BUSINESS STATISTICS

(DBBS21)

(BACHELOR OF BUSINESS ADMINISTRATION)

ACHARYA NAGARJUNA UNIVERSITY

CENTRE FOR DISTANCE EDUCATION

NAGARJUNA NAGAR,

GUNTUR

ANDHRA PRADESH

Lesson - 1

NEED FOR INFORMATION IN DECISION MAKING - PRIMARY VS SECONDARY DATA

Objective

After going through this lesson you will learn:

- The concept of business decisions.
- The concept of primary data and secondary data their uses, advantages and disadvantages.

Structure

1.1 Introduction

When it comes to growing your business, information is vital oxygen that helps you ascend to great heights.

Decision - making is a crucial part of good business. One part of the answer is good information, and experience in interpreting information. Consultation is seeking the views and expertise of other people also helps, as does the ability to admit one was wrong and change one's mind. There are also aids to decision - making, various techniques which help to make information clearer and better analyzed and to add numerical and objective precision to decision - making (where appropriate) to reduce the amount of subjectivity.

Decision - making increasingly happens at all levels of a business. The Board of Directors may make the grand strategic decisions about investment and direction of future growth and managers may make the more tactical decisions about how their own department may contribute most effectively to the overall business objectives. But quite ordinary employees are increasingly expected to make decisions about he conduct of their own tasks, responses to customers and improvements to business practice. This needs careful recruitment and selection, good training and entightened management.

1.2 Types of Business Decisions

1. **Programmed Decisions:** These are standard decisions which always follow the same routine. As such, they can be written down into a series of fixed steps which any one can follow. They could even be written as computer program.

2. **Non - Programmed Decisions:** These are non - standard and non - routine. Each decision is not quite the same as any previous decision.

3. **Strategic Decisions:** These affect the long- term direction of the business eg whether to take over Company A or Company B.

4. **Tactical Decisions:** These are medium - Term decisions about how to implement strategy eg what kind of marketing to have, or how many extra staff to recruit.

5. **Operational Decisions:** These are short - term decisions (also called administrative decisions) about how to implement the tactics eg which firm to use to make deliveries.

6. **Linear Programming:** Linear Programming models help to explore maximizing or minimizing constraints eg. One can program a computer with information that establishes parameters for minimizing costs subject to certain situations and information about those situations.

7. **Spread - Sheets are Widely used for 'What If' Simulations:** A very large spread - sheet can be used to hold all the known information about, say pricing and the effects of pricing on profits. The different pricing assumptions can be fed into the spread - sheet 'modelling' different pricing strategies. This is a lot quicker and an awful lot cheaper than actually changing prices to see what happens. On the other hand, a spread - sheet is only as good as the information put into it and no spread - sheet can fully reflect the real world. But it is very useful management information to know what might happen to profits 'what if' a skimming strategy or a penetration strategy were used for pricing.

The computer does not take decisions ; managers do. But it helps managers to have quick and reliable quantitative information about the business as it is and the business as it might be in different sets of circumstances. There is, however, a lot of research into 'expert systems' which aim to replicate the way real people (doctors, lawyers, managers and the like) take decisions. The aim is that computers can one day take decisions or at least programmed decisions. For example, an expedition could carry an expert medical system on a lap - top to deal with any medical emergencies even though the nearest doctor is thousands of miles away. Already it is possible, to put a credit card into a 'hole - in - the - wall' machine and get basic legal advice about basic and standard legal problems.

1.2.1 Quality of Decision - Making: Some managers and business make better decisions than others. Good decision - making comes from:

1. Training of managers in decision - making skills. See Developing Managers.
2. Good information in the first place.
3. Management skills in analyzing information and handlong its shortcomings.
4. Experience and natural ability information and handling its shortcomings.
5. Risk and attitudes to risk.
6. Human factors. People are people. Emotional responses come before rational responses, and it is very difficult to get people to make rational decisions about things they feel very strongly about. Rivalries and vested interests also come into it. People simply take different views on the same facts and people also simply make mistakes.
7. **Interdependence:** Businesses are highly interdependent each other, their suppliers and their customers. Decisions are not taken in isolation. The effects of any decision will depend critically on the reactions of other groups in the market.

