur. Following are various probability methods:

5.3.1.1. Simple random sampling: A simple random sample from finite population is a sample selected such that each possible sample combination has equal probability of being chosen. It is also called unrestricted random sampling. **It** is a probability sampling procedure that gives every element in the target population, an equal chance of being selected. As such, it is an **equal probability selection method** (EPSEM).

Simple random sampling can be used in two ways:

- **Simple random sampling without replacement:** In this method the population elements can enter the sample only once i.e. the units once selected is not returned to the population before the next draw.
- **Simple random sampling with replacement:** In this method the population units may enter the sample more than once. Simple random sampling may be with or without replacement.

Methods of selection of a simple random sampling:

The following are the methods of selection of a simple random sampling.

- **Lottery Method:**

This is the most popular and simplest method. In this method all the items of the population are numbered on the separate slips, folded and mixed up in a container. The required numbers of slips are selected at random for the desire sample size. For example, if we want to select 10 students, out of the class of 50 students, then we must write the roll numbers of all the 50 students on slips and mix them. Then we make a random selection of 10 students.

This method is mostly used in lottery draws. If the universe is infinite this method is inapplicable.

- **Table of Random numbers:**

As the lottery method cannot be used, when the population is infinite, the alternative method is the table of random numbers. There are several standard tables of random numbers but Tippett' s table is most popular among them. The numbers in a table of random numbers are not arranged in any particular pattern. They may be read in any manner, i.e., horizontally, vertically, diagonally, forward, or backward. In using a table of random numbers, the researcher should blindly select a starting point and then systematically proceed down (or up) the columns or rows of numbers in the table. A random number table is so constructed that all digits 0 to 9 appear independent of each other with equal frequency. If we have to select a sample from population of size N= 100, then the three digit numbers are used by combining digits three by three to give the numbers from 001 to 100.

Procedure to select a sample using random number table:

Units of the population from which a sample is required are assigned with equal number of digits. When the size of the population is less than thousand, three digit number 000,001, 002,….. 999 are assigned. We may start at any place and may go on in any direction such as column wise or row- wise in a random number table. But the pattern of selecting numbers needs to be uniform. On the basis of the size of the population and the random number table available with us, we proceed according to our convenience. If any random number is greater than the population size N, then N can be subtracted from the random number drawn. This can be repeatedly until the number is less than N or equal to N.

Example 1:

In an area there are 500 families. Using the following extract from a table of random numbers select a sample of 15 families in that area.

4652 3819 8431 2150 2352 2472 0043 3488

9031 7617 1220 4129 7148 1943 4890 1749

2030 2327 7353 6007 9410 9179 2722 8445

0641 1489 0828 0385 8488 0422 7209 4950

Solution:

In the above random number table we can start from any row or column and read three digit numbers continuously row-wise or column wise.

Now suppose we start from the third row, the numbers are:

203 023 277 353 600 794 109 179 272 284 450 641 148 908 280

Since some numbers are greater than 500, we subtract 500 from

those numbers and we rewrite the selected numbers as follows:

203 023 277 353 100 294 109 179 272 284 450 141 148 408 280

Merits of using random numbers:

- Personal bias is eliminated as a selection depends solely on chance.
- A random sample is in general a representative sample for a homogenous population.
- There is no need for the thorough knowledge of the units of the population.
- The accuracy of a sample can be tested by examining another sample from the same universe when the universe is unknown.

Limitations:

- Preparing lots or using random number tables is tedious when the population is large.
- When there is large difference between the units of population, the simple random sampling may not be a representative sample.
- It is generally seen that the units of a simple random sample lie apart geographically. The cost and time of collection of data are more.

5.3.1.2 Stratified Random Sampling: Of all the methods of sampling the procedure commonly used in surveys is stratified sampling. This technique is mainly used to reduce the population heterogeneity and to increase the efficiency of the estimates. Stratification means division into groups. In this method the population is divided into a number of Subgroups or strata. The strata should be so formed that each stratum is homogeneous as far as possible and strata should be heterogeneous. Then from each strata a simple random sample may be selected and these are combined together to form the required sample from the population.

Types of Stratified Sampling:

There are two types of stratified sampling. They are **proportional** and **non-proportional**. In the proportional sampling, equal and proportionate representation is given to subgroups or strata. If the number of items is large in a group, the sample will have a higher size and vice versa.

- **Proportionate stratified sampling**

Sampling method in which elements are selected from strata in exact proportion to their representation in the population. The population size is denoted by N and the sample size is denoted by ' n' the sample size is allocated to each stratum in such a way that the sample fractions is a constant for each stratum. That is given by

n/N = c. So in this method each stratum is represented according to its size.

- **Disproportionate stratified sampling**

The method of sampling in which elements are selected from strata in different proportions from those appear in the population. In non-proportionate sample, equal representation is given to all the sub-strata regardless of their existence in the population.

Example 2:

A sample of 50 students is to be drawn from a population consisting of 500 students belonging to two institutions A and B. The number of students in the institution A is 200 and the institution B is 300. How will you draw the sample using proportional allocation?

Solution:

There are two strata in this case with sizes $N_1 = 200$ and $N_2 = 300$ and the total population $N = N_1 + N_2 = 500$

The sample size(n) is 50.

If n_1 and n_2 are the sample sizes,

$n_1 = (n*N_1)/N = (50 *200)/500=20$

$n_2 = (n*N_2)/N = (50 *300)/500=30$

Thus the sample sizes are 20 from A and 30 from B. Then the units from each institution are to be selected by simple random sampling.

Merits:

- ❖ It is more representative.
- ❖ It ensures greater accuracy.
- ❖ Greater geographical concentration reduces time and expenses.

- When the original population is badly skewed, this method is appropriate.

Limitations:

- The division of the population into homogeneous strata requires more money, time and statistical experience which is a difficult one.
- Proportionate stratification requires accurate information on proportion of population in each stratum
- Improper stratification leads to bias, if the different strata overlap such a sample will not be a representative one.

5.3.1.3 Systematic Sampling: In this method of sampling, sample elements are selected from a list. This method is frequently used method of sampling when a complete list of the population is available. It is widely employed because of its ease and convenience. In this method, the first unit is selected with the help of random numbers and the rest get selected automatically according to some pre designed pattern. With systematic random sampling, every k^{th} element in the frame is selected for the sample, with the starting point among the first k elements determined at random. For example, if we want to select a sample of 50 students from 500 students under this method k^{th}item is picked up from the sampling frame and k is called the **sampling interval**.

Sampling interval (k) = Sample size /Population size

$$= n/N = 50/500 = 10$$

k = 10 is the sampling interval. Systematic sample consists of selecting a random number say i and then every K th unit subsequently. Suppose the random number ' i' is 5, then we select 5, 15, 25, 35, 45,……. The random number ' i' is called random start. The technique will generate k systematic samples with equal probability.

Merits :

- This method is simple and convenient.
- Time and work is reduced much.
- If proper care is taken result will be accurate.
- It can be used in infinite population.

Limitations:

- Systematic sampling may not represent the whole population.
- There is a chance of personal bias of the investigators.
- Systematic sampling is preferably used when the information is to be collected house in blocks, entries in a register etc.

5.3.1.4 Cluster sampling: A method of sampling in which elements are selected in two or more stages, with the first stage being the random selection of naturally occurring clusters and the last stage being the random selection of elements within clusters. In this method, population is divided into non-overlapping or heterogeneous groups or clusters and a sample of clusters is taken to represent the population. However, there is homogeneity among the clusters while internally they are heterogeneous. A Cluster contains a wide range of elements that are good representatives of the population. In case the clusters are too large, a second set of clusters is taken from each original cluster .e.g. In a consumer survey, a country is divided in to cities, cities in to areas and then individuals are selected from selected areas. As the sample is being selected in to different stages, this method is known as Multi- Stage Sampling. Cluster sampling is useful when population is widely dispersed.

Merits:

- It involves less cost as well as time.
- As compared to stratified sampling, it is easy to obtain clusters.
- This is best method in general because of absence of sampling frame.

Limitations:

- A cluster may not be the true representative of the population when the elements of cluster are similar.

5.3.2 Non Probability Methods: In non- probability methods the probability of selection of each unit is unknown. Following are various non- probability methods:

5.3.2.1 Convenience Sampling: This method just considers the convenience of the researcher as the sole criterion. No effort is made to choose a representative sample. It is mostly used in exploratory research.e.g.choosing students to conduct an experiment, visiting nearby shops in locality to know which brand of the product is mostly preferred by residents.

Merits:

- Less time consuming
- Less cost is involved
- Sampling units are easily available

  Cluster sampling does not require a sampling frame of all of the elements in the target population.

Demerits:

- ❖ It suffers from selection bias.
- ❖ Sample is not representative of population.

5.3.2.2 Judgemental Sampling: In this method, the selection of the sampling unit is based on the judgement of the researcher, which is believed to be more representative. This method is used when the required information is possessed by a limited number of people i.e. small sample.

Merits:

- ❖ It is a quick and convenient method.
- ❖ Less cost is involved
- ❖ In case the judgement of the researcher is correct, the sample chosen is better representative.

Demerits:

- ❖ Results based on such sample cannot be generalized.
- ❖ Highly subjective method as it is entirely based on researcher's own judgement.

5.3.2.3 Quota Sampling: This is one of the non-probability method in which sub groups are formed like stratified sampling but no random method is used to select the sample from each group. Sample elements are selected on the basis of judgement or convenience. A quota is generally taken in proportion of sub groups in the population. In most of the researches, quotas are taken on the basis of demographics like age, income etc.

Merits:

- ❖ This is the best method in non-probability techniques as far as the representativeness is concerned.
- ❖ It involves relatively lesser cost

Demerits:

- ❖ Selection bias is present.

5.3.2.4 Snowball Sampling: In this method, an initial group of respondents is selected. Then these respondents are asked to identify others who belong to same target group. Thus subsequent respondents are selected on the basis of referrals. This method is generally used when one wants to estimate rare kind of characteristics.

Merits:

- ❖ This is a useful method in case of rare population as respondents can be located easily.

Limitations:

- ❖ Biasness is present in selection.
- ❖ Results based on such sample cannot be generalized.

5.4. Sample Size and Errors:

There are two types of errors in sampling i.e. sampling and non-sampling errors. With the increase in sample size, sampling errors decrease while non-sampling errors can be reduced by better supervision.

- *Sampling errors:* It is the difference between the parameter and statistics. It occurs because a part of the population has been used to estimate the population parameter. The difference is simply because of **chance** and the measure used to estimate the sampling error is the **standard error**. Such errors are known as random errors.

- *Non Sampling errors:* The accuracy of an estimate is also affected by errors arising from other causes such as incomplete coverage and faulty procedures of estimation, faulty sampling, observational errors, compiling errors etc. These errors are termed as non-sampling errors. Such errors are known as systematic errors. They can occur both in census and sample studies.

5.5. Summary: Sampling is a powerful tool for social science research. The process of selecting samples from the population is termed as sampling design. There are two types of sampling designs i.e. probability sampling design and non-probability designs. Probability sampling methods allow a researcher to use the laws of chance, or probability, to draw samples from which population parameters can be estimated with a high degree of confidence. But researchers do not come by representative samples easily. Well-designed samples require careful planning, some advance knowledge about the population to be sampled, and adherence to systematic selection procedures—all so that the selection procedures are not biased. There are four methods of probability sample designs: simple random sampling, stratified sampling, systematic sampling, and cluster sampling. Similarly there are four methods of non- probability sample designs: convenience sampling, judgemental sampling, quota sampling, and snowball sampling. Each method has its own merits and limitations and is used at appropriate situation. A researcher comes across two kinds of errors. Random errors i.e. sampling errors and systematic errors i.e. non-sampling errors. With the increase in sample size, sampling errors decrease while non-sampling errors can be reduced by better supervision.

5.6. Glossary:

- Population**:** The entire set of individuals or other entities to which study findings are to be generalized.
- Sample**:** A subset of a population that is used to study the population as a whole.
- Elements**:** The individual members of the population whose characteristics are to be measured.
- Sampling Frame**:** The list of all elements or other units from which sample is to be chosen. e.g. telephone directory
- Sampling Unit**:** It is the basic unit containing the elements of the population to be sampled.
- Sample Size**:** The number of units in a sample is called sample size

5.7. Suggested Readings:

- Malhotra, N.K. and Dash, S., Marketing Research: An applied orientation, Pearson Education
- Bajpai, N., Business Research Methods, Pearson Education
- Chawla, D. and Sondhi, N., Research Methodology- Concepts and Cases, Vikas Publishing House Pvt. Ltd.
- Kothari, Research Methodology

5.8. Model Questions:

1. What are the principal differences and similarities between the major categories of probability sampling: simple random sampling, systematic sampling, stratified sampling, and cluster sampling?

2. What are the similarities and differences between stratified sampling and quota sampling?

3. What is a sampling frame? Is it necessary to use a sampling frame in selecting a probability sample? Justify your answer.

4. What are the similarities and differences between cluster sampling and stratified sampling?

5. What are the principal differences and similarities between the major categories of non-probability sampling techniques?

6. Differentiate Sampling errors and Non-Sampling errors.

Chapter 6: Sampling Distribution and Theory of Estimation

6.0. Objectives

- To understand the concept of sampling distribution of sample mean.
- To understand the concept of sampling distribution of sample proportion.
- To know the concept of parameter and statistics.
- To know the concept of sample error.
- To know the concept of Estimation.
- To distinguish between point estimation and Interval estimation.
- To understand the concept of confidence limits and estimate them.
- To construct a confidence interval for the population mean when the population standard deviation is known
- To construct a confidence interval for the population mean when the population standard deviation is unknown
- To construct a confidence interval for the population proportion

6.1. Introduction

6.1.1 Sample

The set of observations which are taken from some source for the purpose of obtaining information about that source is called a sample, whereas the source of these observations is called the population. Each number in the population is called an observation and the number of observations in a sample is known as sample size.

6.1.2 Parameter

The value which describes the entire population characteristic is called a parameter. A parameter is denoted by 'μ' as mean and 'σ' as standard deviation.

6.1.3 Statistics

The values which are obtained from the sample data are called sample statistics. A sample statistics is denoted by '$\bar{x}$' as mean and's' as standard deviation.

The value of every statistic varies from one sample to another while the value of a parameter remains constant.

6.2. Sampling Distribution

When the results of the sample are collected (statistics), they are used to make the inferences about the population parameter. All the values of the statistics together with their corresponding frequencies constitute the sampling distribution. The values of the statistics varies from sample to sample. This variation in the values of the statistics is known as sampling fluctuation or sampling variation.

6.2.1 Sampling Distribution of Sample Mean

The sample mean is one of the common statistics used in the inferential process. A probability distribution of all possible sample means of a given sample is sampling distribution of the sample mean. Suppose, a population consists of four numbers i.e. 1,2,3,4; then the number of samples each containing two numbers without replacement will be 6 i.e.; (1,2), (1,3), (1,4), (2,3), (3,4) = 6 = $4C_2$.

In case of with replacement the number of samples will be 16 i.e., ; (1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4), (4,1), (4,2), (4,3), (4,4) = 4^2

Arithmetic mean of the samples are:

Sample	**A.M.**
1,2	1.5
1,3	2
1,4	2.5
2,3	2.5
2,4	3
3,4	3.5

Sampling Distribution of Mean

Mean	1.5	2	2.5	3	3.5
Frequency	1	1	2	1	1

Mean of the sampling distribution of Mean

$$= \frac{1.5*1+2*1+2.5*2+3*1+3.5*1}{1+1++2+1+1}$$

$$= \frac{15}{6} = 2.5 = \text{Mean of the population}$$

Thus, the mean of the population and the mean of the sampling distribution of the mean are the same.

Example: XYZ industries have seven departmental managers (considered the population). The hourly earnings of each manager are given in table below:

Hourly earnings of the managers of XYZ Industries

Departmental Manager	Hourly Earnings
A	$7
B	$7
C	$8
D	$8
E	$7
F	$8
G	$9

1. What is the population mean?
2. What is the sampling distribution of the sample mean for the samples of size 2?
3. What is the mean of the sampling distribution?

Solution: 1. The population mean is $7.71, found by:

$\mu = \Sigma X/N = (\$7+\$7+\$8+\$8+\$7+\$8+\$9)/7 = \$7.71$

2. To arrive at the sampling distribution of the sample mean, we need to select all possible samples of 2 without replacement from the population, then compute mean of each sample. There are 21 possible samples.

$N_{Cn} = N! / [n!(N-n)!]$

$= 7! / [2! (7-2)!] = 21$

Where N=7 is the number of items in the population and n=2 is the number of items in the sample.

The 21 sample means from all possible samples of 2 that can be drawn from the population are shown in the following table.

Sample Mean for All Possible Samples of 2 Employees Each

Sample	Employees	Hourly earnings $	Sum	Mean
1	A, B	7, 7	14	7
2	A, C	7, 8	15	7.5
3	A, D	7, 8	15	7.5
4	A, E	7, 7	14	7
5	A, F	7, 8	15	7.5
6	A, G	7, 9	16	8
7	B, C	7, 8	15	7.5
8	B, D	7, 8	15	7.5
9	B, E	7, 7	14	7
10	B, F	7, 8	15	7.5
11	B, G	7, 9	16	8
12	C, D	8, 8	16	8
13	C, E	8, 7	15	7.5
14	C, F	8, 8	16	8

15	C, G	8, 9	17	8.5
16	D, E	8, 7	15	7.5
17	D, F	8, 8	16	8
18	D, G	8, 9	17	8.5
19	E, F	7, 8	15	7.5
20	E, G	7, 9	16	8
21	F, G	8, 9	17	8.5

These 21 sample means are used to construct a probability distribution. This is the sampling distribution of the sample mean as shown in following table.

Sampling Distribution of the Sample Mean for n = 2

Sample Mean	Number of Means	Probability
$ 7	3	.1429
$7.5	9	.4285
$8	6	.2857
$8.5	3	.1429
	21	1.0000

3. The mean of the sampling distribution of the sample mean is obtained by summing the various sample means and dividing the sum by number of samples.

= Sum of all Sample means/Total number of samples

= ($7 + $7.5 +$8+$8.5)/21 = $7.71

The above example illustrates the important relationships between the population distribution and the sampling distribution of the sample mean:

1. The mean of the sample means is exactly equal to the population mean.
2. The dispersion of the sampling distribution of the sample means is narrower than the population distribution.
3. The sampling distribution of the sample means tend to become bell-shaped and to approximate the normal probability distribution

6.2.2 Sampling Distribution of Sample Proportion

When data are measurable like weight, income, time etc. sample mean is an appropriate measure. But in case data contains countable items like number of credit card holders, the sample proportion p is an appropriate measure. Sample proportion, p is used to make inference about the population proportion, p. The sample proportion can be obtained by dividing the frequency with which a given characteristic occurs in a sample by the number of units in the sample; i.e;

$$\overline{p} = \frac{x}{n}$$

Where, x = number of units in the sample having the given characteristics.

n = number of units in the sample

The mean of the sample proportion for all the samples of size 'n' drawn from the population is 'p' and standard deviation is $\sqrt{\frac{pq}{n}}$.

For the large sample size ($np \geq 5$) the sampling distribution of proportion can be approximated by a normal probability distribution.

$$Z = \frac{\overline{p} - p}{\sqrt{\frac{pq}{n}}}$$

Where, $\overline{p}$ = sample proportion

p = population proportion

q = 1 - p

n = sample size

If a population consists of 5 elements 2,3,6,5,18. The population proportion of even numbers = $\frac{3}{5} = 0.6$. The random samples of size 3 drawn without replacement are 10 i.e., ${}^{5}C_{3}$

Sample	Proportion of even numbers	Sampling Distribution of Proportion
(2,3,6)	2|3	
(2,3,5)	1|3	
(2,3,18)	2|3	2/3
(2,6,5)	2|3	
(2,6,18)	1	1/3
(2,5,18)	2|3	
(3,6,5)	1|3	
(3,6,8)	2|3	1
(3,5,18)	1|3	
(6,5,18)	2|3	

Sampling distribution of proportion

Proportion	1|3	2|3	1
Probability	3|10	6|10	1|10

Mean of the sampling distribution of proportion

$$= \frac{1}{3} * \frac{3}{10} + \frac{2}{3} * \frac{6}{10} + 1 * \frac{1}{10}$$

$= {}^{6}|_{10} = 0.6$

Thus, the proportion of even numbers in the population and the mean of the distribution of proportion is the same.

6.3. Standard Error

The standard error of a statistics i.e., sample mean or sample proportion or sample standard deviation is the standard deviation of the sampling distribution of that statistics. It is used in Hypothesis testing in order to know the accuracy of the sample.

The S.E. of $\bar{x}$ i.e.; $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Where, σ = population standard deviation

n = sample size

Smaller the value of standard error, more accurate is the sample result.

The S.E. of $\bar{p}$ i.e., $\sigma_{\bar{p}} = \sqrt{\frac{pq}{n}}$

Where, p = population proportion

q = 1 – p

n = Sample size

6.4 Sampling Error

Sample is a fair and unbiased representation of the population, thus used to estimate population characteristic. However, since the sample is a portion of the population, it is unlikely that the sample mean would be exactly equal to the population mean. Thus the difference will be there in between a sample statistic and its population parameter. This difference is known as sampling error.

Self-Assessment Exercises 1:

1. The average life of a cutting tool is 41.5 hrs. with standard deviation of 2.5 hrs. What is the probability that a simple random sample of size 50 drawn from this population will have a mean between 40.5 hrs and 42 hrs?

2. A population of items has an unknown distribution but a known mean and standard deviation of 50 and 100, respectively. Based upon a randomly drawn sample of 81 items drawn from the population, what is the probability that the sample arithmetic mean does not exceed 40?
3. The quality control department of a toy manufacturing company, at the time of dispatch of toys discovered that 30 percent of the products are defective if a random sample of 500 toys is drawn with replacement from the population, what is the probability that the sample proportion will be less than or equal to 25 percent defective.

4. State True or False.

   i) The sampling distribution provides the basis for statistical inference when sample results are analysed

   ii) The sampling distribution of mean is the probability density function that describes the distribution of the possible values of a sample mean.

   iii) The expected value (mean) is equal to the population mean from which the sample is chosen.

   iv) As the sample size is increased, the sampling distribution of the mean approaches the normal distribution regardless of the population distribution.

   v) Standard error of the mean is the standard deviation of the sampling distribution of the mean.

6.5 Introduction to Theory of Estimation

The theory of estimation in statistics deals with estimation of population parameters like mean of a statistical distribution. It is assumed that the concerned variable of the population follows a certain distribution with some parameter(s).

Sampling is used to draw inferences about the population, thus, it helps to estimate the population parameter. An *estimator* of a population parameter is a random variable that depends upon the sample information, while an *estimate* is a specific value of that random variable.

6.6 Point Estimation

A point estimate is a statistic taken from a sample that is used to estimate a population parameter. It is the value of a single sample statistic such as a sample mean. It is not possible to determine that 'best' point estimator in all circumstances, e.g., if many random samples are taken from the population, the point estimates derived from those samples are likely to vary.

6.6.1 Properties of Point Estimator

- ***Consistency***

Consistency leads to increase the accuracy of the estimator. Lower the differences between the estimator and the population parameter, more accurate is the result.

- ***Unbiased Estimator***

The estimator which tends to be near the population parameter value, even in the case of small sample is an unbiased estimator and is always desirable.

- ***Efficiency***

An estimator with a smaller variance is more near to population parameter, such an estimator is an efficient estimator.

- ***Sufficiency***

A statistics is said to be sufficient for a parameter if it extracts the maximum information from the sample relating to the parameter.

6.7 Interval Estimation

Due to variations in the sample statistics, estimation of the population parameters with an interval estimate is preferred to a point estimate. An interval estimate is the range of values that are likely to include in a parameter. In case of interval estimates, one needs to indicate the confidence of correctly estimating the value of population parameter. Thus, one can say that

there is a specified confidence that parameter is somewhere in the range of numbers defined by that interval. The intervals or limits are known as confidence interval or confidence limits. Thus, confidence interval is a range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability. The specified probability is called the level of confidence.

6.8 Confidence Intervals for Various Parameters

6.8.1 Confidence Interval for Population Mean (when σ is known)

$(100-\sigma)\%$ confidence Interval is $\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Where, $\bar{x}$ is sample mean; σ is population standard deviation; $Z_{\alpha/2}$ is the point on the normal curve area beyond which on either side is $\alpha/2$.

6.8.2 Confidence Interval for Population Mean (when σ is unknown)

In case population standard deviation is unknown, it has to be estimated for the sample.

Thus, $(100-\sigma)\%$ confidence Interval is

$$\bar{x} \pm t_{\alpha/2},(n-1)\frac{S}{\sqrt{n}}$$

Where, $\bar{x}$ is sample mean; *s* is the sample standard deviation; n is the sample size and $t_{\alpha/2},(n-1)$ is the value in *t* table with (n-1) degrees of freedom area beyond which on either side is $\alpha/2$.

Example: If the average number of customers coming to a bank's branch is found to be 20 per hour, determine the 95% confidence limits for the number of customers in an hour.

Solution: It is assumed that the variable representing the number of customers coming in an hour follows Poisson distribution whose mean is given as 20. Then the 95% confidence limits for average number of customers in an hour are

$20 \pm 1.96\ (20/5)^{1/2}$

$20 \pm 1.96 * 2$

20 ± 3.92

16.08 and 23.92 are the required confidence limits.

Example: The Punjab State Management Association wishes to have information on the mean income of managers in the textile industry. A random sample of 256 managers reveals a sample mean of INR 45, 420. The standard deviation of this population is INR 2050. The association wishes to know the answers to the following questions:

1. What is the population mean?
2. What is the reasonable range of values for the population mean?

Solution: Generally, distributions of income and salary are positively skewed, because a few individuals earn considerably more than others, thus skewing the distribution in the positive direction. Fortunately, the central limit theorem stipulates that if we select a large sample, the distribution of sample means will follow the normal distribution. In this instance, a sample of 256 middle managers is large enough that we can assume that the sampling distribution will follow the normal distribution.

1. What is the population mean?

In this case, we do not know population mean. We do know the sample men is INR 45,420. Hence, our best estimate of the unknown population value is the corresponding sample statistic. Thus, the sample mean of INR 45,420 is a point estimate of the unknown population mean.

2. What is the reasonable range of values for the population mean?

The association decides to use the 95% level of confidence. To determine the corresponding confidence intervals we use formula

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 45420 \pm 1.96 * 2050/\sqrt{256} = 45420 \pm 251$$

The endpoints are INR 45169 and INR 45671. These endpoints are called confidence limits. The degree of confidence or the level of confidence is 95% and the confidence interval is from INR 45,169 to INR 45671. The ±INR 251 is often referred to as the margin of error.

6.8.3. Confidence Interval for Population Proportion Mean

Confidence Interval for the population proportion at σ is

$$\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

Where $\bar{p}$ is sample proportion, *p* is population proportion, q = 1-p, n is sample size.

Example: The union representing the B Co. Ltd. is considering a proposal to merge with the T Co. Ltd. According to B Co. union bylaws, at least three fourths of the union membership must approve any merger. A random sample of 2,000 current B Co. members reveals that 1600 plan to vote for the merger proposal.

1. What is the estimate of the population proportion?

2. Develop a 95% confidence interval for the population proportion.

Solution:

1. Calculate the sample proportion

$\bar{p} = 1600/2000 = 0.80$

Thus, we estimate that 80% of the population favors the merger proposal.

2. Determine the 95% confidence interval using formula $\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{pq}{n}}$.

= 0.80 ± 1.96 $\sqrt{\frac{0.80 * 0.20}{2000}}$ (The z value corresponding to the 95% level of confidence is 1.96)

= 0.80 ± 0.018

The endpoints of the confidence interval are .782 and .818.

Self-Assessment Exercises 2:

(1) In order to introduce some incentive for higher balance in savings accounts, a random sample of size 64 savings accounts at a bank's branch was studied to estimate the average monthly balance in savings bank accounts. The mean and standard deviation were found to be Rs. 8,500 and Rs, 2000, respectively. Find (a) 90%, (b) 95%, (c) 99% confidence intervals for the population mean.

(2) For assessing the number of monthly transactions in credit cards issued by a bank, transactions in 25 cards are analysed. The analysis revealed an average of 7.4

transactions and sample standard deviation of 2.25 transactions. Find confidence limits for the monthly number of transactions by all the credit card holders of the bank?

(3) A random sample of 100 items is taken, producing a sample mean of 49. The population standard deviation is 4.49. Construct a 90% confidence interval to estimate the population mean.

(4) Coopers & Lybrand surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management succession plan in place. A spokesperson for Cooper & Lybrand said that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocus a company for long enough to cause it to lose its momentum.

Use the data given to compute a 92% confidence interval to estimate the proportion of all fast-growing small companies that have a management succession plan.

(5) A clothing company produces men's jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men's jeans market in Oklahoma City that prefers boot cut jeans, the analyst takes a random sample of 423 jeans sales from the company's two Oklahoma City retail outlets. Only 72 of the sales were for boot cut jeans. Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot cut jeans.

6.9 Summary

A probability distribution of all possible sample means of a given sample size is known as sampling distribution of the sample mean. The mean of the distribution of sample means is equal to the population mean. The standard error of the mean measures the variation in the samplg distribution of the sample mean. The difference between a sample statistics and its corresponding population parameter is known as sampling error.

An estimator of a population parameter is a random variable that depends upon the sample information, while an estimate is a specific value of that random variable. A point estimate is a single value derived from a sample and used to estimate a population value. An interval estimate is the range of values constructed from sample data and the population parameter is likely to occur within that range at a specified probability (Level of Confidence).

6.10 Glossary:

- **Confidence Interval or Limits:** The interval or limits within which the true value of the parameter lies.
- **Confidence Level:** It is expressed in percentage e.g. 95%, and indicates the degree of confidence that the true value of the parameter lies in the specified interval.
- **Estimator:** A function of sample values to estimate a parameter of a population.
- **Point Estimation:** A single value estimate like 10.
- **Interval Estimation:** Estimate in the form of an interval, say from 12 to 20.
- **Consistent Estimator:** A property which implies that the estimate tends to the true value of the parameter as the sample size increases.
- **Efficient Estimator:** A property which implies that the variance of the estimator is minimum as compared to any other estimator.
- **Unbiased Estimator:** A property which implies that its expected value is equal to the population value.
- **Sampling Distribution:** A probability distribution consisting of all possible instance as an estimate of a population parameter.
- **Parameter:** A numerical characteristics of a population, such as a population mean μ, a population standard deviation σ, a population proportion p, and so on.
- **Sample Statistics:** A sample characteristic, such as a sample mean, a sample standard deviation, a sample proportion, and so on. The value of the sample statistic is used to estimate the value of the corresponding population parameter.
- **Standard Error:** Standard Deviation of an estimate from a sample

6.11 Answers of SAQ's 1:

1. [Ans. 0.9184]
2. [Ans. 0.1841]
3. [Ans. 0.0083]
4. [Ans: i. T ii. T iii. T iv. T v. T]

Answers of SAQ's 2:

1. [Ans. (a) 8911.25 ; (b) 8990 ; (c) 9144]
2. [Ans. 6.473 – 8.327]
3. [Ans. 48.26 - 49.74]
4. [Ans. 0.45 – 0.57]
5. [Ans. 0.14 – 0.20]

6.12 Suggested Readings

- Black, K. Business Statistics for Contempory Decision Making, Fifth Edition, Wiley India.
- Bajpai, N., Business Statistics, Pearson Education
- Srivastava, T.N. and Rego,S., Statistics for Management, Fourth Reprint, Tata McGraw Hill Companies.
- Thukral, J.K. Business Statistics, Second Edition, TAXMANN'S
- Kothari. C.R., . Research Methodology

6.13 Terminal and Model Questions

1. What is the concept of sampling distribution? State the importance of sampling distribution in inferential statistics.
2. Is the standard deviation of sampling distribution of mean the same as the standard deviation of the population? Explain.
3. The Star TV network is considering replacing one of its prime - time crime investigation show with a new family-oriented comedy show. Before a final decision is made, network executives commission a sample of 400 viewers. After viewing the comedy, 250 indicated they would watch the new show and suggested it to replace the crime investigation show.

   Estimate the value of population proportion.

   Develop a 99% confidence interval for the population proportion.

   Interpret your findings**.**
4. A population consists of five values: 2, 2, 4, 4, and 8.
   a. List all samples of size 2, and compute the mean of each sample.
   b. Compute the mean of the distribution of sample means and population mean. Compare the two values.

c. Compare the dispersion in the population with that of the sample means.

5. A population consists of five values: 12, 12, 14, 15, and 20.

    a. List all samples of size 3, and compute the mean of each sample.

    b. Compute the mean of the distribution of sample means and population mean. Compare the two values.

    c. Compare the dispersion in the population with that of the sample means.

6. A random sample of 15 items is taken, producing a sample mean of 2.364 with a sample variance of 0.81. Assume x is normally distributed and construct a 90% confidence interval for the population mean.
7. For a random sample of 36 items and a sample mean of 211, compute a 95% confidence interval for μ if the population standard deviation is 23.
8. The distribution of the annual earnings of the employees of a cement factory is negatively skewed. This distribution has a mean of Rs. 25,000 and standard deviation of Rs. 3000. If a researcher draws a random sample of size 50, what is the probability that their average earnings will be more than Rs. 26,000?
9. In a departmental store, the mean expenditure per customer is Rs. 1850 with a standard deviation of Rs. 750. If a random sample of 100 customers is selected, what is the probability that the sample average expenditure per customer for this sample is more than Rs. 2000.
10. In a grocery store, the mean expenditure per customer is Rs. 2000 with a standard deviation of Rs. 300. If a random sample of 50 customers is selected, what is the probability that the sample average expenditure per customer is more than Rs. 2500?

.

CHAPTER 7: TESTING OF HYPOTHESIS

Structure

7.0 Objectives

- To understand the meaning, types and process of Hypothesis Testing.
- To know how to develop one tail and two tail Hypothesis to be tested by taking samples and arriving at statistical conclusions.
- To understand the difference between parametric tests and non-parametric tests.
- To understand the applications of z- test and t- test in case of single sample.
- To understand the applications of z- test and t- test in case of two samples.

7.1 Introduction

In business, a businessman has to take various types of decisions depending upon his situations and requirements. In order to decide he takes the help of research. Researcher develops some hypothesis that can be explored so that the decisions can be taken in this light. Hypothesis is nothing but an assumption. Suppose a researcher wants to know whether the new machinery has improved the production per month. He will assume that new machinery has not changed the production pattern. To evaluate, he will check the production after the new machine has been put in use.

7.1.1 Hypothesis is of two types

- Null Hypothesis(H_0)
- Alternate Hypothesis(H_a)

Null Hypothesis is a hypothesis that will be maintained unless there is a strong evidence against it, i.e. a hypothesis of no difference. It is denoted by H_0. H_0 expresses default believes that we accept in the absence of data.

Alternate Hypothesis is the one which is to be accepted in case null hypothesis is rejected. Thus, it is a hypothesis that contradicts H_0. It is denoted by H_a.

7.1.2 One Tail and Two Tail Test

Parametric tests are based on the assumption of normal distribution. H_0 is rejected if the sample statistics is significantly higher or lower than population parameter. Thus, in two tail test the rejection region is equally distributed in both tails of the sampling distribution. In one tail test, rejection region is located in one tail only which may be right or left depending upon the formulation of hypothesis.

For example, suppose the null hypothesis is that the wages of men and women are equal. A two-tailed alternative would simply state that the wages are not equal – implying that men could make more

wages than women, or they could make less. A one-tailed alternative would be that men would make more wages than women. The latter is a stronger statement and requires more theory, because here not only one is claiming that there is a difference, but also the direction of the difference. Thus such kinds of problems are taken as directional problems. **Directional tests**, or **one-tailed tests**, are hypothesis tests where the alternative hypothesis is stated as greater than (>) or less than (<) a value stated in the null hypothesis. Hence, the researcher is interested in a specific alternative from the null hypothesis.

7.1.3 Types of errors

During hypothesis testing a Researcher might take a wrong decision.

Possible action	**Possible alternatives**	
	H_0 is true	**H_0 is false**
Accept H_0	Correct	Incorrect(type II error)
Reject H_0	Incorrect (type I error)	Correct

Type I Error is the error of taking wrong decision i.e. the probability of rejecting a true hypothesis. It is also known as the **level of significance** and is denoted by α.

α = Probability of Type I error = P(rejecting H_0 | H_0 is true)

Typical values chosen for α are .05 or .01. So, for example, if α = .05, there is a 5% chance that, when the null hypothesis is true, we reject it.

Type II Error is the error of accepting a false null hypothesis. It is denoted by β. We accept the null hypothesis when it is not true, we commit type II error.

β= Probability of Type II error = P (accepting H_0 | H_0 is false) α and $\beta$are not independent of each other - as one increases, the other decreases. The decrease in one type of error leads to an increase in the other. Thus, a proper balance of the two errors is required.

However, increase in N causes both to decrease, since sampling error is reduced.

7.1.4 Power of a Test

It measures how well hypothesis test is working. It is a probability that the test will correctly reject a false null hypothesis. It is denoted by '1-β'.

7.1.5 Parametric Vs Non- Parametric Tests

Parametric tests are statistical techniques to test a hypothesis based on assumption that population is normally distributed and samples have been selected randomly from the normal population. Moreover sample data should be quantitative in nature i.e. it should be in interval or ratio scale.

Non- Parametric tests are applicable if data doesn't follow a normal distribution. They are known as distribution free tests. These are applicable to nominal and ordinal data. They do not require a sample to be very large in size.

7.2 Procedure for Hypothesis Testing

1. **Setting up of Hypothesis:** Null and Alternate hypothesis are set up as a first step of Hypothesis testing.
2. **Specify the Level of Significance (α):** $\alpha$is specified in advance, generally taken as 5% or 1%.
3. **Choosing the Test Statistics:** An appropriate test depending upon the nature of the problem is selected.

   Value of test statistics = (observed value – expected value)/standard error of the test
4. **Making Decision:** in order to take the decision i.e., whether to accept or reject H_0, critical values (table values) are used and compared with a calculated value (from step 3).

   If calculated value < critical value, H_0 is accepted.

   If calculated value > critical value, H_0 is rejected.

7.3 Parametric Tests

7.3.1 Hypothesis Testing for One Sample: When we know the mean and standard deviation in a single population, we can use the **one–independent sample *z* test**.

The **one–independent sample *z* test** is a statistical procedure used to test hypotheses concerning the mean in a single population with a known variance**.**

7.3.1.1 Test of Significance of Mean (Large sample): if sample size (n) ≥ 30 and standard deviation α is known, Sampling distribution of a Statistic approximates a normal distribution and z- test is used.

Null Hypothesis: H_0: $\mu = \mu_0$

Alternate Hypothesis: H_a: $\mu \neq \mu_0$

$$z = \frac{\bar{x} - \mu}{S.E\ of\ \bar{x}}$$

Where $\bar{x}$ is sample mean and μ is population mean.

$$S.E\ of\ \bar{x} = \frac{\sigma}{\sqrt{n}}$$

Where σ is population standard deviation and n is sample size.

Calculated value of Z is compared with the tabulated value.

If $|Z_{cal}| <$ tab v, H_0 is accepted

If $|Z_{cal}| >$ tab v, H_0 is rejected

Critical Value (Tabulated Value) of Z

Level of Significance (α)	**1%**	**5%**
Two Tail Test	2.58	1.96
One Tail Test	2.33	1.645

Example: The mean life of a sample of 100 electric tubes produced by a company is found to be 1570 hours with a S.D. of 120 hours. If μis the mean life time of all the tubes produced by the company, test the hypothesis μ= 1600 hours using a level of significance of 0.05.

Solution: The null hypothesis is H_0: μ= 1600 hours

The alternative hypothesis is H_a: μ≠1600 hours

The test statistic is $z = \frac{\bar{x}-\mu}{S.E\ of\ \bar{x}}$; where $S.E\ of\ \bar{x} = \frac{\sigma}{\sqrt{n}}$

So, $S.E\ of\ \bar{x} = \frac{\sigma}{\sqrt{n}}$ = 120/10 = 12 hours

Thus, Z = (1570 – 1600) / 12 = -30 / 12 = -2.5 and IZI = I-2.5I = 2.5 > 1.96

This shows that Z = -2.5 lies in the critical region IZI > 1.96

Hence, null hypothesis is rejected and we conclude that the mean life of the population of electric tubes cannot be taken as 1600 hours.

7.3.1.2 Test of Significance of Mean (Small sample): When n< 30 and standard deviationσ is unknown (t- test).

In case sample size i.e. n< 30, standard deviation (σ) is unknown, z- test is not appropriate to use. In such a case, t- test is applied, t- test is applied even in case of large sample if population standard deviation (α) is unknown. In this case, the appropriate formula is:

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n-1}}$$

Where $\bar{x}$ is the sample mean; S is sample standard deviation and (n-1) is degrees of freedom.

Degrees of freedom: For a fixed number of mean, the number of free choices is called the degree of freedom. In general, degree of freedom is defined as:

Number of frequencies – Number of independent constraints

i.e. n-1

Calculated t value is compared with table value of t at a particular level of significance and degree of freedom. If $|t_{cal}| <$ tab value, H_0 is accepted.

7.3.1.3 Hypothesis testing for a population proportion

In business situations, most of the information is expressed in percentage form, e.g. market share, proportion of defectives etc. In this case, the z- test for population proportion is applicable. It is based on the difference between sample proportion and population proportion.

H_0: $P=P_0$

H_a: $P \neq P_0$; $P < P_0$; $P > P_0$

$$Z = \frac{p - P}{S.E\ of\ p}$$

S.E of p = $\sqrt{\frac{P*Q}{n}}$

Where p is a sample proportion, P is the population proportion, n is the sample size.

Example: In a sample of 750 products manufactured by a company, the number of defective products was observed to be 30. The purchaser, however, claimed that 5% of their product is defective. Is the claim justified?

Solution:

P_0 = expected proportion of defective articles = 5/100 =0.05

P = observed proportion of defective articles = 30/750 = 0.04

Let the null hypothesis H_0: P = 0.05

Then alternative hypothesis is H_a: P<0.05

The test statistic is

$$Z = \frac{p - P}{S.E\ of\ p}$$

S.E of $p = \sqrt{\frac{P*Q}{n}}$ = $[(0.05*0.95)/750]^{1/2}$ = $(0.0475/750)^{1/2}$ = 0.007956

$Z = \frac{0.04-0.05}{0.007956}$ = - 1.323

$|z| = 1.323$

Calculated value of Z (1.323)> tabulated value (1.646). H_0 is accepted.

Thus purchaser's claim is justified.

Self- Assessment Exercise 1:

1. A soft drink vending machine is set to that the amount of drink dispensed is a random variable with a mean 200 ml and a standard deviation of 15 ml. what is the probability that the average amount dispensed in a random sample of size 36 is at least 204 ml?
2. A simple random sample of size 66 was drawn in the process of estimating the mean annual income of 950 families of a certain township. The mean and standard deviation of the sample were found to be Rs. 4, 730 and Rs. 7.65 respectively. Find a 95% confidence interval for the population mean.
3. Traditionally 35% of all loans by a National Bank have been to members of minority groups. During the past year the bank has undertaken efforts to increase this proportion. Of 150 loans currently outstanding 56 are identified as having been made to minorities. Has the bank been successful in its efforts to attract more minority customers? Test the hypothesis using the 5% level of significance.

4. A company can claim that the weight of their product is 10 kgs. A sample of items taken from a lot supplied by the company has shown the following weights:

   10.2, 9.7, 10.3, 10, 9.8, 9.7, 9.6, 9.6, 9.7, 9.4

   Is there any statistical evidence to support the claim of the company about the weight of the item?

5. A brand of matches is sold in boxes on which it is claimed that the average contents are 40 matches. A check on the pack of 5 boxes gives the following results:

   41, 39, 37, 40, 38

   Test the manufacturer's claim.

7.3.2 Hypothesis testing for the difference between two population means

In the business world many a times situations arise when the need is to deal with two samples instead of one e.g. to know the difference in sales of product in two different regions, difference of output of two production processes etc. Thus researcher will have to take two different samples from two different populations. The two are then compared and inference is drawn on the basis of sample means. The two population means are denoted as μ_1 and μ_2. To make an inference, sample of n_1 units from population 1 and sample of n_2 units from population 2 are randomly selected.

7.3.2.1 Test for equality of two means: The two samples are selected independently so known as Independent Random Samples. It is assumed that two population standard deviations σ_1 and σ_2 are known. In case both n_1 and $n_2 \geq 30$(large samples), Z- test formula for difference between mean values of two populations is used. For testing equality of two means the following procedure is applied.

H_0: $\mu_1=\mu_2$

H_a: $\mu_1 \neq \mu_2$, $\mu_1 < \mu_2$, $\mu_1 > \mu_2$

Where μ_1 and μ_2 are population means.

$$Z = \frac{\overline{x_1} - \overline{x_2}}{S.E\ of\ (\overline{x_1} - \overline{x_2})}$$

S.E of $(\overline{x_1} - \overline{x_2}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Where σ_1 and σ_2 are Standard deviation of two populations.

7.3.2.2 Test for equality of two means when σ_1 and σ_2 are unknown

Let x_1 and x_2 be sample means of two independent samples of sizes n_1 and n_2 drawn from two normal populations with means as μ_1 and μ_2 respectively and n_1 and n_2 each<30.In order to test whether the two population means are equal ,t-test is applied. As $\sigma_1$andσ_2 are unknown s_1 and s_2 (being unbiased estimators)are used which are sample standard deviations for sample 1 and sample 2 respectively.

H_0: $\mu_1=\mu_2$

H_a: $\mu_1=\mu_2$, $\mu_1<\mu_2$, $\mu_1>\mu_2$

$$t = \frac{\bar{x_1} - \bar{x}_2}{S\sqrt{1/n_1 + 1/n_2}}$$

Where is distribution with (n_1+n_2-2) degrees of freedom

S.E of $(\bar{x_1} - \bar{x_2}) = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Where, $S = \sqrt{\frac{n_1 s_2^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$

Calculated value of t is compared with table value at given α and d.f. If calculated value of t is less than table value, H_0 is accepted, otherwise it is rejected.

Example: A sample of size 10 is drawn from each of two normal populations with the same unknown variance and following results are obtained:

	Mean	Variance
Sample I:	10	30
Sample II:	8	14

Test at 5% significance level if the two populations have the same mean.

Solution: n_1 = 10, n_2 = 10, $\bar{x}_1$= 10, $\bar{x}_2$= 8, s_1^2 = 30, s_2^2 = 14

Since the sample sizes are small and σ_1, σ_2 are not known, t-test will be used.

Let μ_1 and μ_2 be the two population means.

Let the null hypothesis be H_0: $\mu_1=\mu_2$

Then the alternative hypothesis will be H_a: $\mu_1 \neq \mu_2$

$$t = \frac{\overline{x_1} - \overline{x}_2}{S\sqrt{1/n_1 + 1/n_2}}$$

$S^2 = (n_1 s_1^2 + n_2 s_2^2)/ r_1 + n_2 - 2 = (10*30) + (10*14)/10 + 10 - 2 = 440/18 = 24.44$; so, $S = \sqrt{24.44}$

$t = 2/\sqrt{24.44} * \sqrt{2/10} = 2/2 = 1$

d.f. = 10+10-2 = 18

Tabulated value of t = 2.101

Since tabulated value of t is > calculated value of t;

H_0 is accepted and we conclude that the difference between the two means is not significant.

7.3.2.3 Hypothesis testing for difference between two related population means: Matched pairs (Paired t- test)

A businessman sometimes may be interested to study the effect some action; e.g. effect of advertising or sales promotion campaign on the volume of sales of a product, effect of medicine in reducing the level of disease, effect of incentives/training in improving the performance of the employees etc. In such situations, the measurement or collection of information is done at two times. i.e. one before the introduction of variable and one after that .Then the two kinds of data is matched to see the significant difference in the pairs of data using paired t –test. Let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ be the data collected before and after the introduction of variable whose effect is to be measured. Difference is taken as:

$D_i = x_i - y_i$ or $y_i - x_i$; where I = 1,2,. .....n

H_0: $\mu_1 = \mu_2$ or d=0

H_a: $\mu_1 \neq \mu_2, \mu_1 < \mu_2, \mu_1 > \mu_2$ or $d \neq 0, d < 0, d > 0$

$t = \frac{\bar{d}}{S/\sqrt{n}}$; where $\bar{d} = \sum \frac{d}{n}$ and $s = \sqrt{\sum \frac{d^2 - n(\bar{d})^2}{n-1}}$; It follows (n-1) degrees of freedom

Example: A test was administered to 10 students before and after they were given coaching. The results are given below:

Students	1	2	3	4	5	6	7	8	9	10

Marks before coaching	167	124	157	155	163	154	156	168	133	143
Marks after coaching	170	138	158	158	156	167	168	172	142	138

Test whether there is any change in marks after coaching.

Solution: Let the null hypothesis be H_0: $\mu_1=\mu_2$ i.e., there is no significant effect of the training.

Then the alternative hypothesis is H_1: $\mu_1\neq\mu_2$

Students	Marks before coaching(x)	Marks after coaching(y)	d = y-x	d^2
1	167	170	3	9
2	124	138	14	196
3	157	158	1	1
4	155	158	3	9
5	163	156	-7	49
6	154	167	13	169
7	156	168	12	144
8	168	172	4	16
9	133	142	9	81
10	143	138	-5	25
Total			47	699

$t = \frac{\bar{d}}{S/\sqrt{n}}$; where $\bar{d} = \sum\frac{d}{n}$ and $s = \sqrt{\sum\frac{d^2-n(\bar{d})^2}{n-1}}$

so, $\bar{d}$= 47/10 = 4.7; and $S^2 = 699 - 10 * (4.7)^2 = 53.122$

$S = \sqrt{53.122}$ = 7.29 *(approx.)* ; df = n-1 = 10-1 = 9

$t = \frac{\bar{d}}{S/\sqrt{n}}$ = 4.7 * $\sqrt{10/7.29}$ = 4.7 * 3.16 / 7.29 = 2.0373

Tabulated value i.e. $t_{0.005, 9} = 3.25$

Thus, 2.0373 < 3.25

Hence, H_0 is accepted and we conclude that there is no significant change in IQ after the training programme.

7.3.2.4 Hypothesis testing for the difference in two population Proportions

This method is used when a businessman or researcher is interested to compare market shares of products in two different regions or proportions of people consuming a particular product or service etc. On the basis of difference in sample proportions a researcher can estimate the difference in population proportions.

Let p_1 and p_2 be the two sample proportions. P_1 and P_2 are the two population proportions, n_1 and n_2 be two sample sizes,

$H_0: P_1 = P_2$

$H_a: P_1 \neq P_2, P_1 < P_2, P_1 > P_2$

$$Z = \frac{p_1 - p_2}{S.E\ of\ (p_1 - p_2)};$$

S.E of $(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Where $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

And q= 1-p

Example: Before an increase in excise duty on tea, 400 people were tea drinkers in a sample of 600 people. Do you think that there is a significant decrease in the consumption of tea after the increase in the excise duty?

Solution: Let P_1 and P_2 be the population proportions of people before and after an increase in excise duty who were found to be tea drinkers.

Let null hypothesis be H_0: $P_1 = P_2$

Then alternative hypothesis is H_1: $P_1 > P_2$

$$z = \frac{p_1 - p_2}{S.E\ of\ (p_1 - p_2)}$$

We have S. E of $(p_1-p_2)= \sqrt{pq(\frac{1}{n_1}+\frac{1}{n_2})}$

Where $p = \frac{n_1p_1+n_2p_2}{n_1+n_2}$

Here, $n_1 = 500$, $p_1 = 400/500 = 0.8$, $n_2 = 600$, $p_2 = 400/600 = 2/3 = 0.67$

p = (500*0.8 + 600*2/3)/500+600 = 800/1100 = 8/11

q = 1 – p = 1 – 8/11 = 3/11

S.E. = √8/11*3/11(1/500+1/600) = √24/121*11/3000 = 0.027

$Z = (p_1 – p)_2$ / S.E. = (0.8 – 0.67) / 0.027 = 0.13/0.027 = 4.8148 > 1.645

Since, calculated value is > tabulated value; thus, we reject the null hypothesis.

Thus we conclude that there is a significant decrease in the consumption of tea after the increase of excise duty.

Self- Assessment Exercise 2:

(1) A new drug manufacturer wants to market a new drug only if he could be quite sure that the mean temperature of a healthy person taking the drug could not rise above 98.6°F otherwise he will withhold the drug. The drug is administered to a random sample of 17 healthy persons. The mean temperature was found to be 98.4°F with a standard deviation of 0.6°F. Assuming that the distribution of the temperature is normal and α= 0.01, what should the manufacturer do?

(2) Two brands of truck tyres are being compared by a transport firm. A random sample of 50 tyres of brand X had an average life of 45000 km, while the average life of a random sample of 40 brand Y tyres was 46500 km. assuming that σ_1 was 2000 km and σ_2 was 1500 km, is there any significant difference in quality at 1% level of significance.

(3) IQ test was administered to 5 persons before and after they were trained. The results are given below:

Candidates	**I**	**II**	**III**	**IV**	**V**
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is any change in IQ after training programme.

(4) Accountants were given intensive coaching and four tests were conducted in a month. The scores of tests 1 and 4 are given below:

Serial no. of accountants	1	2	3	4	5	6	7	8	9	10
Marks in 1st test	50	42	51	42	60	41	70	55	62	38
Marks in 4th test	62	40	61	52	68	51	64	63	72	50

Test whether there is any improvement in the performance of the accountants.

7.4 Summary

Hypothesis is an assumption taken by any researcher. Hypothesis testing is a technique of decision making i.e. to accept or reject a hypothesis on the basis of sample information. Hypothesis testing can be one tailed or two tailed depending upon the requirement of the situation. In one tail tests rejection area lies on one side of the sampling distribution while in two tailed tests, rejection region lies equally on both tails. In case a null hypothesis (H_0) is true but rejected through the procedure, type I error occurs. If H_0 is accepted even when it is false, type II error is committed. There are various applications of the parametric tests conducted in hypothesis testing. Z-test and t-test are used in case of Hypothesis testing of one sample as well as two samples. For a large sample i.e. $n > 30$, Z-test is used. For a small sample, if the population S.D. is known, Z-test is used, while if population S.D. is unknown, t-test is used. However, in both the cases we take the assumption that the sample has been drawn from a normal population. In order to examine the equality of two population means, Z-test is applicable in case of large samples. In case of small sample and when population variances are known, Z-test is applicable. But if population variances are unknown, t-test of means for independent samples is used. For populations that are related on some measure i.e. before and after, matched pair t-test is used.

7.5 Glossary

- **Hypothesis:** A statement or assumption about a parameter of population.
- **Null Hypothesis:** it is the hypothesis of no difference.
- **Alternate Hypothesis:** Statement to be accepted if null hypothesis is rejected.
- **Confidence interval or limit:** Intervals or limits within true value of parameter lies.
- **Confidence Level:** Expressed in percentage, e.g. 95% and indicate the degree of confidence that the true value lies in the specified interval.
- **Power of a test (1-β):** Refers to the probability that the test will lead to rejection of a statement when it is false. Equal to type 1 error.

- **Standard error:** Standard Deviation of an Estimate from a sample.
- **Type I Error (α):** Refers to a test of significance. Probability that the test will lead to rejection of a statement when it is true. Denoted by a Greek letter α
- **Type II Error (β):** Refers to a test of significance. Probability that the test will lead to accepting a statement when it is false. Denoted by a Greek letter β.
- **One–independent sample *z* test:** is a statistical procedure used to test hypotheses concerning the mean in a single population with a known variance.
- **Test statistic: It** is a mathematical formula that allows researchers to determine the likelihood or probability of obtaining sample outcomes if the null hypothesis were true. The value of a test statistic can be used to make inferences concerning the value of population parameters stated in the null hypothesis.

True or False:

1. Type I error is the probability of accepting null hypothesis when it is true.
2. Type II error is more harmful than type II error.
3. For a given level of significance, we can reduce β by increasing the sample size.
4. The value of test statistic that defines the rejection region is called critical region for the test.
5. Accepting a null hypothesis when it is false is called Type II error.
6. Alternative hypotheses specify the value that the researcher believes to hold true.
7. For testing the value of the population mean, a Z- test should be used when the sample size is small and the population standard deviations are unknown.
8. If a hypothesis is rejected at 5% level, it must also be rejected at 1% level.
9. The alternative hypothesis H_1: $\mu \neq 35$ is an example of a two tailed test.
10. A Z-test could be to test population mean when population standard deviation is known, though sample size is small.
11. Whenever the degrees of freedom exceed 30, the t distribution can be approximated y Z distribution.
12. The degrees of freedom in the two sample t test for testing the equality of means is given by n_1+n_2-2.
13. The paired sample t test could be used when on the same respondent two observations are taken, one before the experiment and other after the experiment.

14. The sample standard deviation could be used as an unbiased estimate of the population standard deviation.
15. A paired difference test is appropriate when the two samples being tested are dependent samples.
16. In testing hypotheses about the difference of two means, suppose that the sample sizes are large. If we do not know the actual standard deviations of the two populations, we can use the sample standard deviation as the estimates.
17. When doing a two-tailed test for the difference between means, when a null hypothesis of $\mu_1 = \mu_2$, the hypothesized difference between the two population means is zero.
18. Two-sample tests are used to reach conclusions about the relationships between two populations.
19. If the sample sizes are too small to use the normal distribution for a test of the difference of two population proportions, you should use the t distribution.
20. To compare two population means using small samples, you should always pool the two sample variances.
21. Paired difference tests of means can be based on either the normal or the t distributions, depending on the sample sizes.

    [Answers: T, T, F, F, T, T, F, T, T, T, T, T, T, T, T, F, T, T, F, F, T]

7.6 Suggested Answers to Self -Assessment Exercise:

Self- Assessment Exercise 1:

(1) 0.0548; (2) 4728.2; (3) 0.6; (4) 0.55; (5) 1.264

Self- Assessment Exercise 2:

(1) t = 1.37; (2) Z = 4.06; (3) t = 0.82; (4) t = 19.37

7.7 Suggested Readings:

- Black, K. Business Statistics for Contempory Decision Making, Fifth Edition, Wiley India.
- Bajpai, N., Business Statistics, Pearson Education

- Srivastava, T.N. and Rego,S., Statistics for Management, Fourth Reprint, Tata McGraw Hill Companies.
- Chawala, D. and Saundi, N., Research Methodology Concepts and Cases, Vikas Publishing house pvt ltd.
- Kothari. C.R., . Research Methodology

7.8 Model Questions:

1. A random sample of size 20 drawn from a normal population with standard deviation 5.2 has a mean 16.9. Test at 5% signal level that the population mean is 1.5. Also obtain 99% confidence limits for mean.
2. A sample of 900 items has mean 3.4 and standard deviation 2.61. Can the sample be regarded as drawn from a population with mean 3.25 at 5% level of significance?
3. The means of two random samples sizes of 200 and 300 drawn from two populations having standard deviations 8 and 9 respectively are 46 and 48.5. Test the equality of the means of the two populations. (use 1% level of significance)
4. In a certain district A, 450 persons were considered regular consumers of tea out of a sample of 1000 persons. In another district B, 400 were regular consumers of tea out of a sample of 800 persons. Do these facts reveal a significant difference between the two districts as far as tea-drinking habit is concerned?
5. A machine produced 30 defective articles in a batch of 500. After overhauling it produced 15 defective articles in a batch of 300. Has the machine improved? (take $\alpha = 0.05$)
6. The means of two large samples of sizes 800 and 1600 are 7.5 and 75.0 respectively. Test the equality of means of the two populations each with standard deviations 2.4 at 5% level.
7. In a sample of 600 students in a certain college 400 are found to use Ball-Point pens. In another college from a sample of 900 students 450 were found to use Ball-Point pens. Test whether the two colleges are significantly different with respect to habit of using Ball-Point pens.
8. For a random sample of size 10 from a normal population, the mean is 12.1 and the standard deviation is 3.2. Is it reasonable to suppose that the population mean is 14.57? Test at 5% significance level.
9. IQ test was administered to 5 persons before and after they were trained. The results are given below:

Candidates	1	2	3	4	5
IQ before training	210	220	223	232	225
IQ after training	220	218	225	236	221

Test whether there is any change in IQ after the training programme.

10. The means of two random samples of sizes 500 and 600 drawn from two populations having standard deviations 10 and 12 respectively are 60 and 65. Test the equality of the means of the two populations.

CHAPTER 8: ANALYSIS OF VARIANCE AND NON-PARAMETRIC TESTS

Structure

8.0 Objectives

- To understand the meaning of analysis of variance.
- To know the method of conducting one-way ANOVA.
- To know the method of conducting two-way ANOVA.
- To understand the concept of chi-square test.

- To know the various applications of chi-square test.
- To understand the meaning and application of binomial test.

8.1 Introduction

This chapter deals with analysis of variance (ANOVA) and non-parametric tests like chi-square and binomial test. ANOVA has various applications in the field of economics, psychology, sociology, business and industry for example; to find the difference in average fat content of various fruits, to compare the average output of various plants, etc. Non-parametric tests are a set of statistical tests applicable in case normality assumption is not fulfilled or sample size is relatively small. The tests are known as distribution free tests as they do not require any assumption regarding the shape of population distribution from where the sample is drawn.

8.2 Analysis of Variance (ANOVA)

Analysis of Variance is a technique used to test the equality of three or more population means by comparing the sample variances. It is based on the following assumptions:

1. Each population should have a normal distribution.
2. Populations have equal variances i.e. $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_n^2$.
3. Each sample taken from the population should be randomly drawn and should be independent of each other.

Total variation in the sample data can be due to two reasons:

I. Variance between the samples
II. Variance within the samples

Variance between the samples is due to difference among the sample means and is due to assignable causes, while variance within the samples is the difference due to chance errors.

There are two types of methods used in ANOVA:

- Analysis of Variance for One-Way Classification
- Analysis of Variance for Two-Way Classification

8.2.1 One-Way ANOVA

Let the null hypothesis be H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

Then the alternate hypothesis is H_a: all μ_I (I = 1, 2, 3, . . . , k) are not equal.