1.2.2 A Customer Can Obtain Information From Several Sources:

1. **Personal Sources:** Family, Friends, Neighbors etc....
2. **Commercial Sources:** Advertising, Salespeople, Retailers, Dealers, Packaging, Point - of - sale displays.
3. **Public Sources:** News Papers, Radio, Television, Consumer Organizations, Specialist Magazines.
4. **Experiential Sources:** Handling, Examining, Using the product.

The usefulness and influence of these sources of information will varey by product and by customer. Research suggests that customer's value and respect personal sources more than commercial sources (the influence of "word of mouth"). The challenge for the marketing team is to identify which information sources are most influential in their target markets.

In the evaluation stage, the customer must choose between the alternative brands, products and services.

1.2.3 What Types of Data and / or Information are Needed?

The specific types of information and/or data needed to conduct a secondary analysis will depend, obviously, on the focus of your study. For CARE purposes, secondary data analysis is usually conducted to gain a more in depth understanding of the food and livelihood security status of people living in various countries regions where CARE works. This type of socioeconomic characterization is commonly referred to as a country or poverty profile. It involves collecting information, statistics and other relevant data at various levels of aggregation in order to conduct a situational analysis of the area the types of secondary data and information to be collected and summarized should include data related to the following areas:

- Demographic (population, population growth rate rural / urban, ethnic groups, migration trends etc.,)
- Agroecological climatic zones.
- Poverty levels (poverty and absolute)
- Employment and wages (formal and informal)
- Livelihood systems (rural, urban, on - farm, off - farm, informal, etc.,)
- Agricultural variables and practices (rainfall, crops, soil types and uses irrigation etc.,)
- Health (malnutrition, infant mortality, immunization rate, fertility rate, contraceptive prevalence rate etc.,)
- Health Services (#/level, services by level, facility - to - population ratio etc.,)
- Education (adult literacy rate, school enrollment, drop - out rates, male - to - female ratio etc.,), Schools (#/level, school - to - population ratio etc.,)
- Infrastructure (roads, electricity, communication, water, Sanitation etc.,)
- Environmental status and problems and
- Local political environment and access

Special attention should be given to collecting disaggregated data. That is data that is broken down in the following ways; Gender, Age, Ethnicity, Location, etc.,

Even when highly disaggregated, however these "raw" data points alone are often only static or indirect measures of the situation or problems that exist in countries and regions - partial or imperfect reflections of reality. It is through interpretation and analysis that these pieces of information allow us to gain a better understanding of a specific situation population sector etc.,

Analysis of data gives you the information that you need to make judgements, recommend areas of intervention design follow - up studies. Cross analyzing data will also help you understand not only what is happing in a particular area but also why it is happening.

1.3 Data

What are Data?

Data may be considered to be one of the vital fluids of modern civilization. Data are used to make decisions, to support decisions already made, to provide reasons why certain events happen and to make predictions on events to come.

Definition of DATA

The word DATA is defined as things known or assumed facts and figures from which conclusions can be inferred. Broadly data is raw information and this can be qualitative as well as quantitative. The source can be anything from hearsay to the result of elegant and painstaking research and investigation. The terms of reporting can be descriptive, numerical or various combinations of both. The transition from data to knowledge may be considered to consist of the hierarchal sequence.

$$\text{Data} \xrightarrow{\text{analysis}} \text{Information} \xrightarrow{\text{model}} \text{knowledge}$$

Ordinarily, some kind of analysis required to convert data into information. A model is typically required to interpret numerical information to provide knowledge about a specific subject of interest. Also, data may be acquired, analyzed and used to test a model of a particular problem.