Assuming that H_0 is true, the test statistic F = MSB/MSW or MSC/MSE follows F - distribution with degrees of freedom $\nu_1 = k - 1$ and $\nu_2 = N - k$

If the calculated value of F > the table value at αlevel, we reject H_0, otherwise we accept H_0 at αlevel.

A table showing the source of variation, sum of squares, d.f., mean square and mean square ratio is called ANOVA Table.

One-Way ANOVA Table				
Source of Variation	Sum of Squares	d.f.	Mean Squares	Test Statistics
Between	SSB	k - 1	MSB = SSB/(k-1)	F = MSB/MSW -
Within	SSW	N - k	MSW = SSW/(N-k)	
Total	SST	N - 1	-	

Sum of squares of variations between the samples (SSB) = $\sum n_i (X_i \text{ bar} - X \text{ bar})^2$

Where, X_i bar (I = 1, 2, . . . , k) is the sample mean ; n_i is the sample size, X bar is the grand mean of the sample means and k is the number of samples.

Steps:

1. Find mean of each sample i.e. X_1 bar, X_2 bar, X_3 bar, . . . , X_k bar.
2. Find grand sample mean i.e. X double bar = (X_1 bar + X_2 bar + . . . + X_k bar) / k
3. Find the deviations of the means of the different samples from the grand mean i.e. X_1 bar - X bar, X_2 bar - X bar, . . . , X_k bar - X bar.
4. SSB = $n_1(X_1 \text{ bar} - X \text{ bar})^2 + n_2(X_2 \text{ bar} - X \text{ bar})^2 + \ldots + n_k (X_k \text{ bar} - X \text{ bar})^2$

Sum of squares of variation within samples (SSW) = $\sum(X_1 - X_1 \text{ bar})^2 + (X_2 - X \text{ bar})^2 + \ldots + (X_k - X_k \text{ bar})^2$

Steps:

1. Calculate sample mean of all the k samples i.e. X_1 bar, X_2 bar, X_3 bar, . . . , X_k bar_k.
2. Find the deviations of the various items of k samples from the mean values of their respective samples.
3. Square all the deviations and total them.

8.2.1.1 Short-Cut Method

Steps:

1. In this method firstly, T is calculated as $T = \sum X_1 + \sum X_2 + \ldots + \sum X_k$.
2. Correction Factor T^2/N is calculated.
3. Compute $SST = \sum X_1^2 + \sum X_2^2 + \ldots + \sum X_k^2 - T^2/N$.
4. Compute $SSB = (\sum X_1)^2/n_1 + (\sum X_2)^2/n_2 + \ldots + (\sum X_k)^2/n_k - T^2/N$.
5. Compute SSW = SST-SSB.
6. $MSB = SSB/\nu_1$ and $MSW = SSW/\nu_2$.

Example: The following data shows the number of personal loan applications processed per day by four employees of a bank observed over a number of days:

Employee 1	11	14	12	15	13
Employee 2	15	18	13	16	-
Employee 3	13	15	17	12	13
Employee 4	14	18	17	15	-

Find out if the mean number of personal loan applications processed per day is the same for the four employees. Use 5% level of significance.

Solution: This problem requires the application of ANOVA:

Employee 1: Mean (x_1) = (11+14+12+15+13)/5 = 13

Employee 2: Mean (x_2) = (15+18+13+16)/4 = 15.5

Employee 3: Mean (x_3) = (13+15+17+12+13)/5 = 14

Employee 4: Mean (x_4) = (14+18+17+15)/4 = 16

Grand Mean (x double bar) = (5/18)(13) + (4/18)(15.5) + (5/18)(14) + (4/18)(16) = 14.5

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

(The mean number of applications is the same for the four employees.)

H_a: The mean number of applications is not the same for at least two employees.

Variance between the Sample Mean = $5(13-14.5)^2 + 4(15.5-14.5)^2 + 5(14-14.5)^2 + 4(16-14.5)^2$

$$= 11.25 + 4 + 1.25 + 9 = 25.5$$

$$MSB = 25.5 / k\text{-}1 = 25.5 / 3 = 8.50$$

Variance within Samples = $\sum(x_j - \text{x bar}_i)^2 = 49$; where,

$\sum(x_1 - x_1 \text{ bar})^2 = (11-13)^2 + (14-13)^2 + (12-13)^2 + (15-13)^2 + (13-13)^2 = 10$

$\sum(x_2 - x_2 \text{ bar})^2 = (15-15.5)^2 + (18-15.5)^2 + (13-15.5)^2 + (16-15.5)^2 = 13$

$\sum(x_3 - x_3 \text{ bar})^2 = (13-14)^2 + (15-14)^2 + (17-14)^2 + (12-14)^2 + (13-14)^2 = 16$

$\sum(x_4 - x_4 \text{ bar})^2 = (14-16)^2 + (18-16)^2 + (17-16)^2 + (15-16)^2 = 10$

MSW = 49/14 = 3.50

F-Ratio = MSB/MSW = 8.50/3.50 = 2.429

Degrees of Freedom in the numerator of the F-Ratio = $\nu_1 = (k-1) = (4-1) = 3$

Degrees of Freedom in the denominator of the F-Ratio = $\nu_2 = (n-k) = (18-4) = 14$

(Where, n = Total number of elements in the sample)

Tabulated value of F-Ratio = 3.34

ANOVA Table				
Source of Variation	Sum of Squares	d.f.	Mean Sum of Squares	F - Ratio
SSB	25.5	3	8.50	2.429
SSW	63	14	4.50	
SST	88.50	17		

Since the calculated value of F-Ratio (2.429) is less than the tabulated value of F-Ratio (3.34) thus we accept the null hypothesis. Thus, we can conclude that there are no significant differences in the mean number of loan applications processed by the four employees.

Self-Assessment Exercise 1:

1. Citrus clean is a new all-purpose cleaner being test marketed by placing displays in three different locations within various supermarkets. The no. of 12-ounce bottles sold from each location within the supermarket is reported below:

A	18	14	19	17

B	12	18	10	16
C	26	28	30	32

At the .05 significance level, is there a difference in the mean numbers of bottles sold at the three locations? Develop an ANOVA table.

2. The following data are the semester tuition charges (Rs) for a sample of private colleges in various regions of the country. At the .05 significance level, can we conclude there is a difference in the mean tuition rates for the various regions?

Northeast (Rs)	Southeast(Rs)	West(Rs)
10	9	7
11	8	8
12	10	6
10	8	7
12		6

Develop the ANOVA table. What is the value of the test statistic?

3. The following are the number of kilometers/liter which a test driver with three different types of cars has obtained randomly on different occasions.

Car 1	15	14.5	14.8			
Car 2	13	12.5	13.6	13.8	14	
Car 3	12.8	13.2	12.7	12.6	12.9	13

Using a 5% level of significance, perform a one-way ANOVA to examine the hypothesis that the difference in the average mileage in the three types of cars can be attributed to chance.

8.2.2 Two-Way ANOVA

When the data is classified according to two different factors, two-way ANOVA is used. One factor is taken row wise and the other column wise.

In this case SST = SSC + SSR + SSE

Where; SST= total sum of squares

SSC= the sum of squares between the columns

SSR= the sum of squares between the rows

SSE= the sum of squares for the residuals or residual variance or the sum of squares due to error.

Since there are two factors so two different hypotheses are taken and tested individually.

Two-Way ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Columns	SSC	c-1	MSC = SSC/(c-1)	F_1 = MSC/MSE F_2 = MSR/MSE
Between Rows	SSR	r-1	MSR = SSR/(r-1)	
Residual (or Errors)	SSE	(c-1)(r-1)	MSE = SSE/[(c-1)(r-1)]	
Total	SST	N-1	…	…

Example: A company which produces stationary items wants to diversify into the photocopy paper manufacturing business. The company has decided to first test market the in three areas termed as north area, central area and the south area. The company takes a random sample of five salesmen S1, S2, S3, S4 and S5 for this purpose. The sales volume generated by these five salesmen (in thousand rupees) and total sales in different regions is given as follows:

Region	**Salesmen**					
	S1	**S2**	**S3**	**S4**	**S5**	**Total**
North	24	30	26	23	32	135
Central	22	32	27	25	31	137
South	23	28	25	22	32	130
Total	69	90	78	70	95	402

Use ANOVA to examine:

(1) Whether the salesmen significantly differ in performance?

(2) Whether there is a significant difference in terms of sales capacity between the regions?

Take 95% as confidence level for testing the hypothesis.

Solution: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: All salesmen are not equal in terms of sales performance

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: All regions are not equal in sales performance

Region	**S1**	**S2**	**S3**	**S4**	**S5**	**Total**	**Means**
North	24	30	26	23	32	135	27
Central	22	32	27	25	31	137	27.4
South	23	28	25	22	32	130	26
Total	69	90	78	70	95	402	
Means	23	30	26	23.33	31.66		26.3

$SSC = 3\ [(23-26.8)^2 + (30-26.8)^2 + (26-26.8)^2 + (23.33-26.8)^2 + (31.66-26.8)^2] = 183.066$

$SSR = 5\ [(27-26.8)^2 + (27.4-26.8)^2 + (26--26.8)^2] = 5.2$

$SST = [(24-26.8)^2 + (22-26.8)^2 + (23-26.8)^2 + \ldots + (32-26.8)^2] = 200.40$

$SSE = SST - (SSC + SSR) = 200.40 - (183.066 + 5.2) = 12.134$

ANOVA summary table

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Squares**	**F Value**
SSC	183.066	5-1 = 4	MSC = 45.766	F_1 = MSC/MSE = 30.17 F_2 = MSR/MSE = 1.71
SSR	5.2	3-1 = 2	MSR = 2.6	
SSE	12.1333	(5-1)(3-1) = 8	MSE = 1.5166	
SST	200.40	n-1 = 15-1 = 14		

At 95% confidence level, critical value obtained from the F table is $F_{0.05,\ 4,\ 8} = 3.84$ and $F_{0.05,\ 2,\ 8} = 4.46$. The calculated value of F_1 is 30.17, which is greater than the tabulated value and falls in the rejection region. Hence, the null hypothesis is rejected and the alternate hypothesis is accepted. The calculated value of F_2 is 1.71, which is less than the tabulated value and falls in the acceptance region. Hence, the null hypothesis is accepted and alternative hypothesis is rejected. So, this shows that there is a significant difference in the performance of the salesmen in terms of generation of sales. Whereas, there is no significant difference in the capacity of generating sales, for all the three regions.

Self-Assessment Exercise 2:

1. The following table presents the number of the defective pieces produced by three workmen operating in turn on three different machines:

	Machine 1	**Machine 2**	**Machine 3**
Workman 1	27	34	23
Workman 2	29	32	25
Workman 3	22	30	22

   Conduct a two-way ANOVA to test at 5% level of significance, whether:

   (i) The difference among the means obtained for the three workmen are significant.

   (ii) The difference among the means obtained for the three machines are significant.

2. A farmer applies three types of fertilizers on 4 separate plots. The figures on yield per acre are tabulated below:

Fertilizers / Plots	**Yield**				**Total**
	A	**B**	**C**	**D**	
1	6	4	8	6	24
2	7	6	6	9	28
3	8	5	10	9	32
Total	21	15	24	24	84

   Find out if the plots are materially different in fertility, as also, if the three fertilizers make any material difference in yield.

3. For the following data representing the number of units of production per day turned out by workers using five machines, set-up the ANOVA table:

Workers	**Machine Type**			
	A	**B**	**C**	**D**

I	4	-2	7	-4
II	6	0	12	3
III	-6	-4	4	-8
IV	3	-2	6	-7
V	-2	2	9	-1

8.3 Non-Parametric Test

These tests are used in case the assumption that the samples are drawn from a normally distributed population is not met. Thus, they are called distribution free tests. In these tests, nominal or ordinal data are used.

8.4 Chi-Square Test

Chi-square test was given by Karl Pearson in 1900. Chi-square distribution looks like normal distribution which is skewed to the right. It is a continuous distribution which assumes only positive values i.e. from 0 to infinity. The shape of the curve changes with the degrees of freedom. Data in chi-square test is often in terms of frequencies.

Conditions for Chi-Square Test:

- The number of observations should be sufficiently large i.e. $n \geq 50$.
- Expected frequency in any cell should not be less than 5.

The statistics χ^2 is defined as

$$\chi^2 = \sum \{(O_i - E_i)^2 / E_i\} = (O_1 - E_1)^2 / E_1 + (O_2 - E_2)^2 / E_2 + \ldots + (O_n - E_n)^2$$

Where $O_1, O_2, \ldots, O_n$ is a set of observed (i.e. experimental) frequencies and $E_1, E_2, \ldots, E_n$ the corresponding set of expected (i.e. theoretical or hypothetical) frequencies.

8.4.1 Applications of Chi-Square Test

- Test for Goodness of fit
- Test of Independence
- Test of Homogeneity

8.4.1.1 Chi-Square Goodness of fit test is used to analyze probabilities of multinomial distribution trials along a single dimension. It enables us to ascertain whether the known probability distributions such as Binomial, Poisson or Normal distribution fit with an actual sample distribution. The chi-square goodness

of fit test compares the expected, or theoretical, frequencies of categories from a population distribution to the observed, or actual, frequencies from a distribution to determine whether there is a difference between what was expected and what was observed. The formula compares the frequency of the expected values across the distribution. The test loses one degree of freedom because the total number of expected frequencies must equal the number of observed frequencies i.e. the observed total taken from the sample is used as the total for the expected frequencies.

Example: A dice is suspected to be biased. It is rolled 24 times with the following results. Conduct a significant test to see if it is biased.

Outcome	Frequency
1	8
2	4
3	1
4	8
5	3
6	0

Solution: If the dice is unbiased each outcome should appear 4 times. Thus, the expected frequency of each outcome should be 4.

H_0: there is no significant difference between the observed and expected values

H_a: there is a significant difference between the observed and the expected values

Outcome	E	O	O- E	$(O - E)^2$	$(O - E)^2 / E$
1	4	8	4	16	4
2	4	4	0	0	0
3	4	1	-3	9	2.25
4	4	8	4	16	4
5	4	3	-1	1	0.25
6	4	0	-4	16	4

					14.5

$\chi^2 = \sum(O - E)^2/E = 14.5$

Degrees of freedom = 5

Value of $\chi^2_{0.05, 5df} = 11.070$

Since the calculated value is more than the tabulated value, null hypothesis of unbiasedness is rejected and we may conclude that the dice is biased.

8.4.1.2 Chi-Square Test of homogeneity is applicable to the outcome of two or more samples drawn from the same or different populations. H_0: the two samples are homogeneous; H_a: two samples are not homogeneous.

The chi-square contingency test is widely used technique for determining whether there is a statistically significant relationship between two categorical variables i.e. nominal or ordinal variables.

Example: A television company has launched a new product with some advanced features. The company wants to know the opinion of consumers about that product with respect to four characteristics: preferred brand with new features, did not prefer brand with new features, preferred only a few new features and indifferent. The company has divided consumers into three groups – executives/officers, businessmen and private consultants. It has taken a random sample of size 459 and obtained results as follows:

Consumers / Opinion	**Executives / officers**	**Businessmen**	**Private Consultants**	**Total**
Preferred brands with new features	35	25	40	100
Did not prefer brands with new features	30	45	34	109
Preferred only a few new brands	45	50	25	120
Indifferent	25	55	50	130

Total	135	175	149	459

Use χ^2 test of homogeneity and draw inference from the data.

Solution: The null and alternate hypothesis can be stated as:

H_0: Opinion of all the groups is the same about the product with new features

H_a: Opinion of all the groups is not the same about the product with new features

$\chi^2 = \Sigma_i (f_0 - f_e)^2 / f_e$; where i = 1 to n

With degrees of freedom = (number of rows – 1)(number of columns – 1) = (4-1)(3-1) = 6

Here, $\alpha$is taken as 0.05

Expected and Observed Values

Consumers / Opinion	**Executives / officers**	**Businessmen**	**Private Consultants**
Preferred brands with new features	35(29.41)	25(38.13)	40(32.46)
Did not prefer brands with new features	30(32.06)	45(41.56)	34(35.38)
Preferred only a few new brands	45(35.29)	50(45.75)	25(38.95)
Indifferent	25(38.24)	55(49.56)	50(42.20)

Where, expected frequencies are calculated as: f_{e11} = RT*CT / N = 135*100/459 = 29.41

Calculated value of χ^2 :

f_0	f_e	$(f_0 - f_e)^2 / f_e$
35	29.41	1.0617
30	32.06	0.1322
45	35.29	2.6691
25	38.24	4.5814

25	38.13	4.5192
45	41.56	0.2851
50	45.75	0.3944
55	49.56	0.5961
40	32.46	1.7504
34	35.38	0.0540
25	38.95	4.9987
50	42.20	1.4415
$\Sigma f_0 = 459$		$\Sigma(f_0 - f_e)^2 / f_e = 22.48$

$\chi^2 = \Sigma(f_0 - f_e)^2 / f_e = 22.48$

Tabulated value of $\chi^2_{0.05, 6} = 12.59$

Since, calculated value (22.48) is greater than the tabulated value (12.59)

Thus, the null hypothesis is rejected.

Hence, we conclude that the opinions of all the groups are not the same about the product with the new features.

8.4.1.3 Chi-Square Test of Independence is applicable to results of a single sample classified according to any two attributes. Two attributes are presumed to be independent of each other.

H_0: Two attributes are independent of each other

H_a: Two attributes are not independent of each other.

Example: An international airline analysed the data on 200 randomly selected bookings of seats according to the method used for making reservations and the class of travel. The results are presented below:

Reservation Method

Class of Travel	**Travel Agent**	**Internet**	**Toll-free number**
Business class	18	11	9

Economy class	55	65	42

Test at 5% level of significance the airplane's belief that class of travel is unrelated with the method used for reservations.

Solution: H_0: The class of travel and the method of reservation are unrelated

H_a: The class of travel and the method of reservation are related

Expected frequencies = Row Total * Column Total / Grand Total

Expected	**Observed**	**$(O - E)^2/E$**
14	18	16/14 = 1.14
59	55	16/59 = 0.27
14	11	9/14 = 0.64
62	65	9/62 = 0.145
10	9	1/10 = 0.10
41	42	1/41 = 0.024
		$\sum(O - E)^2/E = 2.32$

Degree of freedom = (c-1) (r-1) = 2

Total value of $\chi^2_{0.05, 2df} = 5.991$

Since the calculated value (2.32) is less than the tabulated value (5.991), the class of travel is not related with the method used for reservation.

Example: A behavioral scientist is conducting a survey to determine if the financial benefits, in terms of salary, influence the level of satisfaction of employees, or whether there are other factors such as work environment which are more important than salary in influencing employee satisfaction. A random sample of 300 employees is given a test to determine their level of satisfaction. Their salary levels are also recorded. This information is as follows:

Level of Satisfaction	**Annual Salary (in '000 Rs.)**			
	Up to 5	**5 – 10**	**More than 10**	**Total**

High	10	10	10	30
Medium	50	45	15	110
Low	40	15	5	60
Total	100	70	30	200

At 5% level of significance, determine whether the level of employee satisfaction is influenced by salary level?

Solution: H_0: Annual salary and the level of satisfaction are independent of each other.

H_a: Annual salary and the level of satisfaction are not independent of each other.

Observed	**Expected**	**$(O - E)^2/E$**
10	15	25/15 = 1.667
10	10.5	025/10.5 = 0.024
10	4.5	30.25/4.5 = 6.722
50	55	25/55 = 0.455
45	38.5	42.25/38.5 = 1.097
15	16.5	2.25/16.5 = 0.136
40	30	100/30 = 3.333
15	21	36/21 = 1.714
5	9	16/9 = 1.778
		$\chi^2 = \mathbf{16.93}$

Tabulated value of $\chi^2_{0.05,\ 4df} = 9.49$.

Since the calculated value i.e. 16.93 is more than the tabulated value i.e. 9.49 thus, the null hypothesis is rejected. This indicates that the annual salary and the level of significance are not independent of each other.

Self-Assessment Exercise 3:

1. At the 0.01 level of significance, can we conclude that the following 400 observations follow a Poisson distribution with m = 3

Number of arrivals per hour	0	1	2	3	4	≥ 5
Number of hours	20	57	98	85	78	62

2. A brand manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey in which the country was divided into four geographic regions, a random sampling of 100 consumers in each region was surveyed with the following results:

	Region				
	NE	**NW**	**SE**	**SW**	**Total**
Purchase the brand	40	55	45	50	190
Do not purchase	60	45	55	50	210
Total	**100**	**100**	**100**	**100**	**400**

Develop a table of the observed and the expected frequencies and calculate the value of χ^2.

3. The American Accounting Association classifies accounts receivable as "current", "late" and "not collectible". Industry figures show that 60% of account receivables are current, 30% are late and 10% are not collectible. Massa abs Barr, a law firm has 500 accounts receivables: 320 are current, 120 are late and 60 are not collectible. Are these numbers in agreement with the industry distributions? Use 5% as the level of significance.

4. A social scientist sampled 140 people and classified them according to income level and whether or not they played a state lottery in the last month. The sample information is reported below. Is it reasonable to conclude that playing the lottery is related to income level? Use 0.05 as significance level.

	Income			
	Low	**Middle**	**High**	**Total**
Played	46	28	21	95
Did not play	14	12	19	45

Total	60	40	40	140

(a) State the null hypothesis and the alternate hypothesis.

(b) Determine the value of χ^2.

(c) Interpret the result.

5. A survey was conducted to know the views of different strata of people on new government policy. Four different samples comprised of 100 professional, 120 farmers, 110 businessmen and 150 students were selected to have their opinions.

	Professionals	**Farmers**	**Businessmen**	**Students**	**Total**
In favor	50	70	20	50	190
Against	40	30	60	80	210
Indifferent	10	20	30	20	80
Total	100	120	110	150	480

8.5 The Binomial Test

There are populations whose elementary units consist of only two classes. For example, male and female, smoker and non-smoker, members and non-members, literate and illiterate, single and married, and so on. In such cases, when a sample is drawn, the observations will fall into either one or the other of the two mutually exclusive classes. Although the value of p(and thus the value of q also) may vary from population to population, it is fixed for any one population. A random sample drawn from this population may not contain exactly the same proportion p of cases in one class and proportion q of cases in the other. Such differences between the observed (sample) and the population values arise because of the chance.

The binomial distribution is the distribution of the proportions. The test used here indicates whether it is reasonable to believe that the sample with the observed proportions has been drawn from a population having a specified value of p.

The probability p, of obtaining x elements in one class and n-x elements in the other class is given by

$P = {}^{n}C_{r}\, p^{x} q^{n-x}$

Where p is the proportion of cases expected in one of the categories and q = 1-p. Also,

$\sum_i {}^nC_i\, p^i q^{n-i}$; where i = 0,1,2, . . . x

gives the probability of obtaining the observer value.

The rejection region consists of all values of x for which the probability associated with their occurrence under H_0 is equal to or less than α. Alternatively, statistical table can be used to read the associated probability and H_0 is rejected if this probability is less than α. When the sample size, n, is larger than 25, the binomial distribution is approximated to the normal distribution. Within certain limitations, the sampling distribution of x is approximately normal, with mean np and standard deviation $(npq)^{1/2}$ and, therefore, H_0 may be tested with the test statistic

$Z = (x – np)/(npq)^{1/2}$

Z is approximately normally distributed with mean zero and unit standard deviation. Since binomial distribution is a discrete distribution and normal distribution is a continuous distribution, the approximation becomes more accurate when the correction for continuity is incorporated. Thus, Z can also be written as

$Z = (x \pm 0.5 – np)/(npq)^{1/2}$

Where x+0.5 is used when x< np, and x- 0.5 is used when x> np. The procedure involved can be summarized as below:

(I) Determine the sample size or the number of cases observed.

(II) Determine the frequencies of the observed occurrences in each of the two classes.

(III) If n < 25, small sample formula is applicable to test the hypothesis. If n > 25, Z distribution formula is applicable.

Example: In a study of the effects of managerial stress, a researcher selected 16 managers who were extensively trained to do a specific task through two different methods, A and B. Half of the group was taught method A first, while the other half method B first and then interchanging the training methods, each group being formed randomly. At the end of the training, each manager was tested to complete the task and was classified according to the method (A or B) used to complete the task, as shown in the table below. Test whether the managers used, under stress, the first-learned method to complete the task.

Method used by the Managers to complete the task

	Method Used	
	First- learned	**Second-learned**
No of managers	12	4

Solution: The null hypothesis to be tested here is that no difference exists between the probability (p_1) of using the first- learned method and the probability (p_2) of using the second- learned method under stress, i.e.

H_0: $p_1 = p_2 = ½$

H_a: $p_1 > p_2$.

From the given data, we have n = 16 and x = 4, the smaller frequency, thus, small sample formula will be used. The associated probability is given by:

$P = \sum_i {}^nC_i \, p^i q^{n-i}$; where i = 0, 1, 2, 3, 4

= 0.0384

As $P < 0.05 = \alpha$, we reject the null hypothesis. That is, under stress the managers use the first- learned method to complete the task.

8.6 Summary

Analysis of Variance technique is used to test the equality of more than two population means. As per the technique, the total variations in the dependent variable are composed of two components i.e. one which can be attributed to specific cause and the other may be attributed to chance. In analysis of variance the dependent variable is metric whereas the independent variable is categorical. The assumption in ANOVA is that each sample is drawn from a normal population and each of these populations has an equal variance. One-Way ANOVA is used to analyze the effect of a single variable. In case the effect of two independent variables is to be analysed, two-way ANOVA is used. Non-Parametric Tests are used when the assumption of normality does not hold good. They are used in case of nominal and ordinal data. One of them is chi-square test which is used to ascertain whether the theoretical probability distribution coincides with the empirical sample distribution. This is known as chi-square goodness of fit test. Another application of chi-square test is the test of independence where contingency table for determining the independence of the two variables is used. Moreover, chi-square test of homogeneity is used to determine whether two or more populations are homogeneous with

respect to some characteristics of interest. Binomial test is another non-parametric test which is used to test whether the sample with the observed frequencies has been drawn from a population having a specified value of *p*. In case sample size is less than 25, small sample formula is applicable. But when the sample size is more than 25, binomial distribution approximates the normal distribution, thus, Z test statistics is applicable.

8.7 Glossary

- **Analysis of Variance (ANOVA):** A statistical technique used to test the equality of three or more sample means and thus make inferences as to whether the samples come from populations having the same means.
- **Chi-Square Distribution:** A family of probability distributions, differentiated by their degrees of freedom, used to test a number of different hypotheses about variances, proportions and distributional goodness of fit.
- **Contingency Table:** A table having R rows and C columns. Each row corresponds to a level of one variable, each column to a level of another variable. Entries in the body of the table are the frequencies with which each variable combination occurred.
- **Expected Frequencies:** The frequencies we would expect to see in a contingency table or frequency distribution if the null hypothesis is true.
- **F Distribution:** A family of distributions differentiated by two parameters (df-numerator, df-denominator), used primarily to test hypotheses regarding variances.
- **F Ratio:** A ratio used in the analysis of variance, among other tests, to compare the magnitude of two estimates of the population variance to determine whether the two estimates are approximately equal; in ANOVA, the ratio of between-column variance to within-column variance is used.
- **Goodness of Fit Test:** A statistical test for determining whether there is a significant difference between an observed frequency distribution and the theoretical probability distribution hypothesized to describe the observed distribution.

True or False:

(1) Analysis of variance may be used to test whether the means of more than two populations can be considered equal.

(2) When comparing the variances of two populations, it is convenient to look at the difference in the sample variance, just as we looked at the difference in sample means to make inferences about population means.

(3) When the chi-square distribution is used as a test of independence, the number of degrees of freedom is related to both the number of rows and the number of columns of the contingency table.

(4) Chi-square may be used as a test to decide whether a particular distribution closely approximates a sample from some population. We refer to such tests as goodness of fit tests.

(5) When using a chi-square test, we must ensure an adequate sample size so that we can avoid any tendency for the value of the chi-square statistic to be over-estimated.

(6) Chi-square tests enable us to test whether more than two population proportions can be considered equal.

(7) A 3 X 5 contingency table has 3 columns and 5 rows.

(8) The expected frequency for any cell in a contingency table can be immediately calculated once we know only the row and column totals for that cell.

(9) Sample sizes in analysis of variance need not be equal.

(10) The smaller the value of F statistics, the more we tend to believe there is a difference among the various samples.

(11) Using analysis of variance, it is possible to compare the means of more than two populations simultaneously.

(12) The equality of variances between the sample and within the samples is compared using an F statistic in one way ANOVA.

(13) The degree of freedom corresponding to the total sum of squares equals the number of observations less one

(14) In analysis of variance, the null hypothesis is that the means of all the categories are not equal.

[ans: T, F, T, T, T, T, F, F, T, F, T, T, T, F]

8.8 Suggested Answers for Self-Assessment Exercises:

Self-Assessment Exercise 1:

(1) F = 30.65; (2) F = 25.43; (3) 1.901

Self-Assessment Exercise 2:

(1) $F_1 = 4.96$ $F_2 = 21.28$; (2) $F_1 = 4.76$ $F_2 = 5.14$; (3) $F_1 = 18.39$ $F_2 = 6.58$

Self-Assessment Exercise 3:

(1) $\chi^2 = 4.809$; (2) $\chi^2 = 5.012$; (3) $\chi^2 = 9.33$;

(4) (a) H_0: there is no relation in income and whether the person played the lottery. H_a: there is a relation in income and whether the person played the lottery; (b) $\chi^2 = 6.544$; (c) reject H_0

(5) $\chi^2 = 54.57$

8.9 Suggested Readings:

- Black, K. Business Statistics for Contempory Decision Making, Fifth Edition, Wiley India.
- Bajpai, N., Business Statistics, Pearson Education
- Srivastava, T.N. and Rego,S., Statistics for Management, Fourth Reprint, Tata McGraw Hill Companies.
- Chawala, D. and Saundi, N., Research Methodology Concepts and Cases, Vikas Publishing house pvt ltd.
- Kothari. C.R., . Research Methodology

8.10 Model and Terminal Questions:

(1) What is a χ^2 test? Point out its applications. Under what conditions is this test applicable?

(2) What do you understand by ANOVA? Write its major assumptions.

(3) Three different training programmes were tested for their impact on performance of groups of workers in a large company. After four months, each worker was rated in terms of job performance. The results are summarized in the table below:

Training Programme

Impact on Performance	A	B	C
No Improvement	20	39	31

Moderate Improvement	45	89	36
Great Improvement	35	22	33

Does the data suggest differences among the programmes relating to their impact on the performance of workers?

(4) Is the type of soft drink ordered with pizza at a restaurant independent of the age of the customers? A random poll of 309 customers is taken, resulting in the following contingency table of observed values. Use $\alpha = 0.05$ to determine whether the two variables are independent.

Preferred Beverages					
Age		A	B	C	Total
	21-34	26	95	18	139
	35-55	41	40	20	101
	>55	24	13	32	69
	Total	91	148	70	309

(5) Is the transportation mode used in ship goods independent of type of industry? Suppose the following contingency table represents frequency counts of types of transportation used by the publishing and the computer hardware industries. Analyse the data by using the chi-square test of independence to determine whether type of industry is independent of transportation mode. Let $\alpha = 0.05$.

Industry	**Transportation Mode**			
		Air	Train	Truck
	Publishing	32	12	41
	Computer Hardware	5	6	24

(6) A sample of employees at a large chemical plant was asked to indicate a preference for one of three pension plans. The results are given in the following table. Does it seem that there is a relationship between the pension plan selected and the job classification of the employees? Use 1% level of significance.

Pension plan

Job Class	Plan A	Plan B	Plan C
Supervisor	10	13	29
Clerical	19	80	19
Labour	81	57	22

(7) A book publisher wants to investigate the type of books selected for recreational reading by adult men and women. A random sample provided the following information. At 5% significance level, should we conclude that gender is related or unrelated to the type of book selected?

Gender	**Mystery**	**Romance**	**Self-Help**	**Total**
Men	250	100	190	540
Women	130	170	200	500

(8) The manager of a computer software company wishes to study the number of hours senior executives by type of industry spend at their desktop computers. The manager selected a sample of five executives from each of three industries. At the 0.05 significance level, can she conclude there is a difference in the mean number of hours spent per week by industry?

Banking	**Retail**	**Insurance**
12	8	10
10	8	8
10	6	6
12	8	8
10	10	10

(9) Chapin Manufacturing Company operates 24 hours a day, five days a week. The workers rotate shifts each week. Management is interested in whether there is a difference in the number of units produced when the employees work on various shifts. A sample of five workers is selected and their output recorded on each shift. At the 0.05 significance level, can we conclude there is a difference in the mean production rate by shift or by employee?

	Units Produced		
Employee	**Day**	**Afternoon**	**Night**
A	31	25	35
B	33	26	33
C	28	24	30
D	30	29	28
E	28	26	27

(10) A physician who specializes in weight control has three different diets she recommends. As an experiment, she randomly selected 15 patients and then assigned 5 to each diet. After three weeks the following weight losses, in pounds, were noted. At the 0.05 significance level, can she conclude that there is a difference in the mean amount of weight loss among the three diets?

Plan A	**Plan B**	**Plan C**
5	6	7
7	7	8
4	7	9
5	5	8
4	6	9

(11) A Consumer Organisation wants to know whether there is a difference in the price of a particular toy at three different types of stores. The price of the toy was checked in a sample of five discount stores, five variety stores and five department stores. The results are shown below. Use 5% level of significance.

Discount	**Variety**	**Department**
12	15	19
13	17	17

14	14	16
12	18	20
15	17	19

(12)The number of automobile accidents per week in a certain city was as follows:

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that the accident conditions were the same during the 10–week period?

Chapter 9 Association of Attributes

Structure

9.0 Objectives

- To understand the meaning association of attributes.
- To understand the various notations to be used in association of attributes.
- To understand the various ways of classifying class frequencies to be used in association of attributes.
- To understand the consistency of data.
- To know about independence of attributes.
- To understand the methods of measuring association of attributes.

9.1. Introduction

Data used in any type of analysis can be of two types, i.e. quantitative and qualitative. When we measure the actual magnitude like weight, income, production, amount of rainfall, etc. the characteristics of all these can be measured quantitatively and they are known as variables. On the other hand, certain characteristics like honesty, literacy, deafness, intelligence, etc. are qualitative in nature, these are known as attributes. In such cases we can only measure the presence or absence of a particular characteristic.

While dealing with statistics of attributes, classification of the data is done on the basis of presence or absence of a particular attribute. For example; in order to study the attribute literacy, the classification is done as literate and illiterate. The analysis of such type of data is done with the help of the theory of attributes.

9.2. Notations

Certain symbols are used to represent different classes in the theory of attributes.

The presence of the attribute is represented by capital letters A, B, etc. on the other hand the absence of the attribute is denoted by small Greek letters like; α, β, etc. for example; if 'A' represents literate persons then 'α' would represent illiterate persons. Thus, 'A' and 'α' are said to be complimentary classes. The combination of attributes are also used by grouping together various classes like; AB, Aβ, αB, $\alpha\beta$. The number of observations in different classes is called class frequency and is denoted by enclosing class notation in brackets like; (AB), ($\alpha\beta$), etc.

9.3. Classes and Class Frequency

Different attributes are known as different classes and the number of observations assigned to these classes is termed as class frequencies. The class frequencies are denoted by bracketing the class intervals, for example; (A) represents the frequency of A and (B) represents the frequency of B.

9.3.1. Order of Class and Class Frequencies

The order of a class depends upon the number of attributes specified. The class represented by one attribute is known as the class of the first order, a class with two attributes as the class of the second order and a class represented by 'n' attributes is known as a class of the n^{th} order is a class frequency of first order, (AB) is a class frequency of second order and so on. The total number of observations i.e. N is known as the frequency of the zero order since no attributes are specified. For N attributes the total number of class frequencies of different orders is as follows:

Order	0	1	2	...	r	...	n
No. of Frequencies	1	2n	${}^nC_2\,2^2$	...	${}^nC_r\,2^r$	...	2^n

9.3.2. Relation between Class Frequencies

All the class frequencies of various orders are not independent of each other and any class frequency can always be expressed in terms of class frequencies of higher order. Thus,

$N = (A) + (\alpha) = (B) + (\beta) = (C) + (\gamma)$, etc.

Also,

$(A) = (AB) + (A\beta) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$

$(B) = (A\beta) + (\alpha\beta) = (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$, etc.

The classes of higher order are known as the ultimate classes and their frequencies are the ultimate class frequencies. The total number of classes of the ultimate order is determined by the formula, 2^n where, n = no. of classes. The frequencies of the positive, negative ultimate classes can be known from the following table:

	A	α	**Total**
B	(AB)	(αB)	**(B)**

β	$(A\beta)$	$(\alpha\beta)$	**(β)**
Total	**(A)**	**(α)**	**N**

Example 1: Given the following ultimate class frequencies, find the frequencies of positive class.

$(ABC) = 149$, $(AB\gamma) = 738$, $(A\beta C) = 225$, $(A\beta\gamma) = 1{,}196$

$(\alpha BC) = 204$, $(\alpha B\gamma) = 1{,}762$, $(\alpha\beta C) = 171$, $(\alpha\beta\gamma) = 21{,}842$

Solution: $(A) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) = 2{,}308$

$(B) = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma) = 2{,}853$

$(C) = (ABC) + (A\beta C) + (\alpha\beta C) + (\alpha\beta C) = 749$

$(AB) = (ABC) + (AB\gamma) = 887$

$(AC) = (ABC) + (A\beta C) = 374$

$(BC) = (ABC) + (\alpha BC) = 353$

And $N = [(ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)] = 26{,}287$

9.4. Consistency of Data

Consistency of a set of class frequencies may be defined as the property that none of the frequency is negative. If any of the ultimate class frequencies comes out to be negative, the given data are inconsistent. Thus, the necessary and sufficient condition for the consistency of a set of independent class frequencies is that no ultimate class frequency is negative. For a single attribute A the condition is:

(i) $(A) \geq 0$

(ii) $(\alpha) \geq 0 \Rightarrow A \leq N$

For two attributes A and B the condition is:

(i) $(AB) \geq 0$

(ii) $(A\beta) \geq 0 \Rightarrow AB \leq A$

(iii) $(\alpha B) \geq 0 \Rightarrow AB \leq B$

(iv) $(\alpha\beta) \geq 0 \Rightarrow AB \geq (A) + (B) - N$

Example 2:

Examine the consistency of the following data:

N = 2000, (A) = 1200, (B) = 1000, (AB) = 100

Solution:

As per formula $(\alpha\beta) = N - (A) - (B) + (AB)$

$= 2000 – 1200 – 1000 + 100$

$= -100$

As $(\alpha\beta)$ is negative, so the data is inconsistent.

Example 3:

From the following, find out whether the data are consistent or not:

(i). (A)=150,(B)=100,(AB)=60,N=500

(ii). (A)=150,(B)=100,(AB)=140,N=500

Solution:

(i) We are given (A) = 150,(B) = 100 ,(AB) =60,N=500

Substituting the values in the nine square table

	A	α	Total
B	(AB) 60	(αB) 40	(B) 100
β	(Aβ) 90	($\alpha\beta$) 310	(β) 400
Total	(A) 150	(α) 350	N 500

From the table

$(A\beta) = (A)-(AB) = 150-60=90$

$(\alpha B) = (B)-(AB) = 100-60=40$

$(\alpha\beta) = (\alpha)-(\alpha B) = 350-40 =310$

Since all the ultimate class frequencies are positive, we conclude that the given data are consistent.

(ii) We are given (A) = 100,(B) = 150 ,(AB) =140,N=500

Substituting the values in the nine square table

	A	α	Total
B	(AB) 140	(αB) 10	(B) 150
β	(Aβ) -40	($\alpha\beta$) 390	(β) 350
Total	(A) 100	(α) 400	N 500

From the table

(Aβ) = (A)-(AB) = 100-140= -40

(αB) = (B)-(AB) = 150-140=10

($\alpha\beta$) = (α)-(αB) = 400-10 =390

Thus one of the ultimate class frequencies i.e. (Aβ) is negative; we conclude that the given data are inconsistent.

9.5. Independence of Attributes

Two attributes are said to be independent if there exists no relationship between them. In such a case the proportion of A's amongst B's is same as among β's and the proportion of B's amongst A's is same as α's. If the attributes A and B are independent, the proportion of AB's in the populations is equal to the product of the proportions of A's and B's in the population.

Example 4:

Find if A and B are independent, positively associated or negatively associated in the following cases:

1. N = 1000, (A) = 470, (B) = 620, (AB) = 320
2. (AB) = 256, ($\alpha\beta$) = 768, (Aβ) = 48, ($\alpha\beta$) = 144

Solution:

1. (AB) = 320 and (A)(B) / N = 470*620 / 1000 = 291.40

   Since (AB) > (A)(B) / N , (A) and (B) are positively associated.
2. (A) = (AB) + (Aβ) = 256 + 48 = 304

   (A) = (AB) + (αB) = 56 + 768 = 1024

   N = (AB) + (Aβ) + (αB) +($\alpha\beta$) = 256 + 48 + 768 + 144 = 1216

   Since, (A)(B) / N = 304*1024 / 1216 = 256 and (AB) = (A)(B) / N = 256

Thus, (A) and (B) are independent.

9.6. Association of Attributes

Two attributes A and B are said to be associated if they are not independent.

If, (AB) > (A) (B) / N; this implies both A and B are positively associated.

If, (AB) < (A) (B) / N; this implies both A and B are negatively associated.

9.7. Methods of Association

9.7.1. Proportion Method: If there is no relationship between attributes A and B, it is expected that there is same proportion of As amongst the Bs as well as among the βs.

Example 5:

Out of 3,000 unskilled workers of a factory, 2000 come from rural areas and out of 1200 skilled workers, 300 come from rural areas. Determine the association between skill and residence by the method of proportions.

Solution:

Let A denote skilled workers, αdenote unskilled workers

Let B denote workers from rural areas, βdenote workers from urban areas

We are given:

(B) = 1200, (α) = 3000, (B) = 2000, ($\alpha\beta$) = 2000, (AB) =300

According to the method of proportions, two attributes A and B are said to be independent if:

$\frac{(AB)}{(A)} = \frac{(\alpha\beta)}{(\alpha)}$

In the given case: $\frac{(AB)}{(A)} = \frac{300}{1200} = 0.25$

$\frac{(\alpha\beta)}{(\alpha)} = \frac{2000}{3000} = 0.67$

Since, $\frac{(AB)}{(A)}$ is less than $\frac{(\alpha\beta)}{(\alpha)}$, there is negative association between skill and residence.

9.7.2. Comparison of Observed and Expected Frequencies Method: Actual observations are compared with expected and if they are equal the attributes are said to be independent. If actual observations are more than the expected, the attributes are positively associated and if the actual observation is less than the expected, the attributes are negatively associated.

9.7.3. Yule's Coefficient of Association: This is the most popular method of determining association as it determines the nature as well as degree. It is denoted by Q and its value lies between ± 1.

$$\mathbf{Q = (AB)(\alpha\beta) - (A\beta)(\alpha B) / (AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Example 6:

According to a survey, the following results were obtained:

	Male	Female
No. of candidates appeared at an examination	800	200
Married	150	50
Married and Successful	70	20
Unmarried and successful	550	110

Find the association between Marital status and Success in examination both for Males and Females

Solution:

Let A denote married, α denote unmarried

Let B denote successful, β denote not successful

We will find coefficient of association both for males and females

Males: (A) = 150,(AB) = 70,(αB) = 550 and N= 800

Substituting the values in the nine square table

	A	α	Total
B	(AB) 70	(αB) 550	(B) 620
β	(Aβ) 80	($\alpha\beta$) 100	(β) 180
Total	(A) 150	(α) 650	N 800

From the table

(Aβ) =80 and ($\alpha\beta$) =100

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{(70)(100) - (80)(550)}{(70)(100) + (80)(550)}$$

$$= \frac{7000 - 44000}{7000 + 44000}$$

= -0.725

Females: (A) = 50, (AB) =20, (αB) = 110 and N= 200

Substituting the values in the nine square table

	A	α	Total
B	(AB) 20	(αB) 110	(B) 130
β	(Aβ) 30	($\alpha\beta$) 40	(β) 70
Total	(A) 50	(α) 150	N 200

From the table

(Aβ) =30 and ($\alpha\beta$) =40

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{(20)(40) - (30)(110)}{(20)(40) + (30)(110)}$$

$$= \frac{800 - 3300}{800 + 3300}$$

= -0.61

It is clear from the answer that marital status and success in examination are negatively associated both for males and females.

Example 7:

A professor examined 280 students in Statistics and Accounting and found that 160 failed in Statistics, 140 failed in Accounting and 80 failed in both the subjects. Is there any association between failure in Statistics and Accounting?

Solution:

Let A denotes students who failed in Statistics and B denotes the students who failed in Accounting.

Substituting the values in the nine square table:

	A	α	Total
B	(AB) 80	(αB) 60	(B) 140
β	(Aβ) 80	($\alpha\beta$) 60	(β) 140
Total	(A) 160	(α) 120	N 280

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$= \frac{(80)(60) - (80)(60)}{(80)(60) + (80)(60)}$$

$$= \frac{4800-4800}{4800+4800}$$

$$= 0$$

Since, Yule's Coefficient of Association is 0, there is no association between failure in statistics and Accounting.

9.7.4. Coefficient of Colligation: This is denoted by $\gamma$and helps to find out Yule's coefficient i.e. Q.

$$\gamma = 1 - \sqrt{(A\beta)(\alpha B)/(AB)(\alpha\beta)} \;/\; 1 + \sqrt{(A\beta)(\alpha B)/(AB)(\alpha\beta)}$$

$$Q = 2\gamma/1+\gamma^2$$

9.7.5. Coefficient of Contingency: Contingency Table is a frequency table in which the information is classified according to two or more attributes. The coefficient of contingency is calculated as:

$$C = \sqrt{\chi^2 / N+\chi^2}$$

The value of χ^2 is determined as

$$\chi^2 = (O-E)^2/ E$$

Check your progress:

1) Complete the following table and find Yule's coefficient of association

N=800, (A)=470, (β)=450 and (AB)=230

2) In a class test in which 135 candidates were examined for excellence in Maths and Science, it was discovered that 75 students failed in Maths and 80 failed in science and 50 failed in both. Find if there is any association between failing in Maths and science and also state the magnitude of association.

3) The following summary data relate to the adult population of a small village:

Adult population	600
No. of employed	240
Literate Adult Population Employed	80
No. of literates	200

Determine whether literacy and employment are associated or not.

4) According to a survey, the following results were obtained:

	Male	Female
No. of candidates appeared at an examination	800	200
Married	150	50
Married and Successful	70	20
Unmarried and successful	550	110

Find the association between Marital status and Success in examination both for Males and Females.

5) A professor examined 280 students in Statistics and Accounting and found that 160 failed in Statistics, 140 failed in Accounting and 80 failed in both the subjects. Is there any association between failure in Statistics and Accounting?

9.8 Answer key to Check your progress:

1) 0.253, 2) 0, 3) 0, 4) -0.61, 5) 0

9.9 Terminal and Model Questions:

1. What do you understand by consistency of given data? How do you check it?
2. What do you understand by illusory association?
3. What do you mean by independence of attributes? Give a criterion of independence for attributes A and B.
4. What is association of attributes? How is it measured?

5. When are the two attributes said to be (i) positively associated, and (ii) negatively associated? Also define complete association and disassociation of two attributes.

6. 800 candidates of both genders appeared at an examination. The boys outnumbered the girls by 15% of the total. The number of candidates who passed exceeds the number failed by 480.equal number of boys and girls failed in the examination. Prepare a 2×2 table and find the coefficient of association. Comment.

7. Find whether the attributes α and β are also positively associated, negatively associated or independent, given (AB) = 500, (α) = 800, (B) =600, N=1500.

8. A survey was conducted in respect of marital status and success in examinations. Out of 2000 persons who appeared for an examination, 80% of them were boys, and the rest were girls. Among 300 married boys, 140 were successful,1100 boys were successful among unmarried boys. In respect of 10 married girls 40were successful, 200 unmarried girls were successful. Construct two separate nine square tables and find out the Yule's coefficient of association to discuss the association between marital status and passing of examination.

9. In a series of houses actually invaded by smallpox, 70% of the inhabitants are attacked and 85% have been vaccinated. What is the lowest percentage of the vaccinated that must have been attacked?

10. At a competitive examination at which 600 graduates appeared, boys outnumbered the girls by 96.Those qualifying for interview exceeded in number those failing to qualify by 310.The number of Science graduate boys interviewed was 300 while among Arts graduate girls there were 25 who failed to qualify for the interview. Altogether there were only 135 Arts graduates and 33 among them failed to qualify. Boys who failed to qualify numbered 18.

    Find (i) the number of boys who qualified for interview,

    (i) The total number of science graduate boys appearing, and

    (ii) The number of science graduate girls who qualified.

11. The male population of U.P is 250 lakhs. The number of literate males is 20 lakhs and the total number of criminals is 26 thousand. The number of literate male criminals is 2 thousand. Do you find any association between literacy and criminality?
12. Use proportion method to determine the nature of association between A and B:

	B	β	Total
A	30	50	80
α	20	100	120
Total	50	150	200

Chapter 10: Correlation and Regression Analysis

A king after losing throne in the battle with his relatives was expelled from the state. He along with his wife and daughter were passing through the forest when they were attacked by some dacoits. The king in the process of saving his family lost his life. The queen and daughter fled the scene. In the meantime a father and his son came to the forest. They saw the footprints of the queen and daughter and decided they will marry them depending on the size of their feet. The father decided to marry the lady with bigger feet while son to marry the lady with smaller feet. When they found the ladies, it happened that bigger foot belonged to daughter and smaller to the mother. Why the decision went wrong?

A king's daughter (lets name her Rose) and his courtier's daughter (lets name her Lotus) were very good friends. Once upon a time while playing they entered into a fight and Rose dumped Lotus into a well. When Lotus after the fight reached back home, she complained to her father about Rose's behavior. The courtier got very angry and asked the king to ask Rose to apologize to Lotus. After long deliberation the king asked Rose to act as maid for Lotus. Earlier Lotus was rescued from the well by a handsome young man (lets name him Kamal). Lotus fell in love with him and after the request from her father Kamal agreed to marry her. Rose being maid of Lotus accompanied the couple to their home. After some time Lotus gave birth to a boy (lets name him Elder). During this time without the knowledge of Lotus, Kamal and Rose fell in love with each other and they were also blessed with a boy (lets name him Younger). When Lotus came to know of this she got angry and complained to her father. The father in a fit of anger punished Kamal by making him old and weak. Kamal apologized furiously and also the courtier understood that in this process the real loser would be her daughter as no girl likes her husband to be weak and old. So he told Kamal that he can be young and powerful again if one of his sons' agree to be old and weak in his place. So Kamal firstly asked Elder to take his position which he declined. But Younger agreed to fulfill his fathers' wishes. So Younger became old and weak while his father regained his youth. But after some time Kamal realized his mistake and rectified his mistake by swapping the health conditions with Younger. In gratitude he passed his kingdom to Younger rather than Elder. Why he made his younger son as his successor rather than his elder son? Does the causes of this action were non-fulfillment of wishes by the elder son and otherwise by younger son or Kamal's late realization of his mistake? Which cause played how much role in ensuing an effect? Can this result be used to predict that if Younger faces with similar situation then he would act in a similar fashion? These questions can be answered by studying the concept of regression which helps in understanding the cause and effect relationship between different variables.

10.0 Objectives

To understand

- the techniques of evaluating degree and direction of relationship between two variables.
- the cause and effect linkage.
- the regression line and how to use it in forecasting.
- the goodness of fit of regression model by understanding the meaning of standard error of estimate and coefficient of determination.

10.1 Introduction

Correlation is degree of association between two variables. It indicates the extent to which the variation in the scores on one variable results in a corresponding variation in the scores on the second variable. For example if correlation analysis indicates that marks in Graduate Management Aptitude Test (GMAT) are positively associated with performance in MBA then higher the marks in GMAT higher marks will be scored in MBA.

In this chapter we shall be considering only linear relationship between two variables implying that if two variables are correlated and scores of these variables are plotted on the graph then they will follow a straight line. Such a line is called as *regression line*. A strong correlation indicates that there is only a small amount of error and most of the points lie close to the regression line; a weak correlation indicates that there is a lot of error and the points are more scattered. In the second case it could be concluded that a linear relationship is not a good model for the considered data.

10.2 Measurement of relationship

The importance of correlation being an important measure of finding relationship between two variables is signified by its additional appropriateness than covariance.

The simplest and easiest way to deduce whether two variables are associated is to find whether if one variable deviates from its mean in a particular direction then does other variable also deviates in a similar fashion.

For instance if one variable is X and its mean is X-bar and other variable is Y and its mean is Y-bar then their deviation from respective means is denoted by (X-X bar) and (Y-Y bar). To understand what these deviations infer lets study the following example. The data in table 1 represents five observations of sales and number of advertisements watched.

Table 1

	1	**2**	**3**	**4**	**5**	**Mean**	**S.D.**
No. of advertisements watched (x)	5	4	4	6	8	5.4	1.67
Sales (y)	8	9	10	13	15	11.0	2.92

Covariance (x,y) = $\sum$(x-x bar)*(y-y bar)/n-1

By applying this formula for the above data covariance = 4.25

A positive covariance indicates that deviation of both variables is in the same direction and if one increases then other variable also increases and vice-versa. On the other hand, a negative covariance indicates that if one variable increases then other decreases. But one problem with covariance as a measure of relationship between two variables is that it is not a standardized measure i.e. value changes with change in scale of measurement. For instance, if advertisement expenditure is measured in terms of dollars rather than in rupees then covariance changes.

To rectify this anomaly of lack of standardization, correlation is used. Covariance measure is standardized by dividing the deviations of two variables with their standard deviations. Thus, standardized covariance is known as correlation coefficient (r).

r = covariance (x,y) / {S.D.(x) * S.D.(y)}

$$= \frac{\sum(x-\bar{x})*(y-\bar{y})}{(n-1) * S.D.(x) * S.D.(y)} \qquad = \frac{\sum(x-\bar{x}) * (y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 * (y-\bar{y})^2}}$$

By using this formula r can be calculated as

r = 4.25/1.67 * 2.92 = 0.87.

By standardizing coefficient value has been limited between -1 and +1. Thus, positive sign indicates movement of both variables in the same direction and high magnitude of r indicates high correlation between the variables.

10.3 Types of correlation

Statistically the correlation value ranges between -1 indicating a perfect negative correlation, and +1 indicating a perfect positive correlation. A value of zero indicates no correlation at all. For example, distance travelled by a car is negatively related with petrol remaining in the tank. Productivity is positively related with experience and amount of time spent on work. Whereas, as described in the example above age of mother and daughter has no correlation with foot size.

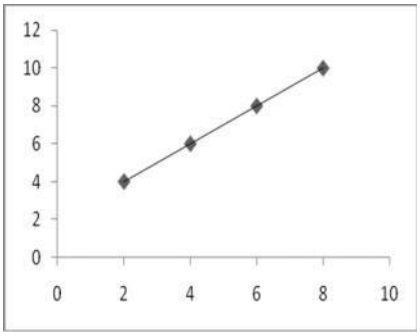
Figure 1 depicts these three different types of correlation by using scatter diagrams. A scatter diagram is simply a graph that plot scores of one variable with the scores of another variable. A scatter diagram tells several things about the data such as whether there exists a relationship between the variables, what kind of relationship it is and whether any cases are distinctly different from the others.

(i) Perfectly positive correlation: A correlation of +1 between two variables indicate a perfect association which implies that if one variable shows increasing or decreasing score then other variable also moves in the same direction.

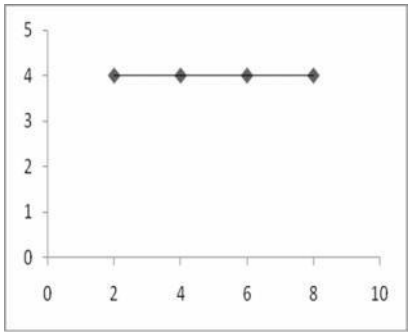
(ii) No correlation: A correlation of 0 between two variables indicate no association.

(iii) Perfectly negative correlation: A correlation of -1 between two variables indicate a perfect association which implies that if one variable shows increasing or decreasing score then other variable moves in the opposite direction.

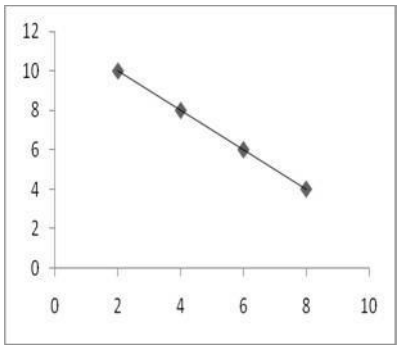
Figure 1



Perfect positive correlation r=+1



No correlation r=0



Perfect negative correlation r=-1

10.4 Pearson coefficient of correlation

The coefficient in equation

r = covariance (x,y) / {S.D.(x) * S.D.(y)}

$$= \frac{\sum(x-\bar{x})*(y-\bar{y})}{(n-1) * S.D.(x) * S.D.(y)} \qquad = \frac{\sum(x-\bar{x}) * (y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 * (y-\bar{y})^2}}$$

is known as Pearson coefficient of correlation. Pearson's correlation relies on a number of assumptions.

- The relationship between the variables is linear.
- The points are evenly distributed along the straight line. This is the assumption of homoscedasticity. If the data has the points unevenly spread along the proposed straight line then the Pearson correlation is not an accurate measure of the association.
- The data are drawn from normally distributed populations.
- The data collected must be interval or ratio.

Example 1: It has been proposed that students who spent maximum time on studying statistics would achieve highest marks. Data for ten students was collected regarding time spent on studying and marks obtained as shown in table 2.

Table 2

Time spent (hrs. per week)	Marks obtained (out of 100)	(x-x bar)	$(x-x\ bar)^2$	(y-y bar)	$(y-y\ bar)^2$	(x-x bar) * (y-y bar)
40	58	11	121	2	4	22
43	73	14	196	17	289	238
18	56	-11	121	0	0	0
10	47	-19	361	-9	81	171
25	58	-4	16	2	4	-8
33	54	4	16	-2	4	-8
27	45	-2	4	-11	121	22
17	32	-12	144	-24	576	288
30	68	1	1	12	144	12
47	69	18	324	13	169	234
x bar=29	**y bar=56**		**1304**		**1392**	**971**

$r = 971 / \sqrt{1304*1392}$

$= 0.72$

These results indicate that as study time increases, statistics exam performance also increases, which is a positive correlation.

Exercise 1:

1. *Explain assumptions of Pearson correlation.*
2. *Determine value of r for the following data*

X	*4*	*6*	*7*	*11*	*14*	*17*	*21*
Y	*18*	*12*	*13*	*8*	*7*	*7*	*4*

10.5 Spearman Rank correlation

In the previous section Pearson correlation was discussed as a technique to measure degree of association between two variables. The primary condition for applying Pearson correlation was the presence of interval data. But when only ordinal data is available then Spearman rank correlation should be applied to achieve the same objective.

The formula for Spearman rank correlation is derived from Pearson correlation coefficient and is as follows:

$r = 1 - \{6\sum d^2/n(n^2 - 1)\}$

where n = number of pairs being correlated

d = difference in rank of each pair

The process of calculating Spearman rank correlation begins by assigning ranks to the ordinal data of two variables under consideration. For one variable the highest value is given rank 1, the next highest is given rank 2 and so on. Similar process is adopted for ranking of second variable. Importantly, the difference in ranks 'd' is found by subtracting the rank of a member of one group from the rank of its corresponding member of the other group. The differences are then squared and summed.

Spearman rank correlation is interpreted similarly as Pearson correlation. Positive correlations indicate that high value of one variable is associated with high value of other variable and negative correlations indicate that high value of one variable is associated with low value of other variable.

Example 2: Is there a correlation between distance travelled by a salesperson and sales achieved. In table 3 sales achieved by nine salesperson in rupees and distance travelled in kms to cover a particular territory is given. Assuming data to be ordinal determine correlation coefficient.

Table 3

Sales (Rs. '000)	Distance (kms)	Rank of sales	Rank of distance	d	d^2
150	1500	9	9	0	0
210	2100	8	8	0	0
285	3200	7	3	4	16
301	2400	6	6	0	0
335	2200	5	7	-2	4
390	2500	4	5	-1	1
400	3300	3	2	1	1
425	3100	2	4	-2	4
440	3600	1	1	0	0
					$\sum d^2 = 26$

Spearman rank coefficient:

$$r = 1 - \{6\sum d^2/n(n^2 - 1)\}$$

$$r = 1 - \{6 * 26/9(81-1)\}$$

$$r = 0.783$$

Positive r indicates sales achieved increases if a salesperson covers more distance per territory. It is not a perfect correlation which implies that other factors than distance travelled plays role in determining sales.

Exercise 2:

1. *What is the difference between interval and ordinal data? Explain by giving example.*
2. *Calculate Spearman rank correlation for following data*

X	*4*	*5*	*8*	*11*	*10*	*7*	*3*	*1*
Y	*6*	*8*	*7*	*10*	*9*	*5*	*2*	*3*

10.6* Regression: *Does money lead to happiness?

Correlation analysis as discussed in previous section indicated the strength of relationship between two variables. But the technique did not tell the direction of relationship. For instance,

- Sales volume is associated with advertisement expenditure. But does increase in advertisement resulted in increase in sales or increase in sales motivated the company to increase expenditure on advertisement.
- Hotel occupancy is related with tourist flow. But does increase in hotels lead to high tourist inflow or more tourists motivates building of more hotels.
- Performance in a job is related with salary. But does increase in salary results in better performance or vice versa?

Whether two variables are associated or not can be evaluated by using correlation analysis but it does not indicate which variable is the cause and which is the effect? The causal relationship can be evaluated by using regression analysis.

Regression analysis is the process of building a model involving two variables that can be used to predict one variable by another variable. The most elementary regression model is called simple regression in which the variable to be predicted is called the dependent variable and variable which predicts is called independent variable. The dependent variable is denoted by 'y' and independent variable is denoted by 'x'. The model is denoted by following equation.

Dependent or predicted variable (y) = intercept (a) + slope (b) * independent or predictor variable (x) .....(1)

10.7 Regression Equation and Coefficients

The first step in regression line is to find whether two variables are associated or not. This can be done by drawing scatter diagram as discussed in previous chapter of correlation. Scatter diagram is the graphical representation of data of two variables. All the data points might not fall in a straight line. For example table 4 shows data regarding advertising expenditure and corresponding sales. The questions are:

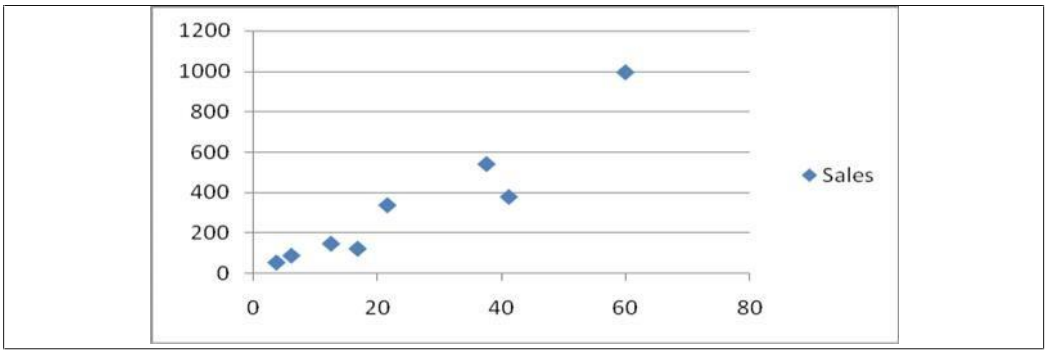
(i) Is there a relationship between two variables?

(ii) Can one variable be used to predict another variable?

Table 4

Advertising expenditure (millions)	12.5	3.7	21.6	60	37.6	6.1	16.8	41.2
Sales	148	55	338	994	541	89	126	379

Fig.2



The relationship between two variables can be deduced by drawing the scatter diagram (Fig.1). It is not a perfect relationship as data points are scattered. But diagram shows a positive relationship between sales and advertisement expenditure indicating that with increase in one variable other variable also increases. But it is important to note here that scatter diagram does not tell causality i.e. whether advertisement expenditure results in sales or is it vice-versa. To understand causality following literature which includes model building and its accuracy has to be appreciated.

For accurate prediction a straight line covering maximum data points should be drawn. Now as data points are scattered so many straight lines can be drawn. This line which touches maximum points and minimizes error is called as regression line. This line as discussed above can be denoted by a model depicted by following equation:

$$y = a + bx$$

where y = dependent variable

a = intercept

b = slope intercept and

x = independent variable.

For instance in the above example if sales is considered as dependent and advertisement expenditure as independent variable then the model would become as:

$$\text{Sales} = a + b * (\text{advertisement expenditure})$$

This type of model is called as deterministic model that produces exact output for a given input. But sales are not a function of advertisement expenditure only. There can be other factors also like distribution, sales people motivation etc. which have an impact on sales. So a more proper model can be as:

$$\text{Sales} = a + b * (\text{advertisement expenditure}) + \text{error}$$

where error denotes other factors than advertisement expenditure which impact sales. This kind of model is called as probabilistic model.

To determine the regression equation values of ‘a’ and ‘b’ has to be determined. This process is termed as least square analysis. This method is used to develop a regression model by producing the minimum sum of squared error values. The error will be minimized if regression line passes through maximum data points. Thus

least squares method is used to find best fit line. A best fit line is one which minimizes the amount of difference between observed data and the line. The data points which does not fall on the line lie either above or below the line. The distance between these points and line is called residual. For some points this residual is positive and for some negative. To avoid cancelling out each other these differences are squared and added. This sum would be least for the best fit line. That's why this method is called as least squares a method of drawing the line. Thus, in this process entire aim would be calculate values of 'a' and 'b'.

Slope intercept 'b' can be calculated by using the following equation:

$b = \sum(x - x\ bar)(y - y\ bar)/\sum(x - x^2)$

y intercept of the regression line can be calculated as:

a = y bar – b * x bar

The regression line depending on which variable is dependent and which variable is independent can give two different results. The following example indicates the importance of establishing causality between two variables.

Example 3:

(a) If sales is dependent variable (y) and advertisement expenditure is independent variable (x) then compute regression equation and also predict sales if advertisement expenditure is 50 millions:

Table 5

Advertisement expenditure (X)	Sales (Y)	x – x bar	y - y bar	(x –x bar)*(y – y bar)	(x – x bar)²
12.5	148	-12.4375	-185.375	2305.602	154.6914
3.7	55	-21.2375	-278.375	5911.989	451.0314
21.6	338	-3.3375	4.625	-15.4359	11.13891
60	994	35.0625	660.625	23163.16	1229.379
37.6	541	12.6625	207.625	2629.052	160.3389
6.1	89	-18.8375	-244.375	4603.414	354.8514
16.8	126	-8.1375	-210.375	1711.927	66.21891
41.2	379	16.2625	45.625	741.9766	264.4689
X bar = 24.93	**Y bar = 333.37**			**Sum = 41051.69**	**Sum = 2692.11**

b = 41.051/2692.11 = 15.24

a = 333.37 – 15.24*24.93 = -46.89

then regression equation becomes y = -46.89 + 15.24 * x

The slope 'b' of the regression equation implies that with every one unit increase in x i.e. advertisement expenditure sales would increase by 15.24.

If expenditure on advertisement is increased to 50 million then sales would be:

y = -46.89 + 15.24 * 50

= 715.61

i.e. sales would be 715 units with increase in advertisement expenditure to 50 million.

(b) If sales is independent variable (x) and advertisement expenditure is dependent variable (y) then compute regression equation and predict expenditure on advertisement if sales were 350 units.

Table 6

Advertisement expenditure (y)	Sales (x)	y – y bar	x - x bar	(x –x bar)*(y – y bar)	(x – x bar)^2
12.5	148	-12.4375	-185.375	2305.602	34363.89
3.7	55	-21.2375	-278.375	5911.989	77492.64
21.6	338	-3.3375	4.625	-15.4359	21.39063
60	994	35.0625	660.625	23163.16	436425.4
37.6	541	12.6625	207.625	2629.052	43108.14
6.1	89	-18.8375	-244.375	4603.414	59719.14
16.8	126	-8.1375	-210.375	1711.927	44257.64
41.2	379	16.2625	45.625	741.9766	2081.641
y bar = 24.93	**x bar = 333.37**			**Sum = 41051.69**	**Sum = 697469.9**

b = 41.051/697469.9 = 0.058

a = 24.93 – 0.058*333.37 = 5.315

then regression equation becomes y = 5.315 + 0.058 * x

The slope 'b' of the regression equation implies that with every one unit increase in x i.e. sales, advertisement expenditure would increase by 0.058.

If sales were 350 units then advertisement expenditure would be:

y = 5.315 + 0.058 * 350

= 25.61

So, by changing dependent and independent variable different regression line can be constructed which could be used for predictive purposes.

Exercise 3:

1. For the following data determine the equation of regression line to predict y from x

X	53	47	41	50	58	62	45	60
y	5	5	7	4	10	12	3	11

2. Slope of regression line represents ……
3. Assessment of regression line needs estimation of constant and slope intercept by using method of ……

10.8 Assessing the accuracy of model: How can it be inferred that proposed model is good?

In the previous section we discussed that an effective model between two variables can be constructed for predictive purposes by using best fit line method. But as shown in equation (4) in every model there would be an error term which should be studied before finalizing the accuracy of the model. This error denotes the difference between observed and expected values. As the analysis is always done on sampled data and decision is interpreted for entire population so a difference between sampled i.e. observed data and expected data for population is expected. In checking the accuracy of model the aim is to assess the impact and amount of error. High and significant error would ask for rectification in the model. This difference between observed 'y' values and expected 'y cap' values given by 'y – y cap' is termed as residual.

For example performance in exam of a student is dependent on the amount of time he/she spends on revising the subject. Presumably, higher the time spent on revising higher he/she will score. To assess this hypothesis a sample of 50 students is taken and data regarding their marks and amount of revision time is recorded. Regression analysis is applied and it was found that this hypothesis was good for not all the students as some students even if spending higher time were not able to improve their performance. Thus, there must be some other factors which cause exam performance. These other factors form part of unexplained variance indicated by residuals, whereas impact of revision time on exam performance is known as explained variance indicated by inclusion of independent variable. Therefore, it becomes necessary to evaluate residuals or unexplained variance to assess the accuracy of model.

The significance of explained and unexplained variance is understood by comprehending the concepts of standard error of estimate and coefficient of determination.

10.8.1 Standard Error of Estimate (s_e)

The standard error of estimate provides a single measure of error which can be used to understand the magnitude of error in the model. Thus, it is used to examine the regression error. So, the first step in estimating accuracy of model is to interpret the error indicated by residuals. As discussed in the previous section residuals denote the difference between observed (y) and expected (y cap) values. The concept is explained by furthering the sales and advertisement expenditure example. This analysis is also indicative of the fact that which of the variable is dependent and which is independent i.e. which of the two regression equations is more representative of the model. The regression equation having less standard error should provide the relationship between dependent and independent variable.

From regression equation analysis by considering sales as dependent and advertisement expenditure as independent variable standard error calculation methodology is depicted in following table.

Example 4:

Table 7

Sales (y)	Advertisement expenditure (x)	y cap	y – y cap	$(y - y\ cap)^2$

148	12.5	143.61	4.39	19.2721
55	3.7	9.498	45.502	2070.432
338	21.6	282.294	55.706	3103.158
994	60	867.51	126.49	15999.72
541	37.6	526.134	14.866	220.998
89	6.1	46.074	42.926	1842.641
126	16.8	209.142	-83.142	6912.592
379	41.2	580.998	-201.998	40803.19
y bar = 333.37	**x bar = 24.93**			**Sum= 70972.01**

The equation of the model is:

Sales (y) = a + b*(advertisement expenditure, x)

From calculations in example 1 value of regression coefficients was found to be

a = -46.89 and b = 15.24

by putting these values the equation becomes

y = -46.89 + 15.24*(x)

for each value of x calculate expected value of y i.e. y-cap. For example,

y cap = -46.89 + 15.24*12.5 = 143.61

Similarly y cap was calculated for all the observations and difference between observed (y) and expected (y cap) was calculated. These values indicate the error or residuals. As the sum of residuals would be approximately equal to zero as positive values would negate the negative error so to avoid this square of residuals is taken and sum calculated.

This total of the residuals squared column is called sum of squares of error (SSE)

Thus, SSE = $\sum(y - y\ cap)^2$ = 70972.01

Now, the question arises how to interpret SSE. A better way of interpreting SSE is through standard error of estimate denoted as s_e. The standard error of estimate is the standard deviation of error. This is calculated to put meaning to the magnitude of error as it is difficult to comprehend whether calculated SSE is high or low. To do this standard error of estimate is calculated by using following formula:

$$s_e = \text{sqrt}(SSE/n-2)$$

where n is number of observations.

For the above example s_e = sqrt(70972.01/8-2)

= 108.75

This standard error of estimate is a good measure to compare the models. The model with a lower s_e would be considered to better fit the data and a better predictor than a model with higher s_e.

10.8.2 Coefficient of determination (R square)

Coefficient of determination depicted as R square is another widely used and effective measure of accuracy of suggested model. It indicates the proportion of variability in the dependent variable (y) explained by the independent variable (x). Higher the value of R square higher is the variability explained. The range of R square is always between 0 and 1. As discussed in previous sections some of variability in the dependent variable could be explained by the introduction of independent variable and remaining goes unexplained. R square is the indicator of explained variation. Thus, higher value of R square is more acceptable. For instance analysis of coefficient of determination can indicate amount of variation in sales because of change in advertisement expenditure or variation in exam performance because of change in revision time.

Meaning and Analysis of R square:

The calculation of R square is explained by using sales and advertisement expenditure example as follows:

Step 1: Initially data was calculated regarding sales of a certain item and it was found there was variability in sales of the item.

Step 2: the researcher wanted to find the reason of sales variability and to create a model where in future sales can be predicted by studying those reasons.

Step 3: Advertisement expenditure was considered to be one of major reasons which causes change in sales volume. The data for advertisement expenditure was collected corresponding to sales and a scatter diagram was plotted. It was found from scatter diagram in addition to information given by correlation analysis that two variables were positively related.

Step 4: Now, for prediction purposes a model was formulated wherein sales were considered to be dependent on advertisement expenditure. By using this model a regression equation was devised with the help of least squares method. This method was selected to find the best fit regression equation with minimum of error.

Step 5: But entire variation in sales data was not explained by advertisement expenditure. The difference between observed and expected values of sales was termed as residuals or unexplained variation. The aim was to have more understanding of variation in dependent variable for more accurate prediction.

Step 6: The variation of dependent variable, in this case, sales data from its mean value is considered to be total variation. This variation is called as Total Sum of Squares denoted as SST and given by following equation:

SST = y – y bar

To explain this variation an independent variable, in this case, advertisement expenditure was introduced. The amount of variation explained as discussed in previous sections is because of regression line. This explained variation because of regression line is denoted by *y cap.* The difference between y cap and mean value is called as Sum of Squares of Regression denoted as SSR and given by following equation:

SSR = y cap – y bar

The remaining variation which goes unexplained is the residual or called as Sum of Squares of Error denoted as SSE and given by:

SSE = y – y cap

Thus complete equation becomes

SST = SSR + SSE

y – y bar = (y cap – y bar) + (y – y cap)

As understood by the definition, coefficient of determination is an indicator of how much of *y* is explanatory and how much of it goes unexplained. So R square represents the amount of variance in the outcome explained by the regression equation relative to how much variation was there to explain in the first place. Therefore, as a percentage it represents the percentage of the variation in the outcome that can be explained by the model:

R square = SSR/SST

or R square = (SST-SSE)/SST

= 1 – (SSE/SST)

From the above equation it is clear that value of R square can never be greater than 1.

Example 5:

Table 8

Sales (y)	Advertisement expenditure (x)	y cap	y – y cap	SSE=(y – y cap)^2	y – y bar	SST = (y – y bar)^2
148	12.5	143.61	4.39	19.2721	-185.75	34503.06
55	3.7	9.498	45.502	2070.432	-278.75	77701.56
338	21.6	282.294	55.706	3103.158	4.25	18.0625
994	60	867.51	126.49	15999.72	660.25	435930.1
541	37.6	526.134	14.866	220.998	207.25	42952.56
89	6.1	46.074	42.926	1842.641	-244.75	59902.56
126	16.8	209.142	-83.142	6912.592	-207.75	43160.06
379	41.2	580.998	-201.998	40803.19	45.25	2047.563
y bar = 333.37	**x bar = 24.93**			**Sum= 70972.01**		**Sum = 696215.5**

R square = 1 – (SSE/SST)

= 1 – (70972.01/696215.5) = 0.898

Interpretation: R square value of 0.898 indicates that 89.8% of variation in sales is explained by advertisement expenditure. This implies accuracy of suggested model. Remaining variation goes unexplained and is depicted as residuals by sum of square of errors.

Exercise 4 ***(True/False)***

1. Standard error of estimate represents the standard deviation of error of regression models.
2. Standard error estimate of a model can be used to interpret the accuracy of a model in isolation.

3. Value of Coefficient of determination can be greater than one.

10.9 Correlation vs. Regression

Understanding of the two concepts facilitates us to infer the similarities and differences between correlation and regression. The major similarity between two statistical tools is that both are measures of association. Higher value of correlation coefficient (r) and higher value of coefficient of determination (R square) indicates strong relationship between two variables. But correlation analysis suffers with two shortcomings. One, it does not tell which variable is the cause and which variable is the effect. Does time spend on revising a subject leads to better performance in exam or better exam performance motivates the candidate to study and revise more? Second, correlation does not have a predictive power. It can be used to link performance with level of stress but it cannot be used to infer that with a specific amount of stress how an employee will perform?

These limitations are removed by regression analysis. The most important procedure in regression analysis is formulation of model having accurate or near accurate predictive power. The steps involved and methodology has been discussed in detail in previous sections.

Coefficient of determination R square is related to coefficient of correlation r. in case of simple linear regression i.e. if there is only one predictor and it is linearly associated with the dependent variable then coefficient of correlation r is square root of coefficient of determination.

coefficient of correlation (r) = square root {coefficient of determination (R square)}

10.10 Summary

There are several different measures to evaluate relationship between two variables. This chapter has only discussed Pearson and Spearman rank correlation. It is necessary to understand that relevant assumptions regarding type of data and its distribution should be checked before selecting a particular technique to find association of two variables. Pearson correlation should be used only for interval data following a normal distribution whereas Spearman rank correlation is a non-parametric test to find association where data is ordinal rather than interval. Both techniques are interpreted in similar fashion as their correlation coefficients lie between +1 and -1. Positive correlation means that as value of one variable increases that of other also increases. Negative correlation means as value of one variable decreases that of other tends to increase. For r values near zero little or no correlation is present.

Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable can be predicted from the other. This chapter discusses bivariate (two variables) regression where there is one independent variable *x* used to predict another dependent variable *y*. The regression model involves building of a regression equation which consists of slope of line as a coefficient of x and a y intercept value as a constant. The accuracy of regression model can be evaluated by using certain statistics.

Standard error of estimate and coefficient of determination are two such statistics which are discussed in this chapter. These statistics involves evaluating observed and expected values of dependent variable. The standard error of estimate is the standard deviation of the error of model. This value of standard error of estimate can be used to analyze magnitude of error instead of studying each residual value. The coefficient of determination is the proportion of total variance of the y variable predicted by x. this value ranges from 0 to 1. High value of coefficient of determination can be interpreted as a significant indicator of accurate model. Lastly, relationship between correlation and regression analysis has been discussed. In the case of bivariate regression correlation coefficient is square root of coefficient of determination.

10.11 Glossary

- **Correlation coefficient:** is a statistic given by Karl Pearson to measure the linear relationship between two variables. its value ranges from -1 to +1 where -1 indicates perfect negative correlation, +1 indicates perfect positive correlation and value of zero implies no relationship between two variables.
- **Spearman's rank correlation:** is a measure of correlation between two ordinal variables.
- **Regression analysis:** is the process of constructing a model involving two variables which can be used to predict dependent variable from independent variable.
- **Least squares method:** is used to create best fit line called as regression line with least error. This method estimates the value of y intercept and slope intercept to create most accurate regression line which is used for prediction of dependent variable.
- **Standard error of estimate:** is the standard deviation of error of regression models which can be used to compare the efficacy of regression models.
- **Coefficient of determination:** represents the proportion of dependent variable explained by independent variable
- **Sum of squares of error:** is the sum of residuals squared in a regression model
- **Total sum of squares (SST):** is the sum of squared deviations about the mean of a set of values

10.12 Answers to check your progress/ Self assessment exercise

Exercise 1:

2. correlation coefficient r = -0.926 implying that with increase in one variable other tends to decrease.

Exercise 2:

2. Spearman rank correlation r = 0.702

Exercise 3:

1. y = -11.335 + 0.355x
2. with increase in a unit of independent variable dependent variable increases by amount of slope of line.

3. Least squares method

Exercise 4:

1. True
2. False
3. False

10.13 References/ Suggested Readings:

- Black, K., *Business Statistics For Contemporary Decision Making,* Fifth Edition, Wiley India,
- Keller, G., *Statistics for Management,* First India Reprint 2009, Cengage Learning India Private Limited.
- Donald R. Cooper & Pamela S. Schindler, *Business Research Methods,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 9th Edition.
- S.P. Gupta, *Business Statistics*, Sultan Chand, New Delhi.

10.14 Terminal and Model Questions

1. Compute the Pearson correlation coefficient to determine the strength of correlation between inflation rate and return on government securities. The data is given in Table 4.

Table 4

Inflation rate (%)	1.57	2.23	2.17	4.53	7.25	9.25	5.00	4.62
Return (%)	3.05	3.93	4.68	6.57	8.27	12.01	10.27	8.45

1. A statistics student asked seven economics students to report their grades in the mathematics and economics course. The results (where 1=F, 2=D, 3=C, 4=B and 5=A) are as shown in Table 5:

Table 5

Mathematics	4	2	5	4	2	2	1
Economics	5	2	3	5	3	3	2

Compute the Spearman rank correlation coefficient.

2. Explain the significance of regression analysis. How it is different from correlation analysis.
3. What do you mean by residuals? Explain the role of analysis of residuals in formulating an accurate regression model.
4. How slope and y intercept in a regression equation can be estimated by method of least squares?
5. The members of a health spa pay annual membership dues of $300 plus a charge of $2 for each visit to the spa. Let y denote the total dollar cost for the year for a member and x the number of visits by the member during the year. Express the relation between x and y mathematically.
6. The director of admissions of a college administered a newly designed entrance test to 20 students selected at random from the newly admitted first year students in a study to determine whether a

students' grade point average (GPA) at the end of first year (y) can be predicted from the entrance test score (x). the results of the study are shown as follows.

S.No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X	5.5	4.8	4.7	3.9	4.5	6.2	6.0	5.2	4.7	4.3
Y	3.1	2.3	3.0	1.9	2.5	3.7	3.4	2.6	2.8	1.6
S.No.	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
X	4.9	5.4	5.0	6.3	4.6	4.3	5.0	5.9	4.1	4.7
Y	2.0	2.9	2.3	3.2	1.8	1.4	2.0	3.8	2.2	1.5

Calculate:

(i) correlation coefficient between two variables.

(ii) regression coefficients and create regression equation.

(iii) Standard error estimate and interpret it.

(iv) Coefficient of determination and interpret it

(v) Relationship between correlation coefficient and coefficient of determination.

CHAPTER 11: INDEX NUMBERS

Structure

11.0 Objectives

- To understand the meaning and uses of index numbers.
- To understand the various methods of constructing index numbers.
- To know various tests of consistency of index numbers.
- To know the concept of base shifting.
- To understand the concept of deflating .
- To understand the concept of splicing.
- To identify the precautions in construction of index numbers.
- To understand the limitations of index numbers.

11.1 Introduction

The value of money changes from time to time. Prices of some commodities may increase and some may decrease over a period of time. As a manager of a business enterprise, the person may be interested to know the changes from one period to another, may it be the cost of raw material, labour, advertising, profit etc. Thus it becomes essential to define an average measure which can compare such differences in prices from one period to another.

An index number is a statistical measure which helps to show such changes. It is a relative measure which describes the average changes in any quantity or prices or values over a period of time in relation to its value at some fixed point in time known as Base period. In other words, it is the ratio of a measure taken for one time period to the same measure taken for another time period. The ratio is multiplied by 100.

$$\text{Index Number} = \frac{\text{Current Period Value}}{\text{Base Period Value}} \text{x } 100$$

Thus the Index number is a number that expresses the relative change in price, quantity or value as compared to a base period.

11.2 Types of Index Numbers

There are three types of Index numbers as discussed below:

- ***Price Index Number***

This show the relative changes in the level of prices over a period of time. e.g. prices of commodity, share prices etc.

- ***Quantity Index Number***

They help to measure the changes in quantity of volume of goods produced or sold e.g. Industrial products in Agriculture production etc.

- ***Value Index Numbers***

These are used to compare changes in the money value of transactions. We calculate value by multiplying price with the quantity e.g. National income, Gross domestic product, etc.

11.3 Methods of Constructing Index Number

11.3.1 Price Index Methods

There are two broad categories of methods to calculate price Index:

11.3.1.1 Simple or Unweighted method

11.3.1.2 Weighted Method

In case of simple method, Index numbers are constructed without assigning weights to different items while in case of weighted method, index numbers are constructed after assigning weights to different items as per their relative importance. In general, quantity is termed as weights. Hence simple method keeps quantity as constant for two periods whereas weighted method allows a change in quantity.

11.3.1.1 Simple or Unweighted Method has 2 types:

- Simple Aggregate Method
- Simple Average of Relatives Method

Simple Aggregate Method

Price Index (P_{01}) is calculated by expressing the aggregate of prices of all the items in the current year as a percentage of the aggregate of prices in the base year.

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

Where, p_1 is price of selected item in the current year and p_0 is the price of the selected item in the base year.

$\sum p_1$ = Aggregate of prices of all the selected items in the current year.

$\sum p_0$ = Aggregate of prices of all the selected items in the base year.

Example: Calculate price index for 2013 using 2011 as the base period by Simple Aggregate Method.

Commodity	Price in 2011(Rs.)	Price in 2013(Rs.)
A	30	40
B	25	28
C	90	108
D	15	24
E	96	120

Solution:

Commodity	Price in 2011 (Rs.)	Price in 2013 (Rs.)
A	30	40
B	25	28
C	90	108
D	15	24
E	96	120
Total	256	320

The price index using simple aggregate method is given by

$$P_{01} = \frac{\sum p_1}{\sum p_0} \times 100$$

$$P_{01} = \frac{320}{256} \times 100$$

= 125

Simple Average of Relatives Method: The simple average of relatives is computed by following steps.

1. Calculate price relative for each selected commodity by using formula:

$$P = \frac{p_1}{p_0} \times 100$$

2. Calculate Price Index for the composite group by taking average of the price relatives.

$$P_{01}(\text{A.M.}) = \frac{1}{n}\sum p = \frac{1}{n}\sum\left(\frac{p_1}{p_0} \times 100\right)$$

Where, n is the total number of commodities

Example: Compute Simple Average of price relative Index from the following data.

Item	1995 Price ($)	2005 Price ($)
Bread, cost per pound	0.77	0.89
Eggs per dozan	1.85	1.84
Milk per gallon	0.88	1.01
Apples per pound	1.46	1.56
Orange Juice, 12 0z	1.58	1.70
Coffee, per pound	4.40	4.62
Total	10.94	11.62

Solution:

The simple Index for bread is 115.6 found by using the formula given in step 1

= (0.89/0.77)*100 = 115.6

Simple index for the other items have been calculated in the same way.

Item	**Simple Index**

Bread, cost per pound	115.6
Eggs per dozan	99.5
Milk per gallon	114.8
Apples per pound	106.8
Orange Juice 12 0z	107.6
Coffee per pound	105.0

The Simple average of price relatives is calculated using formula given in step 2.

= (115.6+99.5+114.8+106.8 +107.6+105.0)/6 = 649.3/6 = 108.2

This indicates that the mean of the group of indexes increased by 8.2% from 1995 to 2005. A positive feature of the simple average of price indexes is that the same value for the index is obtained regardless of the units of measurement.

11.3.1.2 Weighted Price Index Method

In this method each item is assigned a weight according to its importance. Generally, the importance of each item is measured by the amount of quantity.

$$P_{01} = \frac{\sum p_1 w}{\sum p_0 w} x\,100$$

Where, w is the weight given to the item. Weights used are either quantity weights or value weights.

There are two price indexes in weighted method:

- Weighted Aggregate Price Index
- Weighted Average of price relative method
    - **Weighted Aggregate Method**

There are various methods for construction of Index numbers having their own criterions. These have been discussed below:

- **Laspeyres's Price Index(Ls):** Laspeyres's Price Index is a weighted aggregate price index in which base period quantities are considered for weighing price of each commodity for both base period as well as current period.

$$L_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Where, p_0 is the price in the base year, p_1 is the price in the current year, and q_0 is the quantity in the base year.

- **Paasche's Price Index(P_S):** The Paasche's Price Index is a weighted aggregate price index computed by using the quantities for the current year.

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Where, p_0 is the price in the base year, p_1 is the price in the current year, and q_1 is the quantity in the current year.

- ***Dorbish Bowley's Price Index:*** This method is basically the arithmetic mean of Laspeyre's and Paasche's method.

$$\text{(D\&B) } P_{01} = \frac{1}{2}\left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1}\right] \times 100$$

Or $P_{01} = 1/2[\text{Ls} + P_S] * 100$

Where, p_0 is the price in the base year, p_1 is the price in the current year, q_0 is the quantity in the base year and q_1 is the quantity in the current year.

- ***Marshall-Edgeworth's Price Index:*** In this method, the weight for each item is taken as the A.M. of the quantities consumed in the base period and the current period.

$$\text{M E p}_{01} = \frac{\sum p_1 q_0 + \sum p_1 q_1}{\sum p_0 q_0 + \sum p_0 q_1} \times 100$$

Where, p_0 is the price in the base year, p_1 is the price in the current year, q_0 is the quantity in the base year and q_1 is the quantity in the current year.

- ***Fisher's Ideal Price Index:*** This method is basically the geometric mean of the Laspereye's and Paasche's method

$$F\,p_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= (L_S * P_S)^{1/2} \times 100$$

Where, p_0 is the price in the base year, p_1 is the price in the current year, q_0 is the quantity in the base year and q_1 is the quantity in the current year.

This is known as an ideal method as it is based on geometric mean (considered as the best average). Moreover, it takes into account both base year and current year quantities as weights. Fisher's Index number satisfies Time reversal and Factor reversal test required for Index number.

- ***Walch's Price Index:*** In this method, the quantity weight used is the geometric mean of the base and current year quantities:

$$W\,p_{01} = \frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$$

Where, p_0 price in the base year, p_1 is the price in the current year, q_0 is the quantity in the base year and q_1 is the quantity in the current year.

- ***Kelly's Price Index:*** According to this method, ratio of the aggregates with selected weights is used to calculate index values with selected year as base year.

$$K\,p_{01} = \frac{\sum p_1 q}{\sum p_0 q} \times 100$$

Where, p_0 is the price in the base year, p_1 is the price in the current year and q = fixed weight

Example: Find Fisher's, Marshal-Edgeworth, Laspeyres's and Paasche's index number for the following data:

Commodity	2009		2010	
	Price (Rs)	Qty. (kg)	Price (Rs)	Qty. (Kg)
A	20	8	40	9
B	50	10	60	10
C	40	15	60	12

D	10	15	151	15

Solution:

Commodity	2009		2010		p_0q_0	p_0q_1	p_1q_0	p_1q_1
	p_0	q_0	p_1	q_1				
A	20	8	40	9	160	180	320	360
B	50	10	60	10	500	500	600	600
C	40	15	60	12	600	480	900	720
D	10	15	15	15	150	150	225	225
Total					1410	1310	2045	1905

Using Fisher's formula

$$F\,p_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \text{ x } \frac{\sum p_1q_1}{\sum p_0q_1}} \text{ x } 100$$

$$F\,p_{01} = \sqrt{\frac{2045}{1410} \text{ x } \frac{1905}{1310}} \text{ x } 100 = 145.23$$

Using Marshall-Edgeworth's formula

$$M\,E\,p_{01} = \frac{\sum p_1q_0 + \sum p_1q_1}{\sum p_0q_0 + \sum p_0q_1} \text{ x } 100$$

$$= \frac{2045 + 1905}{1410 + 1310} \text{ x } 100$$

= 145.22

Using Laspeyres's formula

$$L_{01} = \frac{\sum p_1q_0}{\sum p_0q_0} \text{ x } 100$$

$$L_{01} = \frac{2045}{1410} \text{ x } 100$$

=145.03

Using Paasche's formula

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \text{ x } 100$$

$$P_{01} = \frac{1905}{1310} \text{x } 100$$

=145.42

- **Weighted Average of Price Relatives Method:**

In this method appropriate weights are assigned to the price relatives according to the relative importance of different commodities in the given group of commodities. The price index is obtained by taking the weighted average of the price relatives. Price index using the weighted arithmetic mean of price relatives is given by:

$$P_{01} = \frac{\sum WP}{\sum W}$$

Where $P = p_1/p_0 * 100$

W (weight assigned to the price relative P)$= p_0 q_0$

Example: The price quotations for five different commodities for the years 2000 and 2005 are given below:

Commodity	Weights	Price	
		2000	2005
A	30	100	90
B	15	20	20
C	20	70	60
D	10	20	15

E	25	40	55

Calculate the price index number for 2005 with 2000 as base year using Weighted A.M. of price relatives.

Solution:

Commodity	Weight	Price		Price Relative	WP
		2000	2005		
A	30	100	90	90/100*100=100	2700
B	15	20	20	20/20*100=100	1500
C	20	70	60	60/70*100=85.71	1714.2
D	10	20	15	15/20*100=75	750
E	25	40	55	55/40*100=137.5	3437.5
	=100				=10101.7

Price index number based on the weighted arithmetic mean of price relatives is given by

P_{01}=10101.7/100=101.017

11.3.2. Quantity Index Number

A quantity index measures the percentage change in production or consumption of an individual commodity or a group of commodities from one time period to another. Here, prices are taken as constant over time to isolate the effect of quantity index as it eliminates the effects of fluctuating prices.

Tip: You can easily calculate these numbers by replacing (p) with (q) and (q) with (p) in price index formulas too.

Laspeyres's Quantity Index i.e., $Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$

Paasche's Quantity Index i.e., $Q_P = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$

Fischer's Quantity Index i.e., $E_p = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_1} x \frac{\sum q_1 p_1}{\sum q_0 p_1}} x\ 100$

11.3.3 Value Index Numbers

Value Index Numbers are designed to compare changes in the money values of the transaction, in two periods of time. The value of commodity is calculated by multiplying its price and quantity.

Value Index, $V = \frac{\sum q_1 p_1}{\sum q_0 p_0} x\ 100$

or $V = \frac{\sum V_1}{\sum V_0}$

Where

V_0 = Value at the base year

V_1 = Value at the current year

Such indexes are not weighted because they take into account both the price and quantity

Example: The price of a commodity increased by 40% between 2009 and 2013 while its production decreased by 20% during the same period. By what percentage did the value of production of the commodity change with respect to its value 2009?

Solution: It is given that

$p_1/p_0 * 100 = 140$

$p_1/p_0 = 1.4$ and

$q_1/q_0 * 100 = 80$

$q_1/q_0 = 0.8$

The value index of the commodity (V) = $p_1q_1/p_0q_0 * 100 = 1.4*0.8*100=112$

The value of the commodity increased by 12% in 2013 with respect to its value in 2009.

11.4 Tests of Consistency

In order to know the most suitable method under given situation, certain tests have been suggested. These are as follows:

11.4.1 Time reversal test

This test was suggested by Prof. Fisher. This is used to test whether a given method will work both backwards and forwards with respect to time. In simple words a price or quantity index for a given period with respect to the preceding period is equal to the reciprocal of the price or quantity index when periods are interchanged.

$P_{01} \times P_{10} = 1$ and $Q_{01} \times Q_{10} = 1$

$$\sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} * \frac{\sum p_1 q_1}{\sum p_0 q_1} * \frac{\sum p_0 q_1}{\sum p_1 q_1} * \frac{\sum p_0 q_0}{\sum p_1 q_0}} = 1$$

Time Reversal Test is satisfied by all the methods except Laspreye's and Paasche's.

11.4.2 Factor Reversal Test

This test was suggested by Prof Fisher. According to this, product of price Index (P_{01}) and Quantity Index (Q_{01}) give the true value ratio. In other words the change in the price when multiplied by the change in quantity should represent the total change in value.

Therefore, $P_{01} \times Q_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$

Where

$\sum p_1 q_1$ is total value in the current year.

$\sum p_0 q_0$ is total value in the base year.

This test is satisfied only by Fisher's Index number method.

Example: Calculate Fisher's Ideal index from the following data and prove that it satisfies both Time Reversal Test and Factor Reversal Test.

Items	Base year		Current Year	
	Price	Quantity	Price	Quantity
A	6	50	10	60
B	2	100	2	120

C	4	60	6	60

Solution: Computation of Fisher Ideal Index

Items	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	6	50	10	60	300	360	500	600
B	2	100	2	120	200	240	200	240
C	4	60	6	60	240	240	360	360
Total					740	840	1060	1200

Solution: Using Fisher's formula

$$F\,p_{01} = \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} * \frac{\sum p_1q_1}{\sum p_0q_1}} * 100$$

$$= \sqrt{\frac{1060}{740} * \frac{1200}{840}} * 100$$

=143.05

Time Reversal Test: $P_{01} \text{ x } P_{10} = 1$

$$\sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} * \frac{\sum p_1q_1}{\sum p_0q_1} * \frac{\sum p_0q_1}{\sum p_1q_1} * \frac{\sum p_0q_0}{\sum p_1q_0}} = 1$$

$$\sqrt{\frac{1060}{740} * \frac{1200}{840} * \frac{840}{1200} * \frac{740}{1060}} = 1$$

Factor Reversal Test: $P_{01} \text{ x } Q_{01} = \frac{\sum p_1q_1}{\sum p_0q_0}$

$$Q_{01} = \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} * \frac{\sum q_1p_1}{\sum q_0p_1}}$$

$$P_{01} \times Q_{01} = \sqrt{\frac{1060}{740} * \frac{1200}{840} * \frac{840}{740} * \frac{1200}{1060}}$$

$$= \sqrt{\frac{1200}{740} * \frac{1200}{740}} = 1200/740 = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Hence its proved that Fisher's ideal index satisfies both Time Reversal Test and Factor Reversal Test.

11.4.3 Circular Test

This test is an extension of the time reversal test for more than two periods.

$P_{01} \times P_{12} \times P_{20} = 1$

This test is satisfied by simple aggregate method, Kelley's method and simple geometric mean of price relatives.

Self-Assessment Exercise:

1. From the following data calculate price index numbers for 2007 with 2006 as base by (a) Laspeyre's Method, (b) Paasche's method, (c)Marshall-Edgeworth method, (d) Dorbish and Bowley,s method, (e) Fisher's method, (f) Walsh's method.

Commodity	2006		2007	
	Prize	Quantity	Prize	Quantity
A	10	100	12	150
B	8	80	10	100
C	5	60	10	72
D	24	30	18	33

2. It is stated that Marshall-Edgeworth index number is a good approximation to the Fisher's ideal index number. Verify it using the following data:

Commodity	2003		2004	
	Price (Rs)	Quantity (kg.)	Price (Rs)	Quanity (kg.)
A	20	8	40	9

B	50	10	60	10
C	40	15	60	12
D	10	15	15	15

3. Find Laspeyre's, Paasche's. and Fisher's quantity index numbers from the following data:

Commodity	Base year		Current year	
	Price (Rs)	Quantity(Kg.)	Price(Rs)	Quantity(Kg.)
A	5	25	6	30
B	10	5	15	4
C	3	40	2	50
D	6	30	8	35

4. Calculate Fisher's ideal Index from the following data and prove that it satisfies both Time Reversal and Factor Reversal Test.

Commodity	Base Year		Current Year	
	Price	Quantity	Price	Quantity
A	6	50	10	60
B	2	100	2	120
C	4	60	6	60

5. Calculate a weighted average of relative quantity index using 1995 as base period.

Commodity	Quantity (in 1000 kg)		Price (Rs. / kg)
	1995	1999	1995
Wheat	29	24	3.80
Corn	3	2.5	2.91
Soya beans	12	14	6.50

6. **State True or False**:

i) The Fisher Ideal Index number is a compromise between two well known indexes -not a right compromise, economically, for the statistician.

ii) Like relatives are based on the idea that one series can be converted into another because time reversibility holds.

iii) Index numbers are the signs and guide posts along the business highway that indicate to the businessman how he should drive or manage.

iv) Index numbers measure change in magnitude of a group of distinct but related variables.

v) Prices should be for the same unit of quantity in index numbers.

vi) Quantity relatives are used to measure changes in the volume of consumption.

vii) Weighting of index number makes them more representative.

viii) Chain indexes don't give the same results as do fixed based index numbers.

ix) Weighted average of relatives and weighted aggregative methods render the same results.

x) Paasche's formula is a weighted aggregate index with quantity weights in the base year.

11.5 Chain Base Index Numbers

All the methods assume that the base period is a fixed period. If the base period is far from the current period, those methods are not relevant. In this case, chain base index numbers are used which compare relative changes in any period with that of the immediately preceding period.

For construction of an index by the chain base method, a series of indexes is computed for each period with preceding period as the base. These methods are called link relatives.

$$\text{Link Relative} = \frac{\text{Current year's Price}}{\text{Previous year's price}} \text{ x } 100$$

Then these link relatives are chained together.

Chain index of the current year = $\frac{\text{Link relative of the current year x Chain index of previous years}}{100}$

In case of only series of observations, fixed base indices and chain base indices will be same.

Example: From the following prices of three groups of commodities for the years 2009 to 2013, find:

(i) the chain base index numbers chained to 2009,

(ii)the fixed base index numbers with 2009 as base year.

Groups	2009	2010	2011	2012	2013
I	4	6	8	10	12
II	16	20	24	30	36
III	8	10	16	20	24

Solution(i): Computation of chain base index numbers

Group	Link Relatives based on preceding years				
	2009	2010	2011	2012	2013
I	100	6/4×100=150	8/6×100=133.33	10/8×100=125	12/10×100=120
II	100	20/16×100=125	24/20×100=120	30/24×100=125	36/30×100=120
III	100	10/8×100=125	16/10×100=160	20/16×100=125	24/20×100=120
Total of Link Relatives	300	400	413.33	375	360
Average of Link Relatives	100	133.33	137.78	125	120

Chain					
Indices	100	100×133.33/100 =133.33	133.33×137.78/100 =183.70	183.70×125/100 =229.63	229.63×120/100 =275.56

(ii)Computation of Fixed Base Index Numbers (2009=100)

Price Relatives (Base2009 = 100)					
Group	2009	2010	2011	2012	2013
I	100	6/4×100=150	8/4×100=200	10/4×100=250	12/4×100=300
II	100	20/16×100=125	24/16×100=150	30/16×100=187.5	36/16×100=225
III	100	10/8×100=125	16/8×100=200	20/8×100=250	24/8×100=300
Total of price relatives	300	400	550	687.5	825
Fixed Base Index(Avg of price relative)	100	133.33	183.33	229.17	275

Conversion of Chain Base Index Numbers to Fixed based Index number.

$$\text{Current year F.B.I.} = \frac{\text{Current year C.B.I. x Previous year F.B.I.}}{100}$$

11.6 Base Shifting

The shifting of the base of an index number series from one period to another is called base shifting it is used if the base year is too far from the current year. Moreover, it helps in comparison.

$$New\ Index\ Number = \frac{Old\ Index\ Number\ of\ Current\ Year}{Old\ Index\ Number\ of\ New Base\ Year} * 100$$

Example: The following are the index numbers of prices based on 2004 prices. Shift the base from 2004 to 2008:

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012
Index No	100	140	260	340	400	450	500	260	240

Solution: Computation of Index Numbers (Base 2008=100)

Year	Index Number (Base2004=100)	Index Number(Base2008=100)
2004	100	100/400 ×100 =25
2005	140	140/400×100=35
2006	260	260/400×100=65
2007	340	340/400×100=85
2008	400	400/400×100=100
2009	450	450/400×100=112.5
2010	500	500/400×100=125
2011	260	260/400×100=65
2012	240	240/400×100=60

11.7 Splicing

There may be a situation when index numbers constructed on the basis of a particular base period is discontinued after some time. Thus a new series of index numbers may be computed using the last period of the first series as the base. There will be two series of index numbers. In order to make the series continuous, the two series are combined. This process of combining

the series is known as splicing. If the old series and the new series are combined in a way that Base of the Continuous series will be that of old series, it is forward splicing. If Base of the series will be that of the new series, it is Backward splicing.

11.8 Deflating

Deflating is the process of making allowances for the effect of changing price levels. During inflation, the purchasing power of money falls, so it becomes necessary to adjust. The purchasing power of money is given by the reciprocal of the index number.

$$\text{Real Income} = \frac{No\min al\ Income}{\Pr ice\ Index} * 100$$

The real income is called Deflated Income.

$$\text{Thus Real Wage Index} = \frac{Index\ of\ Money\ Wage}{Consumer\ \Pr ice\ Index} * 100$$

11.9 Uses of Index Numbers

- Provides economic indicators like prices of commodities, industrial and agricultural products, share prices etc. which helps to know the trends among different states and nations.
- Helps to know changes in the living standards of people. It indicates change in real income.
- Consumer Price Index number helps to decide the amount of dearness allowance (D.A.) to be given to employees.
- Price Index number is useful to analyze the business conditions and to take decisions accordingly.
- Export-Import Index tells about the foreign trade position of the country, thus government formulate policies accordingly.
- Wholesale price Index helps to decide the interest rate on Bank loans, deposits, government securities, Bank rates etc.
- They are helpful to forecast future pricing, marketing and sales policies.
- They also help to deflate time series data on wages, national income etc.

11.10 Precautions in Construction of Index Number

- The purpose of construction of Index Number should be clearly defined.
- Selection of Base period is an important factor in the construction of Index Numbers. It should not be too far from the present time period. It needs to be a stable year too i.e. prices in that year should not be exceptionally high or low.
- Selection of Weights should be according to the quantity produced, quantity consumed and quantity sold.
- Selection of average need to be appropriate. Though Geometric mean is the best average, year due to certain difficulties, Arithmetic mean is taken.
- Selection of Items itself is of great importance. They should be widely consumed, relevant, comparable and representative.
- Selection of sources of data is an important factor.

11.11 Summary

Index Numbers are used in describing relative changes generally expressed in percentage form in economic variables like price, income, production, employment etc. These are helpful to measure the relative change in the level of the variable with respect to time, location or any other reason. Various types of Index Numbers like Price Index Numbers, Quantity Index numbers, Value index numbers are used. In order to measure the relative changes, different methods for construction of Index numbers can be used, each having its own relative advantage. The concept of chain Index number is very useful in case base period has become too old.

Moreover splicing and deflating are useful methods available in case of combining the two series and to record the effect of rising prices respectively.

11.12 Glossary:

- **Index Numbers:** A tool for comparing values of a variable over a period of time.
- **Price Relative:** The ratio of price in the base year.
- **Quantity Relative:** The ratio of quantity in the current year to the quantity I the base year.
- **Value Relative:** The ratio of value in the current year to the value in the base year.
- **Deflating Prices:** The process of dividing price of a commodity by wholesale price index to remove the effect of inflation.

- **Base Shifting:** The process of changing the base of a series of index numbers to facilitate comparison between two series.
- **Splicing of Index Numbers:** The process of joining two series with slight overlapping.
- **Aggregate Price Index:** A composite price index based on the prices of a group of items.
- **Weighted Aggregate Price Index:** A composite price index in which the prices of the items in the composite are weighted by their relative importance.
- **Laspeyres Index:** A weighted aggregate price index in which the weight of each item in its base-period quantity.
- **Paasche Index:** A weighted aggregate price index in which the weight of each item in its current-period quantity.
- **Quantity Index:** An index designed to measure changes in quantities over time.

11.13 Suggested Answers to SAQ's:

1. Answer (a) 118.5, (b) 119.18, (c) 118.68, (d) 118.615, (e) 118.61, (f) 118.64
2. Answer (145.22)
3. Answer (a) 115.79, (b) 113.76, (c) 114.77
4. Answer (a) 143.05, (b) Both the tests satisfied.
5. Answer 96
6. I) F, ii) T, iii) T, iv) T, v) F, vi) T, vii) T, viii) F, ix) T, x) F,

11.14 Suggested Readings:

- Black, K. Business Statistics for Contempory Decision Making, Fifth Edition, Wiley India.
- Srivastava, T.N. and Rego,S., Statistics for Management, Fourth Reprint, Tata McGraw Hill Companies.
- Thukral, J.K. Business Statistics, Second Edition, TAXMANN'S.

11.15 Model Questions:

(1) It is said that 'Index numbers are specialized type of averages'. How far do you agree with this statement? Briefly explain the Time Reversal Test and Factor Reversal Test.

(2) Discuss the various problems faced in the construction of Index Numbers.

(3) What is Fisher's Index? Why is it called 'Ideal'?

(4) An index of clothing prices for 2006 based on 2000 is to be constructed. The clothing items considered are shoes and dresses. The prices and quantities for both years is given below. Use 2000 as the base period and 100 as the base value.

Item	2000		2006	
	Price	Quantity	Price	Quantity
Dress (each)	75	500	85	520
Shoes (pair)	40	1,200	45	1,300

(a) Determine the simple average of the price indexes.

(b) Determine the aggregate price index for the two years.

(c) Determine Laspeyre's price index.

(d) Determine the Paasche price index.

(e) Determine Fisher's ideal index.

(5) Compute – (a) Laspeyre's ; (b) Paasche's ; (c) Fisher's ; (d) Bowley's index numbers from the following table:

Items	p_0	q_0	p_1	q_1
A	10	4	12	6
B	15	6	20	4
C	2	5	5	3
D	4	4	4	4

(6) Compute Cost of living Index number using Laspeyre's and Paasche's methods from the following data:

Items	Unit consumption in Base year	Price in Base year(Rs.)	Unit consumption in Current year	Price in Current year(Rs.)
Rice(per kg.)	150	20	152	23
Wheat(per kg.)	100	14	95	17
Pulses(per kg.)	40	30	37	42
Oil(per litre)	42	60	35	80
Sugar(per kg.)	50	12	44	18
LPG(per cylinder)	12	270	13	330

(7) The following table shows the sales of chemicals in million rupees from 2003-2004 to 2011-2012 in the country. Using 2003-2004 as the base year, compute index values from 2004-2005 to 2011-2012.

Year	Sales (in million Rs.)
2003-2004	62560.6
2004-2005	72301.6
2005-20006	57574.7
2006-2007	98347
2007-2008	128260
2008-2009	152130

2009-2010	178570
2010-2011	222160
2011-2012	249420

(8) Calculate Fisher's price index for the following data and prove that it satisfies both Time Reversal and Factor Reversal test:

Items	Quantity (kg)		Price (Rs. / kg)	
	2001	2002	2001	2002
Wheat	8	10	20	30
Sugar	6	9	14	18
Tea	2	5	15	20

(9) Calculate the Chain base index numbers from the following data:

Prices in Rupees

Commodity	2009	2010	2011	2012	2013
A	12	13	14	12	12
B	13	16	19	14	13
C	14	22	30	18	26
D	15	18	28	21	32

(10) Shift the base from 2004 to 2006 in the data given below:

Year	Index (2004 = 100)

2001	87.27
2002	90.91
2003	95.40
2004	100.00
2005	104.00
2006	106.00
2007	112.00

Chapter 12: Time Series Analysis

A statistics professor wants to predict marks of newly admitted students in statistics course to infer whether the subject was found to be easy or difficult. He has been teaching the subject for last ten years. He collected data regarding marks in the subject of 30 students from last 10 years. Is it possible for him to forecast the marks of incoming students of next year or of the students who will enter the course five years from present? What techniques should he use to forecast for near future i.e. for newly admitted students and to predict for the students who will enter after five years?

12.0 Objectives

The objective of this chapter is to enable the student to understand:

- Meaning of time series and its components.
- Various smoothing techniques used to forecast stationary time series data.
- Trend analysis based on least squares method
- Methods of finding forecast error: Which method is better?

12.1 Introduction

Forecasting is the method of predicting the future. This method is used every day by planners, decision makers and common man. The business decision involves decision about future sales, changes in production levels dependent on accurate prediction of sales, hiring new staff or firing old one or procurement of finances for future operations. For example:

- Business people predict resurgence in economy with establishment of new government.
- IT sector put hold on recruiting predicting slowdown in developed economies.

- Decision makers increases interest rates to curb inflation predicting higher prices because of increase in oil prices.

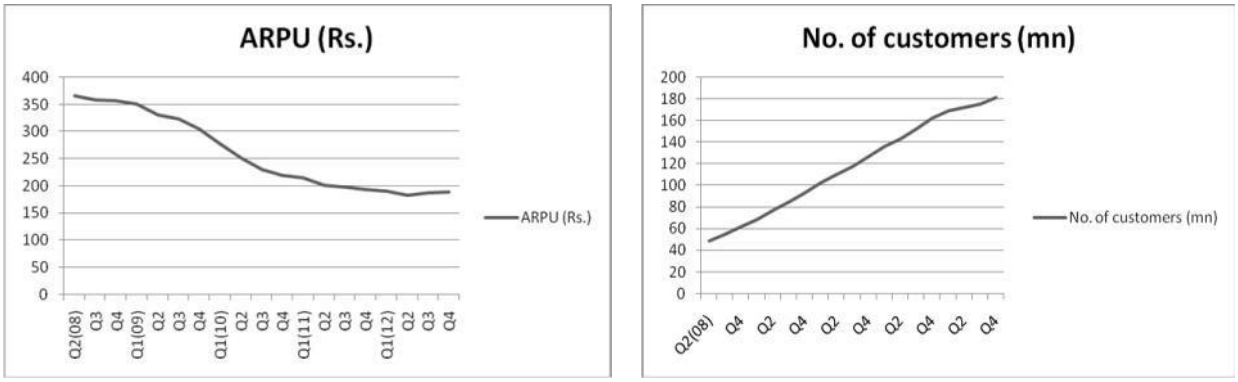
There are numerous situations in business that decision makers encounter on daily basis which involve prediction. But the question is how to predict the future which is uncertain to say the least and how to have accurate forecasts? This chapter deals with some techniques of forecasting time series data. The data which is collected over a continuous period of time is called time series data. For instance, sales over last four quarters or crime rate in different cities over last twelve months etc.

12.2 Time Series components

The data collected over certain time period and for which forecast has to be made can vary in different ways. For instance, sale of laptops shows an increasing trend whereas that of desktops show a decreasing trend; sale of air conditioners show a seasonal variation whereas economy going through cycles of boom and bust show cyclical variation. Generally, time series data is considered to be composed of four components; trend, seasonal, cyclical and irregular.

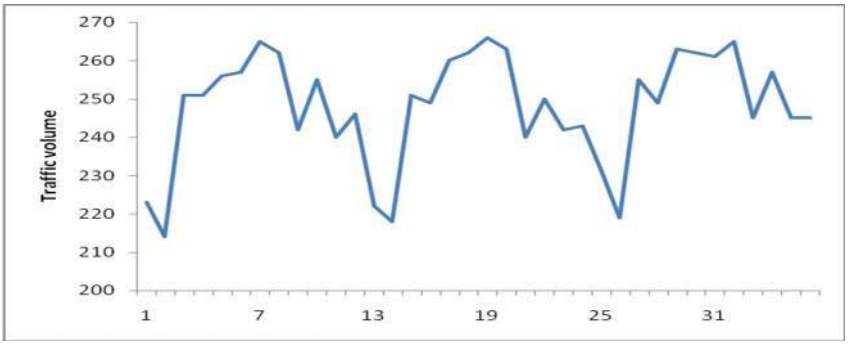
Trend: The general direction of data over a long term is referred to as trend. For instance, trends of sales, employment, and other business and economic series follow various patterns. Some move steadily upward, others decline, and still others stay the same over time. Sale of smart phones for last 2-3 years have shown an increasing trend whereas over the same period sale of feature phones have declined. As shown in Fig. 1 average revenue per user of Bharti Airtel is indicating a declining trend whereas the company has shown an increasing trend in terms of number of customers acquired over a period of four years.

Fig. 1



Cyclical Variation: Cyclical variations are variation in the data with highs and lows through which data move over time periods usually of more than one year. A typical business cycle consists of a period of prosperity followed by periods of recession, depression, and then recovery with no fixed duration of the cycle. The following example as shown in Fig. 2 gives upward and downward movement of traffic volume data over a period of three years.

Fig. 2



According to Fig. 2 traffic volume increases and then decreases and repeats the same process over a period of three years. One important thing to note is that time period in one cycle of upward and downward movement is different from other cycles.

Seasonal Variation: Seasonal variations are also variations of high and low movement of data but over a shorter period of time usually less than one year. The seasonal effects are measured over a period of week or month or quarterly. For example, sale of woolen clothes increase in winter and decreases during summer, hotel occupancy increases during tourist season and falls during off season. Fig. 3 gives an example of earnings of a company which follows seasonal cycle per year. According to Fig. 3 earnings increase in first half of year and then decreases during second half of the year and the cycle gets repeated every year.

Fig. 3

Importantly the difference between cyclical and seasonal variation is twofold: one, the time period of cyclical variation is more than one year whereas for seasonal variations it is less than a year. Secondly, in cyclical variation it is difficult to predict the duration for which high and low movement would remain whereas in case of seasonal variations the duration for upward and downward movement can be predicted. For instance, a country's' economy would go through a cycle of boom and bust but it is difficult to predict for how long these two cycles would remain whereas in case of woolen clothes it is easier to predict the time duration of increase and decrease in sale of woolen clothes.

Irregular variation: Irregular variations are sudden changes in the data which occur over a shorter time periods than seasonal effects. These fluctuations are recorded on a daily basis and rapid or sudden variations are difficult to explain. For instance, variations in stock of a company recorded on a daily basis. It can show abrupt changes in stock prices over a very short term.

Time series data that do not show any of the trend, cyclical or seasonal variations are referred as stationary data. Thus, techniques used to analyze and forecast stationary data deal with only irregular variations. These techniques are discussed in next section under smoothing techniques.

Exercise 1 (True/False)

1. Selection of forecasting technique depends on how data like sales vary over certain time period.
2. Stock market fluctuations of a company during a particular can be considered as trend variation.
3. An economy going through long periods of recession and growth is categorized as seasonal cyclical effect.
4. Sale of milk products is an example of seasonal variation.

12.3 Smoothing techniques

There are various smoothing techniques which attempts to smoothen out the irregular fluctuations in the data for forecasting purpose. Some of the techniques that are discussed are:

- Naïve forecasting model
- Simple moving average
- Weighted moving average
- Exponential smoothing

12.3.1 Naïve Forecasting models are used for short term period forecasts where recent data is considered to be as future data for next period. These models are not relevant for data which shows trend, cyclical or seasonal pattern. Thus, this method is applicable for predicting data on daily or weekly basis. Mathematically the model can be depicted as:

$$F_t = d_{t-1}$$

Where F_t = forecast for next period d_{t-1} = data for previous period

for instance if stock price of a blue chip company was Rs.1500 at the closing today then by using naïve forecasting method its price at next opening would be predicted as Rs.1500.

12.3.2 Simple Moving Average

This method uses simple average method to forecast data. It is called moving because the average value of the data changes with moving time. It is computed and then updated for every new time period being considered. For instance, if we have data of oil prices of last twelve months from January to December and we intend to forecast for 3-month period. Then simple moving average method would calculate average of oil prices of January, February and March to predict for April and then would use data of February, March and April to predict for May and so on. The advantage of this method is that it uses the most recent data.

Mathematically the model can be depicted as:

$F_t = (D_{t-3} + D_{t-2} + D_{t-1})/3$ for a 3-period moving average

Where F_t = forecast for next period

D_{t-1} = data of previous 1 month

D_{t-2} = data of previous 2 month

D_{t-3} = data of previous 3 month

Example 1: For the following data develop forecasts for periods 5 through 10 using 4-month moving averages.

Table 1

Time Period	Value	Forecasted value
1.	27	----
2.	31	----
3.	58	----
4.	63	----
5.	59	44.75
6.	66	52.75
7.	71	61.5
8.	86	64.75
9.	101	70.5
10.	97	81

The first 4-month moving average is

$= (27 + 31 + 58 + 63)/4 = 44.75$

Similarly second moving average would be calculated by leaving the removing the first value and adding the fifth value.

Second 4-month moving average is $= (31 + 58 + 63 + 59)/4 = 52.75$

The first 4-month moving average is shown against 5^{th} time period indicating the forecasted value of this period as it is computed by taking actual values of previous 4-months. This method keeps on repeating and all forecasted values are calculated by taking 4 month as time series data.

12.3.3 Weighted Moving Average

As seen the forecasting method depends on availability of past data. So accuracy of forecast depends on reliability and availability of past data. The reliability of past data goes on decreasing with the passage of time. So if a manager wants to give different importance to data of different time periods then the manager should be using weighted moving average. In simple moving average method equal weight was given to all time series data whereas in this method a forecaster tends to give more weights to recent data and less importance or weights to old data.

Mathematically the formula can be depicted as:

$F_t = w_1(D_{t-1}) + w_2(D_{t-2}) + w_3(D_{t-1})$

Where w1, w2 and w3 are weights assigned to respective time series data.

Example 2: For the data shown in table 1 of example 1 develop forecasts for given periods using 4-month weighted moving average. The recent month should be given weight of 4, the previous month 2 and the other months 1.

The first weighted average is = (4*63 + 2*58 + 1*31 + 1*27)/8

= 53.25

Similarly, second 4-month weighted moving average is = (4*59 + 2* 63 + 1*58 + 1*31)/8

= 56.37

Table 2

Time Period	Value	Forecasted value
1.	27	----
2.	31	----
3.	58	----
4.	63	----
5.	59	53.25
6.	66	56.37
7.	71	62.87
8.	86	67.25
9.	101	76.37
10.	97	89.12

Note that while calculating forecasted data by using weighted moving average the divisor is always the sum of weights. In example 2 sum of weights is 8, that is why average is calculated by dividing the sum of past 4 month data by 8.

12.3.4 Exponential Smoothing

In moving averages method data keeps on moving forward corresponding to movement in time. The disadvantage of this methodology is that it might result in loss of data. Sometimes the previous data that is left behind might have small but significant importance in predicting the future data. To rectify this problem exponential smoothing method is used in which all previous data is used with exponentially decreasing importance in the forecast. The pattern of smoothing depends on a constant value called as exponential smoothing coefficient denoted by α. The value of α lies between 0 and 1.

Mathematically the formula is depicted as:

$F_{t+1} = \alpha(D_t) + (1 - \alpha)F_t$

where F_{t+1} = forecast for next period

F_t = forecast for present period

D_t = actual data for present period

and α = exponential smoothing coefficient

The selection of value of α is the prerogative of the forecaster. If the forecaster wants to give more importance to actual data for present period and less importance to forecast for present period then high value of α would be selected. Whereas, if less importance is to be assigned to actual data of present period and more to forecast of present period then low value of α is selected. Lets understand this by using following example.

Example 3: following are time series data for eight different periods. Use exponential smoothing to forecast for data given. Use the value for the first period as the forecast for second period. Compute forecasts using α as 0.1 and 0.8.

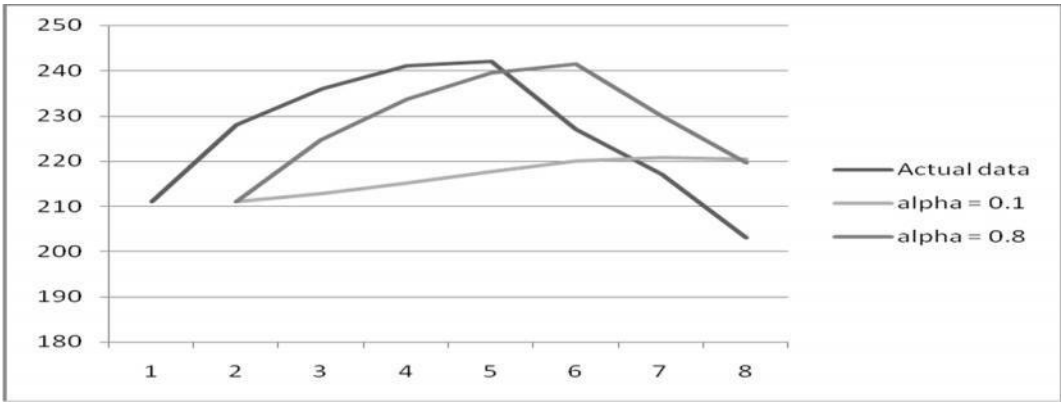
Table 3

Time Period	Actual Value	Forecasted value with α = 0.1	Forecasted value with α = 0.8
1.	211	-----	-----
2.	228	0.1*211 + 0.9*211 = 211	0.8*211 + 0.2*211 = 211
3.	236	0.1*228 + 0.9*211 = 212.7	0.8*228 + 0.2*211 = 224.6
4.	241	0.1*236 + 0.9*212.7 = 215.03	0.8*236 + 0.2*224.6 = 233.72
5.	242	0.1*241 + 0.9*215.03 = 217.62	0.8*241 + 0.2*233.72 = 239.54
6.	227	0.1*242 + 0.9*217.62 = 220.05	0.8*242 + 0.2*239.54 = 241.50
7.	217	0.1*227 + 0.9*220.05 = 220.74	0.8*227 + 0.2*241.5 = 229.9
8	203	0.1*217 + 0.9*220.74 = 220.36	0.8*217 + 0.2*229.9 = 219.58

Note that because no forecast is given for the first time period, so we cannot compute forecast based on exponential smoothing for the second period. Instead actual value for first period is used as the forecast for the second period.

Now according to above illustration, which is the best smoothing coefficient? To understand this quantitatively calculation of forecast errors can provide the result. But that would be discussed in subsequent section. The objective of smoothing coefficient is to smoothen out the irregular fluctuating time series data. Thus, the smoothing coefficient which smoothens out such data more should be considered by the forecaster. This is illustrated graphically in Fig. 4

Fig. 4



According to Fig. 4 smaller exponential smoothing coefficient i.e. alpha = 0.1 smoothens out the irregular data more than alpha = 0.8 does. This is because forecasted data is calculated by multiplying alpha with actual data and 1 – alpha is multiplied with present forecast. So, a forecaster should give less importance to present actual data and more to present forecast if variability is high. Thus, it can be inferred from above example that more

smoothing is obtained with smaller smoothing coefficients and less smoothing is obtained with larger smoothing coefficients.

Exercise 2 (True/False)

1. Simple moving average method gives equal weights to all past data.
2. Weighted simple moving average is more reliable than simple moving average as it is difficult to compute.
3. Exponential smoothing method is easy to compute as it takes into account only recent data of actual and forecasted demand.
4. The selection of value of smoothing coefficient depends on forecast error of previous period.
5. For the following data F6 was calculated as 165. Could a 3-period weighted moving average be used to obtain F6? Is the choice of using exponential smoothing method on the basis that it uses only recent data is correct?

Period	1	2	3	4	5
Demand	70	100	160	100	85
Forecast	90	110	100	110	100

12.4 Linear Trend

The trend data depicts directional data over a long term extending for more than one year. The trend can be increasing or decreasing. The prediction for such time series data which shows a kind of trend can be determined by several statistical methods. This chapter discusses only least squares method for forecasting such data.

The first step in forecasting is to create a linear model where there are two variables: one, is independent or predictor variable and second, is dependent or variable to be predicted. It is important to note that forecasting in such cases should be done where time acts as independent variable which is associated with the dependent variable which needs to be predicted. For instance, sales over a period of time, population growth over certain time period, job performance of employees over last certain years etc.

This model can be depicted by following equation:

$$Y = a + bt$$

where Y is the variable to predicted

t is the independent or predictor variable which is always time in case of time series data

a is the constant or y intercept and

b is the slope of the line

The least squares method fits the best trend line among the two variables under study. The best fit line is that line which results in least possible error between observed and estimated values of the predictor variable. To find the best fit line depicted by above equation values of ‘a’ and ‘b’ needs to be estimated.

Furthermore, often independent variable time is coded to make the equation easier to interpret. In other words, we let t be 1 for the first year, 2 for the second, and so on. When time is coded, we use the following equations to find the slope, b, and the intercept, a, to substitute into the linear trend equation.

The slope $b = \{n\sum tY - (\sum Y)(\sum t)\} / \{n\sum t^2 - (\sum t)^2$

The intercept $a = (\sum Y / n) - b(\sum t/n)$

Example 4: The sales of a small grocery store since 2001 are given below

Year	2001	2002	2003	2004	2005
Sales (million dollars)	7	10	9	11	13

Determine the least squares trend line

Solution: to simplify the calculations the year data is replaced by coded data where 2001 is'1', 2002 is '2' and so on. Thus, following are the calculations for 'a' and 'b'

Table 4

Year (t)	**Sales (million dollars), Y**	**Time (t)**	**tY**	**t^2**
2001	7	1	7	1
2002	10	2	20	4
2003	9	3	27	9
2004	11	4	44	16
2005	13	5	65	25
Sum	50	15	163	55

Now values of the slope by using given formula $b = (5*163 - 50*15) / \{5*55 - (15)^2\}$

$= 1.3$

Value of intercept $a = (50 / 5) - 1.3*(15/5)$

$= 6.1$

So, linear equation becomes $Y = 6.1 + 1.3t$

The value of b = 1.3 indicates that sales increase at a rate of 1.3 million dollars per year and value of 6.1 indicates that during base year i.e. when t = 0 (representing year 2000) sales would be 6.1 million dollars.

By using these values of 'a' and 'b' a trend line can be plotted at different values of t which would be the best fit line with least error.

Example 5: By using data in table 4 what would be the sales in year 2008?

Solution: According to the coded scheme year 2008 would be given value '8'. Substituting this value in formulated linear equation: $Y = 6.1 + 1.3t$

$Y = 6.1 + 1.3(8) = 16.5$

So the estimated sales of the grocery store in year 2008 would be 16.5 million dollars.

Exercise 3 (True/False)

1. The independent variable has to be time variable in trend method for time series data.
2. The stock price of a company can be predicted it is showing an increasing trend for last one month.
3. The accurate calculation of values of y intercept and slope intercept is important for accurate prediction.
4. Value of 'b' can be zero or negative.
5. If the last observed value for following equation is 65 what would be the next value
   Y(t) = 625 – 1.3 Y(t-1)

12.5 Forecast Error

This chapter has explained various techniques of forecasting the time series data. But a forecaster has to decide which forecasting method to use. The obvious answer to this could be the method which gives more accurate forecast or in other words the method which provides least erroneous results should be selected. Forecast error is the difference between forecasted and actual value of the data under study represented by following formula.

Forecast error = actual value (D_t) – forecasted value (F_t)

It is important to understand that actual data is the present data which has occurred and forecasted data is the future data which has to occur. The various methods to evaluate forecast errors that would be discussed are:

- Mean Absolute Decision (MAD)
- Mean Square of Errors (MSE)
- Mean Absolute Percent Error (MAPE)

12.5.1 Mean Absolute Deviation (MAD) is the average of the absolute value of forecast errors.

$$MAD = \text{absolute value } (D_t - F_t) / n$$

For example, by taking data of table 3 MAD is calculated for forecasted value with alpha = 0.1 and with alpha = 0.8

Table 5

Time Period	Actual Value	α = 0.1			α = 0.8		
		Forecasted value	Forecast error	Absolute value	Forecasted value	Forecast error	Absolute value
1.	211	-----	-----	-----	-----	------	-----
2.	228	211	228-211=17	17	211	17	17
3.	236	212.7	23.3	23.3	224.6	11.4	11.4
4.	241	215.03	25.97	25.97	233.72	7.28	7.28
5.	242	217.62	24.38	24.38	239.54	2.46	2.46
6.	227	220.05	6.95	6.95	241.50	-14.5	14.5
7.	217	220.74	-3.74	3.74	229.9	-12.9	12.9
8	203	220.36	-17.36	17.36	219.58	-16.58	16.58
				16.95			**11.73**

MAD (when alpha = 0.1) = 16.95

MAD (when alpha = 0.8) = 11.93

Thus, by using MAD smoothing coefficient alpha = 0.8 gives more accurate forecasts as the results were less erroneous.

12.5.2 Mean Square Error (MSE) is computed by squaring each error and then calculating their average. The squaring is done to negate the negative effect of errors.

$$MSE = \sum(D_t - F_t)^2 / n$$

For example, by taking data of table 3 MSE is calculated for forecasted value with alpha = 0.1 and with alpha = 0.8

Table 6

Time Period	Actual Value	α = 0.1			α = 0.8		
		Forecasted value	Forecast error	Square value	Forecasted value	Forecast error	Square value
1.	211	-----	-----	-----	-----	------	-----
2.	228	211	228-211=17	289	211	17	289
3.	236	212.7	23.3	542.89	224.6	11.4	129.96
4.	241	215.03	25.97	674.4409	233.72	7.28	52.99
5.	242	217.62	24.38	594.3844	239.54	2.46	6.05
6.	227	220.05	6.95	48.3025	241.50	-14.5	210.25
7.	217	220.74	-3.74	13.9876	229.9	-12.9	166.41
8	203	220.36	-17.36	301.3696	219.58	-16.58	274.89
				352.05			**161.36**

MSE (when alpha = 0.1) = 352.05

MSE (when alpha = 0.8) = 161.36

Thus, by using MSE method of finding forecast errors the forecaster should select smoothing coefficient as 0.8 as it results in less error as compared to alpha = 0.1

12.5.3 Mean Absolute Percent Error (MAPE) is the average absolute present error.

$$MAPE = \sum\{\text{absolute value of}(D_t - F_t) / D_t\} / n$$

For example, by taking data of table 3 MAPE is calculated for forecasted value with alpha = 0.1 and with alpha = 0.8

Table 7

Time Period	Actual Value	α = 0.1				α = 0.8			
		Forecasted value	Forecast error	Absolute value of error	% age of forecast error	Forecasted value	Forecast error	Absolute value of error	% age of forecast error
1.	211	-----	-----		-----	-----	------	-----	-----
2.	228	211	228-211=17	17	17/228 = 7.45	211	17	17	7.45
3.	236	212.7	23.3	23.3	9.87	224.6	11.4	11.4	4.83
4.	241	215.03	25.97	25.97	10.77	233.72	7.28	7.28	3.02
5.	242	217.62	24.38	24.38	10.07	239.54	2.46	2.46	1.01
6.	227	220.05	6.95	6.95	3.06	241.50	-14.5	14.5	6.38
7.	217	220.74	-3.74	3.74	1.72	229.9	-12.9	12.9	5.94

8	203	220.36	-17.36	17.36	8.55	219.58	-16.58	16.58	8.16
			Average value		**7.35%**		***Average Value***		**5.26%**

MAPE (when alpha = 0.1) = 7.35%

MAPE (when alpha = 0.8) = 5.26%

Thus by using MAPE the forecaster should select alpha = 0.8 as it results in less erroneous forecast.

From a computational viewpoint the difference between these measures is that MAD weights all errors evenly, MSE weights according to their squared values and MAPE weighs according to the relative error.

Exercise 4 (True/False)

1. MSE value in all cases would always be higher than MAD value.
2. MAPE is better method for computing forecast errors than MAD and MSE.
3. Comparison between efficacy of two forecasting techniques can be done by computing forecast errors for two data.

12.6 Summary

The true test of a forecast is the accuracy of the prediction. Until the actual value is obtained for a given time period the accuracy of the forecast is unknown. This chapter discusses various methods of forecasting for time series data. The data that has been recorded over a certain time period is called time series data. The selection of a particular technique of forecasting depends on how the data is varying and time span for which data is recorded. The selected data can show four different patterns: trend, cyclical, seasonal and irregular variations. Irregular variations occur for short term period whereas trend and cyclical variations are recorded for long term. Seasonal variations which occur in less than a year are considered to be medium term. This chapter has considered in detail the calculations and methodology of irregular fluctuating time series data. Certain models in this respect such as moving averages and exponential smoothing have been discussed. Trend method is also explained mathematically for time series data that varies over a long term and shows either an increasing or decreasing trend. The selection of various forecasting methods depends on its accuracy or the method which results in least erroneous forecasts. Three methods viz. MAD, MSE and MAPE have been discussed.

12.7 Glossary

- **Time series data:** is the data which has been collected regarding a variable over a period of time at regular intervals.
- **Seasonal effects:** indicates the variation of data collected over certain time period of less than one year and shows fluctuations with respect to seasons.
- **Cyclical variation:** represents the variation of data collected over long term of more than one year where data shows fluctuations in the form of peaks and troughs.

- **Irregular fluctuations:** when data is collected over a short term and does not form any patter.
- **Forecast errors:** the difference between estimated and observed data values.
- **Mean Absolute deviation (MAD):** is the average of absolute values of deviations of certain data around its mean.
- **Mean Square Error (MSE):** is the average of squared value of forecast errors of data under consideration.

12.8 Answers to check your progress/ Self assessment exercise

Exercise 1

1. True
2. False
3. True
4. True

Exercise 2

1. True
2. False
3. False
4. True
5. Not possible to get F6 as 165 from D3 = 160, D4 = 100 and D3 = 85 using a 3 period moving average since the largest possible forecast is 160 with the weights w3=1, w4=0 and w5=0. Also, the claim that exponential smoothing method only includes recent data is not true.

Exercise 3

1. True
2. False
3. True
4. True
5. 540.5

Exercise 4

1. True
2. False
3. True

12.9 References/ Suggested Readings

- Black, K., *Business Statistics For Contemporary Decision Making,* Fifth Edition, Wiley India,
- Keller, G., *Statistics for Management,* First India Reprint 2009, Cengage Learning India Private Limited.

- Stevenson, W.J., *Operations Management,* Ninth Edition, Tata McGraw Hill, New Delhi, 2009.
- Donald R. Cooper & Pamela S. Schindler, *Business Research Methods,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 9th Edition.
- S.P. Gupta, *Business Statistics*, Sultan Chand, New Delhi.

12.10 Terminal and Model Questions

1. What advantages does exponential smoothing have over moving averages as a forecasting tool?
2. What factors enter into selecting the value of exponential smoothing coefficient?
3. An electrical contractor's records during the last five weeks indicate the number of job requests:

Week	1	2	3	4	5
Requests	20	22	18	21	22

Predict the number of requests for week using following methods:

(i) Naïve

(ii) A four-period moving average

(iii) Exponential smoothing with alpha = 0.30. Use 20 for week 2 forecast.

4. A cosmetics manufacturer's marketing department has developed a linear trend equation that can be used to predict annual sales of its popular cream

$F_t = 80 + 15t$

Where F_t = annual sales ('000 bottles) and t = 0 corresponds to 2000

(i) Are annual sales increasing or decreasing? By how much?

(ii) Predict annual sales for the year 2010 using the equation.

5. Following are time series data for nine time periods. Use exponential smoothing constants of 0.3 and 0.7 to forecast time periods 3 through 9. Let the value for time period 1 be the forecast for time period 2. Compute the forecasts for period 4 through 9 using a 3-month moving average. Compute the errors for the forecasts and interpret which method is suitable for forecasting?

Time	1	2	3	4	5	6	7	8	9
Value	9.4	8.2	7.9	9.0	9.8	11.0	10.3	9.5	9.1

6. For the following data compute:

Month	1	2	3	4	5	6	7	8	9	10	11	12
Values	10.08	10.05	9.24	9.23	9.69	9.55	9.37	8.55	8.36	8.59	7.99	8.12

(i) Trend in the data by using regression trend analysis.

(ii) Use a 4-month moving average to forecast values.

(iii) Use simple exponential smoothing to forecast values for each of the months by using alpha as 0.3 and then 0.7.

(iv) MAD for forecasts obtained in (ii) and (iii)

Chapter 13: Probability

When Virata, king of Matsya, learnt that that his cows had been stolen by the king of Trigarta, he rode out of his city with his army in hot pursuit of the thieves. Taking advantage of his absence, the king of Hastinapur attacked his city. There was no one around, except women and children, to defend Matsya. Everyone was frightened. Uttar, the young prince decided to protect his people from the enemy. What were his chances of saving the people and winning the war?

To do so he asked for a charioteer. A eunuch called Brihanalla, who served in the women's quarters, offered to help since he had some experience. Though not happy to have a eunuch as his charioteer, the prince, armed with a bow, rode out to face the army of Hastinapur in battle. But when Uttar entered the battlefield and saw the enemy before him, he trembled in fear. In panic, Uttar, jumped off the chariot and began running back towards the city. Why did he do so? Do his chances of winning have diminished or his earlier calculation of chances of winning the war was wrong?

13.0 Objectives

This chapter should help students to understand:

- Different methods of probability
- Types of event
- Different laws of probability: addition law, multiplication law and conditional probability
- Application of Bayes' rule

13.1 Introduction

Decision makers encounter uncertainty in decision making almost on daily basis. For example, a pizza restaurant does not know definitely the demand of pizzas during the weekend. An operations manager wants to know in advance the demand so that the manager can gather resources for increased demand. An HR manager needs this information for new recruitment purposes. A finance manger has to look for sources of finance. This entire decision making depends on chance of increased demand in near future. As most such questions do not have definite answers the decision making is based on uncertainty. Such situations can be dealt to a good extent by assigning quantitative values to the likelihood of an outcome. This chapter is about how to assign probabilities.

Statistics is studied under two different branches: descriptive and inferential statistics. Much of the decision making involves making an inference from small data. Probability forms the basis of inferential statistics as this branch of statistics involves collecting data from a sample and making decision about population. The reason for doing so is non-availability of data of population. So, a decision maker by applying certain statistic tools makes a conclusion about entire population by studying the representative sample. For instance, a quality manger checks the quality of pistons by selecting a small sample and if this sample passes the quality standards then entire population of pistons is considered to be good. But as all pistons were not checked so there is always certain probability of some pistons in the population to be faulty.

13.2 Three approaches to assigning probabilities

13.2.1 Classical method: This method helps to assign probabilities associated with games of chance. It involves an experiment, which is a process that produces outcome and an event, which is an outcome of an experiment. For instance, tossing of coin would result in associating a probability of 0.5 to an outcome i.e. head or tail. In a throw of dice the probability of getting a three is 1/6. Thus, in classical method probability of an event occurred is determined by the ratio of number of items in a population containing the event to the total number of items in the same population. In throwing of dice example, total number of items was 6 and number of items containing the event was one i.e. three occurs only once.

This method is also referred as *a priori* approach method because the probability of an event can be determined in prior to the experiment. For instance, if in a throw of dice probability of occurrence of three was 1/6, then if the dice would be thrown again i.e. experiment is again conducted then this approach of classical method tells us that probability of the same event would be 1/6. Similarly, suppose there are three machines A, B and C producing same products. The manager has prior information that out of total products produced 20% are produced by machine A and they are defective. Now, if an experiment is conducted that all items from three machines are mixed and an item is selected, so what is the probability that the selected item would be defective? According to priori approach the chance of selecting a defective item would be 20%.

This approach does not help the manger to make an inference. For instance, an experiment is conducted in which a coin is thrown ten times and every time head is obtained. The probability of getting a head would be

0.5. So, if a coin is thrown eleventh time can it be said that this time also we will get head? No, because the occurrence of one event does not depend on the other.

13.2.2 Relative Frequency of Occurrence: is associated with long run events. The lack of ability of classical approach in making an inference is rectified by relative frequency method of probability. For instance, in a shopping complex if out of footfall of 1000 only 200 individual did shopping on a particular day. Then probability of an individual to shop would be 200/1000 = 0.2. This figure represents only an estimate as in the long run number of customers visiting the shopping complex and indulging in shopping might vary. The larger the data collected the better would be the estimate.

13.2.3 Subjective approach: this approach is applied when data about past outcomes of an experiment are either not available or show very irregular pattern. For instance, who would win the next match? Which party would form the next government? In such cases, the probabilities are assigned to the outcomes on the basis of individuals' feelings or expert judgments. It is not a scientific approach but the reasoning is based on managers' accumulation of knowledge and experience. So the process of assigning probabilities depends on managers' intuition rather than on mathematical tools.

Exercise 1

1. If I calculate the probability of an event and it turns out to be 0.7, I know that
   i. the event is probably going to happen.
   ii. the event is probably not going to happen.
   iii. the probability of it not happening is .3.
   iv. I made a mistake.
2. If I flip a fair coin 10 times, which of the following is true?
   i. The number of heads will equal the number of tails.
   ii. The probability of all heads is greater than the probability of all tails.
   iii. The probability of HHHHHHHHHH = the probability of HTHTHTHTHT.
   iv. The probability of HHHHHHHHHH < the probability of HTHTHTHTHT.

13.3 Defining Events

The process of producing outcomes is called an experiment. The outcome of the experiment is called an event. Checking the quality of finished goods is an experiment and finding the defective item is an event. Appearing for the exam is an experiment and passing or failing in the exam is an event.

Mutually Exclusive Events: the events which cannot occur simultaneously are called mutually exclusive events. These events can occur in a sequence but not at the same point in time. For instance, a student cannot pass or fail in the same exam at same time. A product cannot be deemed as good or bad simultaneously.

Independent Events: the occurrence of an event does not affect the probability of occurrence of another event. For example, in case of classical approach to probability the probability of occurrence of a head does not depend on probability of same event which occurred earlier. In cases where an item is selected randomly then the outcomes can be dependent or independent. For example, suppose a bag contains 10 red balls. If one ball is drawn randomly what is the probability that it is red. According to classical method the probability of getting one red ball from a total of 10 balls would be 1/10.

Now further there can be two cases:

With replacement: now if the withdrawn ball is restored back in the bag and then again one ball is withdrawn at random, what is the probability of it being red. The selection of one event out of total of 10 would give us probability as again 1/10.

Without replacement: after withdrawing one ball, it is not restored in the bag. Then another ball is withdrawn, what is the probability of it being again red. Now number of balls left in the bag are 9, so probability of getting another red ball would be 1/9. *Thus in this case probability of occurrence of one event is dependent on occurrence of another event.*

Symbolic notation of probability of two independent events is:

P(X!Y) = P(X) denotes the probability of occurrence of X when Y has already occurred but X is not dependent on Y.

P(Y!X) = P(Y) denotes the probability of occurrence of Y when X has already occurred but Y is not dependent on X.

Exhaustive events: depicts the all possible events for an experiment. For instance, if a pair of dice is thrown then how many events represent occurrence of five? The answer would be (1,4), (2,3), (4,1) and (3,2) representing the outcome of five in all possible ways.

Exercise 2

1. Tickets numbered 1 to 20 are mixed up and then a ticket is drawn at random. What is the probability that the ticket drawn has a number which is a multiple of 3 or 5?

(i) ½ (ii) 2/5 (iii) 8/15 (iv) 9/20

2. In a box, there are 8 red, 7 blue and 6 green balls. One ball is picked up randomly. What is the probability that it is neither red nor green?

(i) 1/3 (ii) ¾ (iii) 7/19 (iv) 8/21

3. Which of the following are likely to be dependent events?

i. the weather and the number of books on your shelf
ii. the color of your car and its gas mileage
iii. the weight of your car and its gas mileage

iv. the size of your house and the size of your shoe

4. If I sample with replacement, which of the following may be true?

i. The numerator for the next event's probability changes.

ii. The denominator for the next event's probability changed.

iii. Both the numerator and denominator for the next event's probability change.

iv. None of the values used in calculating the next event's probability change.

13.4 Laws of Probability

Four laws of probability have been presented in this chapter: addition law, multiplication law, conditional probability and Bayes' rule.

13.4.1 Addition Law

Law of addition is used to find the probability of union of two events. If there are two events E1 and E2 then their union probability is represented as P(E1 U E2) indicating occurrence of either of two events or both. Addition law is applied depending on type of events.

(i) Events are mutually exclusive: to illustrate such case consider the example shown in Table 1. Let us try to find probability that a randomly selected physician is less than 35 years old or 55-64 years old. Now according to the data given there are two events: a physician in the age group of less than 35 years (E1) and a physician in the age group of 55-64 years (E2).

P (E1) = 0.18 and P(E2) = 0.14

Now, these two events cannot occur simultaneously i.e. a physician who is less than 35 years old cannot be in 55-64 years age category.

Thus, P(E1 U E2) = P(E1) + P(E2) = 0.18 + 0.14 = 0.32

So, law of addition in case of events which are mutually exclusive is:

P(E1 U E2) = P(E1) + P(E2)

Table 1

		Age (years)					
		<35	35-44	45-54	55-64	>65	
Gender	Male	0.11	0.20	0.19	0.12	0.16	0.78
	Female	0.07	0.08	0.04	0.02	0.01	0.22
		0.18	0.28	0.23	0.14	0.17	***1.00***

(ii) Events are not mutually exclusive: by considering the data in Table 1, let us try to find probability that a randomly selected physician is a man or is 35-44 years old. There are two events: age of physician (E1) and gender of physician (E2).

Probability of an individual being selected to be in the age group of 35-44 years is P(E1) = 0.28

Probability of an individual being selected to be a man is P(E2) = 0.78

So P(E1 U E2) = P(E1) + P(E2) = 0.18 + 0.78 = 0.96

But in this case the two events can occur simultaneously i.e. they are not mutually exclusive. A male physician can belong to the age category of 35-44 years old. The probability of such an event is given as 0.20. This should be subtracted from above calculated probability as probability assigned to two events has already been counted. Thus, P(E1 U E2) = 0.28 + 0.78 – 0.20 = 0.86

So, law of addition in case of events being mutually exclusive is:

P(E1 U E2) = P(E1) + P(E2) – P(E1 and E2)

13.4.2 Conditional Probability

In conditional probability, the probability of an event, E1 is determined which is dependent on occurrence of another event E2. In this case the occurrence of E2 is already known to the researcher. The process involves determining probability of joint events which are not independent. For instance, determining the probability of buying a retirement policy by an individual who belongs in the age category of 35-45 years. The researcher has prior knowledge of an event i.e. age category of the individual. This method has important role in laws of multiplication and Bayes' method. The mathematical expression of conditional probability is:

P(E1!E2) = P(E1 and E2)/P(E2)

where P(E1!E2) is the probability of E1 which has to be determined when probability of occurrence of E2 is known.

P(E1 and E2) is the joint probability of two events under consideration and,

P(E2) is the known probability of event E2

Example2: By taking data of table 1 find the probability of a physician being selected to be women who belongs to age category of 45-54 years.

Solution: E1 represents the event of selection of female physician for which probability is to be determined and E2 represents the event of selection from age group of 45-54 years which is know.

Thus, P(E1!E2) = P(E1 and E2)/P(E2)

= 0.04 / 0.22

= 0.182

Example 3: in a certain city, 30% of the families have a Master card, 20% have an American Express and 25% have a visa card. 8% of the families have both a master card and an American express card. 12% have both a visa and a master card. 6% have both an American express and a visa card.

(i) If a family has a master card, what is the probability that it has a visa card?

(ii) If a family has a visa card, what is the probability that it has a master card?

(iii) If a family has an American express card, what is the probability that it has a visa card?

Solution:

P(Master card) = 0.3, P(American express card) = 0.2, P(Visa card) = 0.25

P(American express and master card) = 0.08

P(Visa and Master card) = 0.12

P(American Express and Visa card) = 0.06

(i) P(Visa card!Master card) = P(Visa and Master card) / P(Master card)

= 0.12/0.3

(ii) P(Master card!Visa card) = P(Master card and Visa card) / P(Visa card)

= 0.12/0.25

(iii) P(Visa card ! American Express card) = P(Visa card and American Express card) / P(American Express card)

= 0.06/0.2

13.4.3 Multiplication law

Law of multiplication is depicted by intersection between two events i.e. when both events must occur. This is represented by joint probability when two events E1 ***and*** E2 occur. For instance, to qualify for being admitted to a business management course the candidate should pass through the entrance exam as well as should have minimum qualifying marks in the graduation. Both events have to happen. The occurrence of either of the event is not sufficient.

The multiplication law is illustrated by taking the data from Table 1:

(i) What is the probability that a randomly selected physician is both a woman and 45-54 years old?

Solution:

(i) In this case there are two events: physician being a woman (E1) and belonging to age group of 45-54 years (E2).

We need to find: P(E1 and E2) which is given as:

P(E1 and E2) = P(E1). P(E2!E1)

Where P(E1) is probability of event E1 and,

P(E2!E1) is probability of event E2 given that E1 has already occurred.

From table 1 P(E1) = 0.22 and

P(E2!E1) = 0.04/0.22

Thus, P(E1 and E2) = 0.22 * (0.04/0.22) = 0.04

So, in case the two events are not independent implying that probability of occurrence of an event is dependent on other then multiplication formula becomes:

P(E1 and E2) = P(E1). P(E2!E1)

In this case the two events were not independent because a selection of woman is categorized by the age group she belongs to.

(ii) In case events are independent:

Independence of events means that P(E1!E2) = P(E1) i.e. probability of occurrence of E1 given that E2 has happened is equal to probability of E1. Similarly P(E2!E1) = P(E2) i.e. prior information of event E2 being already happened has no impact on occurrence of E1.

Example 4: Suppose that a bag has 20 balls out of which 8 are red and 12 are black. Two balls are drawn at random in succession. What is the probability that on both occasions selected balls were red?

Solution: E1 represents drawing of first red ball from the bag.

So P(E1) = 8/20

Now, if this ball is replaced in the bag before the withdrawal of next ball then probability that second ball is red when E1 has already happened, E2 would be

P(E2!E1) = 8/20 = P(E2).

So, these two events are independent as when ball is replaced then probability of occurrence of second event does not depend on prior information of occurrence of first event.

The probability of happening of E1 and E2 in case of independence of events would be

P(E1 and E2) = P(E1).P(E2) = (8/20) * (8/20)

Thus, when events are independent the formula for law of multiplication becomes:

P(E1 and E2) = P(E1).P(E2)

Exercise 3

From the following given joint probabilities associated with smoking and lung disease among 60-65 year old men compute the probability that a randomly selected man:

	He is a smoker	***He is a non-smoker***
He has lung disease	0.12	0.03
He does not have lung disease	0.19	0.66

(i) Is a smoker

(ii) Has along disease

(iii) Has a lung disease given that he is a smoker

(iv) Has lung disease given that he does not smoke

13.5 Bayes' Rule

Bayes' rule is an extension of conditional probability. In case of conditional probability, the probability of an event was determined, the occurrence of which was dependent on the occurrence of another event. The researcher had prior information about the probability of happening of this event. In Bayes' rule this concept is extended to more than one event. If there is more than one event whose occurrence is dependent on known event then these events would interact simultaneously with the dependent event. The determination of probability of one of these events in the presence of other events with respect to the event which has already occurred requires

application of Bayes' rule. For instance, if three individuals A, B and C buy a kilogram of potatoes each and probability of finding a rotten potato in case of A is 1 out of 12, in case of B is 1 out of 6 and in case of C it is 1 out of 3. Then all these potatoes are mixed and randomly one potato is picked. It is given that the selected potato is rotten. What is the probability that it came from A or from B or from C? In this illustration, the researcher had prior knowledge of an event that the selected potato was a rotten one (X). There were other three events for which probability needed to be found i.e. potatoes came from A, B or C. Using this example the mathematical expression of Bayes' theorem for finding out probability that rotten potato came from A in the presence of potatoes of B and C with the prior information that it was a rotten one (event X) is:

P(A!X) = P(A and X) / {P(A and X) + P(B and X) + P(C and X)}

Or

P(A!X) = P(A).P(X!A) / {P(A).P(X!A) + P(B).P(X!B) + P(C).P(X!C)}

The numerator indicates the joint probability of two events whereas the denominator has joint probabilities of all the dependent events.

Example 5: A manufacturing plant produces a product on three different machines A, B and C. Machine A produces 10%, machine B produces 40% and machine C produces 50% of this product. 5%, 12% and 8% of products from machines A, B and C were found to be defective respectively. During quality check one item was picked randomly and was found to be defective. What is the probability that this defective item came from machine A, B or C.

Solution:

P (product was produced by machine A, E1) = 0.10

P (product was produced by machine B, E2) = 0.40

P (product was produced by machine C, E3) = 0.50

If X denotes the event for which prior information is available, in this case it is selection of defective item then,

P (X!E1) = 0.05

P (X!E2) = 0.12

P (X!E3) = 0.08

The probability that defective item came from machine A

P(E1!X) = P(E1 and X) / {P(E1 and X) + P(E2 and X) + P(E3 and X)}

= P(E1).P(X!E1) / {P(E1).P(X!E1) + P(E2).P(X!E2) + P(E3).P(X!E3)}

= 0.10*0.05 / (0.10*0.05 + 0.40*0.12 + 0.50*0.08)

= 0.053

P(E2!X) = P(E2 and X) / {P(E1 and X) + P(E2 and X) + P(E3 and X)}

= P(E2)P(E2!X) / {P(E1).P(X!E1) + P(E2).P(X!E2) + P(E3).P(X!E3)}

= 0.40*0.12 / (0.10*0.05 + 0.40*0.12 + 0.50*0.08)

= 0.516

P(E3!X) = P(E3 and X) / {P(E1 and X) + P(E2 and X) + P(E3 and X)}

= P(E3)P(E3!X) / {P(E1).P(E1!X) + P(E2).P(E2!X) + P(E3).P(E3!X)}

= 0.50*0.08 / (0.10*0.05 + 0.40*0.12 + 0.50*0.08)

= 0.43

Exercise 4

1. A financial analyst estimates that the probability that economy will experience a recession in next 12 months is 25%. It is also believed that if the economy encounters a recession the probability that mutual fund will increase in value is 20%. If there is no recession, the probability that mutual fund will increase in value is 75%. Find the probability that the mutual fund's value will increase.
2. In a survey 8% of customers were not satisfied with the service they received at their last visit to the store. Of those who were not satisfied only 22% return to the store within one year and of those who were satisfied 64% return within a year. A customer was picked randomly and it was found that it is less than one year since he has visited the store. What is the probability that he was satisfied with the service he received?

13.6 Summary

Probability is defined as the likelihood of occurrence of an event. Before assigning probability type of event has to be comprehended. This chapter discusses three types of event: mutually exclusive, independent and exhaustive events. There are different methods of assigning probabilities such as classical method, relative frequency and subjective method. The laws of probability were discussed in the light of different kind of events. Laws of probability such as addition laws, multiplication laws, conditional probability and Bayes' rule give different results depending on the events under consideration. Bayes' rule is an extension of conditional probability which takes into consideration prior probabilities of events occurring and adjusts those probabilities on the basis of information about what occurs subsequently.

13.7 Glossary

- **Probability:** is the likelihood of occurrence of a random event.
- **Random event:** is the event the outcome of which is not biased
- **Classical approach to probability:** is the approach where outcome of an event cannot be inferred from its past occurrence. For example, in a throw of dice probability it cannot be predicted that outcome would be six even if in past experiments the outcome was always six.
- **Subjective approach:** where outcome can be inferred from past experiences. For example, if productivity of few employees was found to be high because of certain incentive, then it can be inferred that if such incentive is given to all employees then productivity can be enhanced.
- **Mutually exclusive events:** the events which cannot occur simultaneously

- **Independent event:** occurrence of one event is not dependent on other
- **Exhaustive events:** at least one event will happen
- **Conditional Probability:** the probability of the occurrence of one event given that another event has occurred.
- **Bayes' rule:** a method involving application of conditional probability used to find the impact of revised probabilities.

13.8 Answers to check your progress/ Self assessment exercise

Exercise 1

1. Iii
2. Iii

Exercise 2

1. Iv
2. i
3. iii
4. ii

Exercise 3

1. 0.31
2. 0.85
3. 0.387
4. 0.044

Exercise 4

1. 0.6125
2. 0.971

13.9 References/ Suggested Readings

- Black, K., *Business Statistics For Contemporary Decision Making,* Fifth Edition, Wiley India,
- Keller, G., *Statistics for Management,* First India Reprint 2009, Cengage Learning India Private Limited.
- Stevenson, W.J., *Operations Management,* Ninth Edition, Tata McGraw Hill, New Delhi, 2009.
- Donald R. Cooper & Pamela S. Schindler, *Business Research Methods,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 9th Edition.
- S.P. Gupta, *Business Statistics*, Sultan Chand, New Delhi.

13.10 Terminal and Model Questions

1. Bayes' rule extends the use of the law of conditional probabilities to allow revision of original probabilities with new information. Explain with example.
2. Suppose 70% of all companies are classified as small companies and rest as large companies. Suppose, 82% of large companies provide training to employees, but only 18% of small companies provide training. A company is randomly selected without knowing if it is a large or small company; however it is determined that the company provides training to employees. What are the prior probabilities that the company is large or a small company? What are the revised probabilities that the company is large or small? Based on your analysis, what is the overall percentage of companies that offer training?
3. Explain the law of multiplication when events are dependent and are independent.
4. In a survey only 42% of small companies offer retirement plans while 61% offer life insurance. Suppose 33% offer both retirement plans and life insurance as benefits. If a small company is selected randomly, determine the following probabilities:
   (i) the company offers a retirement plan given that they offer life insurance.
   (ii) The company offers given that they offer a retirement plan.
   (iii) The company offers a retirement plan or life insurance.

Chapter 14: Probability Distributions

A king had only daughter named Indumati. She was raised as a boy and knew all the arts of warfare. It was clear that whoever will marry Indumati would acquire the entire kingdom. But whenever king used to broach the subject of marriage to her she always gave a hesitant reply. Under pressure from her father Indumati decided to get married only after the prospective groom passes her test. The test involved climbing a steep wall and then jumping from the wall on a three tiered cage which had sharp knives. Indumati decided to marry that person who would come out of the cage unharmed. Because of the difficulty of the test king decided not to reveal its details while sending invitation to prospective grooms. On the day of test several suitors turned up. As it was difficult to put all the suitors to test, Indumati selected few and decided to put them through the difficult test. What is the probability that not more than one suitor out of given sample would pass the test?

14.0 Objectives

The students should be able to capture the following concepts:

- Meaning and significance of probability distributions
- Type of probability distributions
- Characteristics and computation of Binomial distribution
- Characteristics and computation of Poisson distribution
- Characteristics and computation of Normal distribution

14.1 Introduction

Distributions play a significant role for a statistician in deciding which analysis should be applied to the data for decision making. Distributions indicate the behavior of data showing certain characteristics. For instance, in a class of 100, majority of students score average marks, few score very less and few score very high. So most

likely such data show a symmetrical behavior. In an organization very few people earn very high salaries and a large chunk of employees earn average or small wages. Such data most likely shows asymmetrical characteristics. Thus, a manager or a decision maker should understand the distributions of data.

Probability distributions are depiction, either numerically or graphically, of probabilities of all the possible outcomes in an experiment. For example, in an experiment of throwing a dice there can be six possible outcomes and for each outcome a probability of its occurrence can be computed. The selection of a particular kind of probability distribution depends on the type of variable. Firstly, the variable should have a random outcome i.e. its outcome should be dictated by chance rather than by inference. In a toss of coin the occurrence of head or tail is governed by chance rather than by previous outcomes. Secondly, are the outcomes of the variable being counted or measured. An automobile repair workshop encounters number of cars daily. Assigning probability to a randomly selected car being a foreign made car requires *counting* of cars whereas assigning probability to a potato chip pack to be underweight requires *measuring* the weight of potato chip pack.

Depending on the type of data various probability distribution can be categorized into discrete and continuous random distributions. ***Discrete random probability distributions*** involve distributions of those variables the outcome of which can be counted and probability is determined for such outcomes like number of one rupee coins in a bag, number of defective parts in a sample of produced parts, etc. The random variable which can be counted called discrete variable can occur over a sample like finding defective parts in a sample of 20 parts or getting heads when experiment of tossing a coin is conducted for five times. In these cases number of defective parts or number of heads are counted. In the next step probability of obtaining defective parts in the given sample or probability of getting heads is obtained. The random variable can also occur over a given interval. For instance, average number of foreign made cars coming to car repair store in a ten minute interval or average number of chocolate chips in a chocolate cookie. In these cases, time and a cookie is an interval. In the next step probability of getting a certain number of foreign made of cars in the given interval or probability of getting certain number of chocolate chips in the given interval of cookie is obtained.

Continuous random distributions involve distributions of variables which can be measured like time, weight, height, volume etc. It is difficult and infeasible to determine exact amount of sugar in one kilogram of sugar. The weight can be slightly higher or lower than one kilogram. So, probability of the interval in which weight of one kilogram of sugar is computed rather than probability of exact outcome. Probability is not measured in terms of success or failure but whether the outcome is within the given interval or not. For example, determining probability that a student would score in a given interval of percentage would result in assigning him first division or otherwise. In another example, in an army recruitment drive if average height for selection is six feet, then it would be difficult and infeasible to find a candidate with exact height of six feet and ultimately assigning probability to such an event. In this case the experiment is picking of a candidate and random outcome be the height of the selected outcome. Now, height is a measurable data and assigning probability to an exact height outcome would be infeasible. So, it would be more appropriate and feasible to determine probability that a

candidate has a height within an interval of five feet eight inches and six feet two inches. Now, if probability is to be found for more than one candidate then it would result in formulation of continuous probability distribution.

This chapter discusses two kind of discrete probability distributions: Binomial and Poisson, and one continuous probability distribution: normal distribution.

Exercise 1

1. Which one of these variables is a continuous random variable?
   i. The time it takes a randomly selected student to complete an exam.
   ii. The number of tattoos a randomly selected person has.
   iii. The number of women taller than 68 inches in a random sample of 5 women.
   iv. The number of correct guesses on a multiple choice test.
2. Continuous random variables are obtained from data that can be measured rather than counted.

(i) True (ii) False

3. Discrete variables have values that can be measured.

(i) True (ii) False

4. A(n)_____is one in which values are determined by chance.
5. A(n)_____probability distribution consists of the values in which a random variable can assume the corresponding probabilities of the values.

14.2 Binomial distribution

Binomial distribution is a kind of discrete random distribution which involves assigning probabilities to the experiments which can be counted and have only two outcomes. For instance, if the experiment is appearing in an exam then the candidate can have only two outcomes i.e. pass or fail. In an experiment of tossing of a coin the outcome can only be head or tail. Both outcomes cannot occur simultaneously.

14.2.1 Characteristics of Binomial Distribution:

- *The experiment involves n identical trials*

  This means that if throwing a dice is an experiment and a sample of five trials is taken then the experiment should involve only throwing of the dice. The trials should not be different from each other.

- *Each trial has only two possible outcomes as success or failure*

  As the experiment is conducted on the basis of its outcome being random, so an experiment can have only outcomes in terms of either a success or a failure.

- *Each trial is independent of the previous trials*

For instance, in tossing of coin experiment if an outcome is head in the first trial, then it does not have any impact on the occurrence of head in succeeding trials. Thus, in binomial distribution the trials are

mutually exclusive and independent where two trials cannot occur simultaneously and their outcomes are independent of each other. This constraint is possible in cases where the experiment by nature produces independent trials such as tossing of coin, rolling of dice, result of an exam etc. or the experiment is conducted with replacement. For instance if of all items in a bin 5% are known to be defective then probability of finding a defective item when one is withdrawn from the bin is p = 0.05. If this item is not replaced in the bin and second item is withdrawn from the bin then probability of getting a defective item would be different from the first trial as sample size or the number of limited items in the bin has changed. The binomial distribution would not be applicable in such cases as it does not allow probability of an outcome to change from trial to trial. This condition can be relaxed if population is very high. For instance in a potato chip manufacturing facility number of chips produced daily is enormous. So to check its quality a small sample is taken. In such cases even without replacement the independence assumption is generally met.

- *The probability of success denoted by p and that of failure denoted by q =1-p remains constant for every trial.*

This implies that probability of occurrence of one outcome in a trial is independent of occurrence of same outcome in another trial. For instance, probability of getting head in one trial is 0.5, which would be the same incase of other trials.

Some of the examples of binomial distribution are:

- Suppose a vending machine has an error rate of 5%. If 50 individual use this machine, what is the probability that less than 2 individuals encounter error?
- A company promises to deliver pizzas within half an hour. The probability of failure is 10%. To check the success of the policy the manager decides to take a sample of 20 deliveries. What is the probability that less than one delivery failed?
- After logging on to amazon.com the probability of buying is 20%. What is the probability that out of 100 people logging at a particular time less than four people would not buy?

14.2.2 Computation of Binomial probability problem

This would be illustrated by conducting an experiment of tossing a coin three times i.e. number of trials denoted by n = 3. There can be only two possible outcomes, head (H) or tail (T). The outcome of random variable is denoted by X. The probability of getting a head is termed as success and is denoted by p whereas getting tail is failure and is denoted as q = 1-p. For illustration, if trial is conducted for the first time i.e. n=1, the outcome can be head or tail. If the outcome is head with probability p, then second trial is conducted (n=2) and outcome can be head or tail. If outcome is head with probability p, then third trial is conducted (n=3) and outcome can be head or tail. If outcome is head then for three trials the probability would be p*p*p. Similarly, probability for all the cases can be obtained which is shown in table 1.

Table 1

Events	Probability	No. of successes
HHH	p*p*p	3
HHT	p*p*(1-p)	2
HTH	p*(1-p)*p	2
HTT	p*(1-p)*(1-p)	1
THH	(1-p)*p*p	2
THT	(1-p)*p*(1-p)	1
TTH	(1-p)*(1-p)*p	1
TTT	(1-p)*(1-p) *(1-p)	0

Probability of occurring of any event from above illustration can be generalized as

$P(X) = p^x(1-p)^{n-x}$

For instance, probability of 2 successes i.e. probability of getting head twice in three trials implies that when, p=0.5

$P(X=2) = 0.5^3 * (1-0.5)^{3-2}$

But according to table 2 successes can be obtained three times. Thus, final formula in case of binomial distribution becomes

$$\mathbf{P(X) = {}^nC_x * p^x(1-p)^{n-x}}$$

Where C is combination calculated by using formula n! / (x!)*(n-x)!

Example 1: According to government figures 6% of all workers in a city are unemployed. While conducting a small survey for the city what is the probability of getting two or fewer unemployed workers in a sample of 20?

Solution:

Two or fewer unemployed employees indicate 0, 1 or 2. Thus solution becomes

$P(X <=2) = P(X=0) + P(X=1) + P(X=2)$

$= {}^{20}C_0 * 0.06^0(1-0.06)^{20-0} + {}^{20}C_1 * 0.06^0(1-0.06)^{20-1} + {}^{20}C_2 * 0.06^2(1-0.06)^{20-2}$

$= 0.8850$

14.3 Poisson Distribution

Poisson distribution is another discrete random distribution involving assigning probabilities to countable outcomes over an interval. It is different from binomial distribution because here number of trials is not given whereas in binomial distribution trials happen over a given sample space. For instance, number of erroneous calls over a given number of calls is a binomial distribution, whereas number of erroneous calls over a given interval of say five minutes is a Poisson distribution. Another difference between two distributions is that in Poisson distribution instead of sample space events happen over an interval. For instance, number of defects per carpet, number of chocolate chips per cookie, number of customers coming in a bank during lunch hour etc.

14.3.1 Characteristics of Poisson Distribution

- *Occurrence of success in any one interval is independent of that in any other interval.*

For example, number of chips in a cookie is independent of those in another cookie; arrival of customer during an interval is not dependent on arrival of another customer in other interval; a customer passing through a security gate from another customer passing through the same gate at different point in time.

- *Probability of observing more than one success in any one interval is zero.*

This implies that if happening of an event over a continuum ranges from zero to infinity then that continuum should be broken down to smaller intervals in such a way that only one event should happen in that interval. For example, number of customers coming in a bank during lunch hour is a discrete event. Now, suppose on an average 180 customers arrive during one hour. In this case the interval is big i.e. of one hour so probability of observing more than one success i.e. arrival of a customer is different from zero. So, interval is broken down into shorter duration say in seconds. In one second chance of arrival of more than one customer is very rare. Thus, the occurrence of more than one success in that interval is zero.

- *Probability of success in an interval is same for other intervals.*

This implies intervals to be mutually exclusive i.e. events cannot occur simultaneously. For instance, by taking same example when interval of one hour is broken down into 3600 seconds then probability of arrival of one customer out of 180 would be 180/3600 = 0.05. As only one customer can arrive per interval of one second so probability of arrival of another customer in second interval of one second would be same i.e. 0.05. A corollary of this characteristic is that probability of occurrence of an event is interval dependent. If interval is of one second then probability is 0.05 but if interval is of two seconds then probability of occurrence would also be doubled to 0.1 as now two customers might arrive during increased interval.

Some of examples of Poisson distribution are:

- Number of accidents per hour.
- Number of wrong delivery of newspapers per morning.
- Number of customers coming in a restaurant every five minutes on a busy Saturday night.
- Number of defective pens per carton.
- Number of golf players in a small city.
- Number of people having a rare disease in population of ten lakh.

14.3.2 Computation of Poisson Distribution problem

If a Poisson distribution phenomenon is studied over a long period of time, then a long run average of that event can be estimated denoted by **lambda (λ).** A binomial distribution required n and p to describe occurrence of discrete variables, whereas a Poisson distribution can be described by lambda. The formula to calculate probability of occurrences of a discrete variable over a given interval is

$$P(X) = e^{-\lambda} * \lambda^{x} / x!$$

where λ = long run average

e = exponential constant = 2.718

x = number of occurrences per interval for which the probability is being computed.

Example 2: Number of calls received by an operator between 9 and 10 am has a Poisson distribution with a mean of 12. What is the probability that an operator received at least 5 calls during:

(i) 9 and 10 am

(ii) 9 and 9:30 am

(iii) 9 and 9:15 am

Solution:

(i) $P(X>=5) = 1 – [P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)]$

Solving for $P(X=0) = 2.178^{-12} * 12^0 / 0!$

Similarly probability for x = 1, 2, 3 and 4 can be found.

(ii) As discussed earlier lambda is dependent on interval duration. In this case interval duration is halved so value of long run average i.e. average number of events happening in that interval will also be halved. Thus, $\lambda = 6$

(iii) In this case $\lambda = 4$

Calculations are left as exercise for student.

Exercise 2

1. A medical treatment has a success rate of .8. Two patients will be treated with this treatment. Assuming the results are independent for the two patients, what is the probability that neither one of them will be successfully cured?
   (i) 0.5 (ii) 0.36 (iii) 0.2 (iv) *.04 (this is (1 – .8)(1 – .8) = (.2)(.2) = .04)*
2. The probability of a success must remain the same for each trial in a binomial experiment.
   (i) True (ii) False
3. In binomial experiments, the outcomes are usually classified as successes or failures.
   (i) True (ii) False
4. In a binomial experiment, the outcomes of each trial must be dependent on each other.
   (i) True (ii) False
5. When sampling is done without replacement, the binomial distribution does not give exact probabilities because the trials are not independent.
   (i) True (ii) False
6. A coin is tossed five times. Find the probability of getting exactly three heads.
   (i) 0.3750 (ii) 0.1563 (iii) 0.2500 (iv) 0.3125
7. If a student randomly guesses at 20 multiple-choice questions, find the probability that the student gets exactly four correct. Each question has four possible choices.

(i) 0.19 (ii) 0.17 (iii) 0.08 (iv) 0.23

8. One of the requirements for a binomial experiment is that there must be a_____number of trials.
9. The Poisson distribution is used when n is small and p is large.

   (i) True (ii) False

10. Which one of these variables is a binomial random variable?

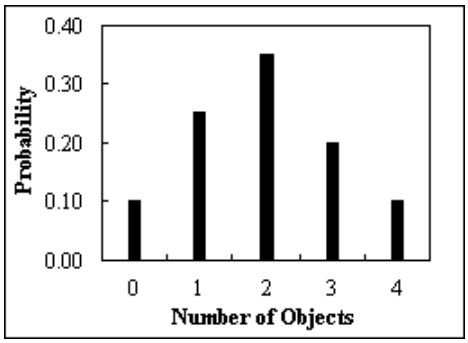
    (i) time it takes a randomly selected student to complete a multiple choice exam

    (ii) number of textbooks a randomly selected student bought this term

    (iii) number of women taller than 68 inches in a random sample of 5 women

    (iv) number of CDs a randomly selected person owns

11. The figure below represents the probability distribution for selecting a number of objects out of a container. Construct a probability distribution from this graph.

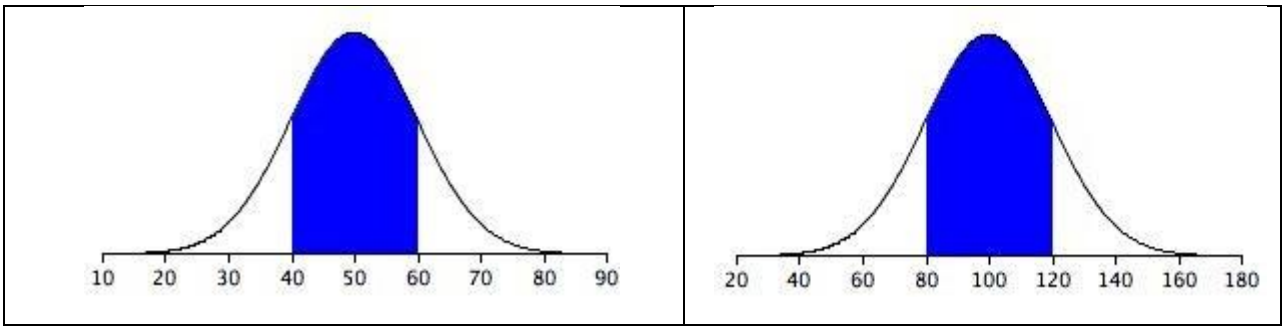


14.4 Continuous probability distribution: Normal Distribution

Normal distribution is the most applied distribution because of its use in various decision making processes. The distribution is appropriate for assigning probability for the occurrence of continuous data. The data which is not counted but measured like, height, weight, length, speed etc. are put under normal distribution. Many researchers use scaling techniques like Likert scale to measure customers' satisfaction or any other behavioral characteristic. These variables are continuous in nature as it measures an individuals' behavior. The data regarding these variables follow normal distribution making it easy for decision maker to reach an inference.

Graphically Normal distributions are represented by bell shaped curve as shown in Fig. 1. Figure 1 shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.

Fig. 1	
Normal distribution with a mean of 50 and standard deviation of 10.	A normal distribution with a mean of 100 and a standard deviation of 20.



The normal distribution shown in Figure 1 is specific example of the general rule that 68% of the area of any normal distribution is within one standard deviation of the mean.

14.4.1 Characteristics of normal distribution:

- It is a continuous distribution.
- It is a symmetrical distribution about its mean implying that each half of the distribution is the mirror image of the other half.
- It is unimodal. The graph has only one peak indicating that in the continuous data there is only one value which has highest frequency.
- Area under the curve is one.

The normal distribution bell shaped curve is described by two parameters: mean and standard deviation. For each value of mean and standard deviation the graph would carry a different shape. If mean of one graph is more and standard deviation is less than other then it would have higher peak and narrow width. Whereas, high standard deviation spreads the graph wide. For a perfect normally distributed graph mean would be equal to zero with one standard deviation.

Because of different normal distribution with different value of mean and standard deviation the generalized or standardized formula for normal distribution is given by z distribution.

$$z = (x - \mu) / \sigma$$

where x depicts the value of data for which probability in comparison to mean value has to be found. If the value of x is less than mean the z score would be negative and if x value is more than mean then z score would be positive. If x value is equal to mean value then z score would be equal to zero implying that the selected data (x) has zero deviations from the mean.

Example 3: In case of GMAT test with a mean of 494 and a standard deviation of 100, what is the probability that a candidate would score

(i) greater than 700

(ii) equal to or less than 550

(iii) between 300 and 600

Solution:

(i) the z score for this problem is:

$z = (700 - 494) / 100$

$= 2.06$

From z tables this value of z gives the probability as 0.4803. But this probability is for the area between mean value and 700. The question asks for probability greater than 700. As we know that in case of normal distribution the area of curve is 1 and mean divides the curve into two equal halves. The probability for one half is 0.5 So, the required probability is

= 0.5 (probability of x greater than mean) – 0.4803 (probability of x between mean and 700)

= 0.0197 (probability of greater than 700)

This means, that probability that a candidate would score greater than 700 would be 1.97%.

(ii) the z score for this problem is:

$z = (550 - 494) / 100$

$= 0.56$

The associated probability of x between mean 494 and 550 from the z tables is 0.2123. but the question asks for finding the probability that a candidate will score less than 550. So the required probability is:

= 0.5 (probability of x less than mean) + 0.2123 (probability of x between 550 and mean)

= 0.7123

This means, that probability that a candidate would score less than 550 would be 71.23%

(iii) the z score for this problem is calculated in two parts:

between 300 and mean:

$z = (300-494) / 100$

$= -1.94$

The associated probability from z tables for x between 300 and 494 is 0.4738.

Between mean and 600

$z = (600-494) / 100$

$= 1.06$

The associated probability for x between 600 and 494 is 0.3554.

So, the required probability is

= 0.4738 (probability of x between 300 and 494) + 0.3554 (probability of x between 494 and 600)

= 0.8292

This means, that probability of a candidate scoring between 300 and 600 would be 82.92%.

Example 4: An agency publishes data on solid waste generation. One year the average number of waste generated per person per day was 3.58 kgs. Suppose the daily amount of waste generated per person is normally distributed,

with a standard deviation of 1.04 kgs. Of the daily amounts of waste generated per person, 67.72% would be greater than what amount?

Solution: the mean and standard deviation are given but x and z are unknown. The problem is to determine specific x value when 0.6772 of the x values are greater than that value. If 0.6772 values are greater than x then 0.6772-0.5000 = 0.1772 are between x and the mean value. According to z table the value of z for 0.1772 is 0.46. because x is less than the mean, the z value actually is -0.46.

Solving the z equation gives:

$z = (x - \mu) / \sigma$

$-0.46 = (x - 3.58) / 1.04$

Thus, $x = 3.10$

So, 67.72% of the daily average amount of solid waste per person weighs more than 3.10 kgs.

Exercise 3

1. Heights of college women have a distribution that can be approximated by a normal curve with a mean of 65 inches and a standard deviation equal to 3 inches. About what proportion of college women are between 65 and 67 inches tall?

   (i) 0.75 (ii) 0.50 (iii) 0.25 (iv) 0.17
2. Suppose that vehicle speeds at an interstate location have a normal distribution with a mean equal to 70 mph and standard deviation equal to 8 mph. What is the z-score for a speed of 64 mph?

   (i) −0.75(ii) +0.75(iii) −6 (iv) +6
3. Pulse rates of adult men are approximately normal with a mean of 70 and a standard deviation of 8. Which choice correctly describes how to find the proportion of men that have a pulse rate greater than 78?

   (i) Find the area to the left of z = 1 under a standard normal curve.

   (ii) Find the area between z = −1 and z = 1 under a standard normal curve.

   (iii) Find the area to the right of z =1 under a standard normal curve.

   (iv) Find the area to the right of z = −1 under a standard normal curve.

14.5 Summary

Probability distributions are important to understand the characteristic of data which would help a researcher to apply different statistics tools. This chapter discusses three probability distributions: binomial, Poisson and normal distributions. The application of these distributions depends on type of data under consideration. Binomial and Poisson distributions are applied where data is countable over a sample space or under given interval. Normal distribution is a continuous distribution where data is measurable like volume, height, consumer behavior etc. The chapter understands these distributions by illustrating certain numerical problems.

14.6 Glossary

- **Discrete distributions:** the probability distributions which involve countable data like, number of 10 Rs. notes etc.
- **Continuous Distributions:** the probability distributions which involve measurable data like height, weight, volume etc.
- **Binomial distribution:** is a kind of discrete distribution in which there are only two possible outcomes on the occurrence of an experiment. It involves specific number of experiments. For instance, determining probability of certain number of defective products from a specific number of manufactured products.
- **Poisson distribution:** is also a discrete distribution where events happen over an interval which could be time or space.
- **Normal Distribution:** it is a kind of continuous distribution which involves measurable data such as many human behaviors and timing or capacity of machines. This is most widely used distribution as many statistical applications are based on normally distributed data.

14.7 Answers to check your progress/ Self assessment exercise

Exercise 1

1. i
2. i
3. ii
4. random variable
5. discrete

Exercise 2

1. iv
2. i
3. i
4. ii
5. i
6. d
7. a
8. fixed
9. b
10. c
11.

No. of objects, x	0	1	2	3	4

P(x)	0.10	0.25	0.35	0.20	0.10

Exercise 3

1. iii
2. i
3. iii

14.8 References/ Suggested Readings

- Black, K., *Business Statistics For Contemporary Decision Making,* Fifth Edition, Wiley India,
- Keller, G., *Statistics for Management,* First India Reprint 2009, Cengage Learning India Private Limited.
- Stevenson, W.J., *Operations Management,* Ninth Edition, Tata McGraw Hill, New Delhi, 2009.
- Donald R. Cooper & Pamela S. Schindler, *Business Research Methods,* Tata McGraw-Hill Publishing Co. Ltd., New Delhi, 9th Edition.
- S.P. Gupta, *Business Statistics*, Sultan Chand, New Delhi.

14.9 Terminal and Model Questions

1. How binomial and Poisson distributions similar and different?
2. Explain discrete and continuous distributions.
3. According to a survey by a consumer magazine 60% of all consumers have dialed an '100' or '200' telephone number for information about some product. Suppose a random sample of 25 consumers is contacted about their buying habits
   (i) What is the probability that 15 or more of these consumers have called '800' or '900' telephone number for information about some product?
   (ii) What is the probability that fewer than 10 of these consumers have called '800' or '900' telephone number for information about some product?
4. The average number of annual trips per family to amusement parks is Poisson distributed with a mean of 0.6 trips per year. What is the probability of randomly selecting a family and finding:
   (i) The family did not make a trip to an amusement park last year?
   (ii) The family took exactly one trip to an amusement park last year?
   (iii) The family took two or more trips to an amusement park last year?
5. According to a report the average monthly household cellular phone bill is $60. Suppose local monthly household cell phone bills are normally distributed with a deviation of $11.35. What is the probability that a randomly selected monthly phone bill is
   (i) More than $85.
   (ii) Between $45 and $70

(iii) Between $65 and $75

(iv) No more than $40.