Statistical Techniques For Data Analysis

Data producers have the obligation to present all pertinent information that would impact on the use of it, to the extent possible. Often, they are in the best position to provide such background information and they may be the only source of information on these matters. When they cannot do so, it may be a condemnation of their competence as metrologists. Of course every possible use of data cannot be envisioned when it is produced, but the details of its production its limitations and quantitative estimates of its reliability always can be presented. Without such data can hardly be classified as useful information.

Users of data cannot be held blameless for any misuse of it, whether they may have been misled by its producer. No data should be used for any purpose unless their reliability is verified. No matter how attractive it may be unevaluated data are virtually worthless and the temptation to use them should be resisted. Data users must be able to evaluate all data that the utilize or depend on reliable sources to provide such information to them.

Why should a marketer need to understand the customer evaluation process?

The answer lies in the kind of information that the marketing team needs to provide customers in different buying situations.

In high - involvement decisions, the marketer needs to provide a good deal of information about the positive consequences of buying. The sales force may need to stress the important attributes of the product, the advantages compared with the competition and may be even encourage 'trial' or 'sampling' of the product in the hope of securing the sale.

Post - Purchase Evaluation - Cognitive Dissonance

The final stage is the post purchase evaluation of the decision. It is common for customers to experience concerns after making a purchase decision. This arises from a concept that is known as "cognitive dissonance". The customer, having bought a product may feel that an alternative would have been preferable. In these circumstances that customer will not repurchase immediately, but is likely to switch brands next time.

To manage the post - purchase stage it is the job of the marketing team to persuade the potential customer that the product will satisfy his or her needs. Then after having made a purchase, the customer should be encouraged that he or she has made the right decision.

1.3.1 Primary Data: Primary data is data that you collect yourself using such methods as:

Direct Observation: lets you focus on details of importance to you, lets you see a system in real rather than theoretical use.

Surveys: Written surveys let you collect considerable quantities of detailed data. You have to either trust the honesty of the people surveyed or build in self - verifying questions.

Interviews: Slow, expensive and they take people away from their regular jobs, but they allow in depth questioning and follow - up questions. They also show non - verbal communication such as face - pulling, fidgeting, shrugging hand gestures, sarcastic expressions that add further meaning to spoken words e.g. "I think it's a GREAT system" could mean vastly different things depending on whether the person was sneering at the time! A problem with interviews is that people might say what they think the interviewer wants to hear ; they might avoid being honestly critical in case their jobs or reputation might suffer.

Logs: (e.g., fault logs, error logs, complaint logs, transaction logs). Good, empirical, objective data sources (usually if they are used well). Can yield lots of valuable data about system performance over time under different conditions.

Primary data can be relied because you know where it came from and what was done to it. It's like cooking something yourself. You know that went into it.

1.3.2 Secondary Data: Secondary data analysis can be literally defined as second - hand analysis. It is the analysis of data of information that was either gathered by someone else (e.g. researchers, institutions, other NGO's etc.) or for some other purpose than the one currently being considered or often a combination of the two.

If secondary research and data analysis is undertaken with care and diligence, it can provide a cost - effective way of gaining a broader understanding of specific phenomena conducting preliminary needs assessments.

Secondary data are also helpful in designing subsequent primary research and as well can provide a base line with which to compare your primary data collection results. Therefore, it is always wise to begin any research activity with a review of the secondary data.

1.3.1 Sources of Secondary Data

Official Statistics: Official statistics are statistics collected by governments and their various agencies, bureaus and departments. These statistics can be useful to researchers because they are an easily obtainable and comprehensive source of information that usually covers long periods of time.

However, because official statistics are often "characterized by unreliability data gaps, over - aggregation, inaccuracies, mutual inconsistencies and lack of timely reporting". It is important to critically analyze official statistics for accuracy and validity. There are several reasons why these problems exist:

1. The scale of official surveys generally requires large numbers of enumerators (interviewers) and in order to reach those numbers enumerators contracted are often under - skilled.
2. The size of the survey area and research team usually prohibits adequate supervision of enumerators and the research process and
3. Resource limitations (human and technical) often prevent timely and accurate reporting of results.

Technical Reports: Technical reports are accounts of work done on research projects. They are written to provide research results to colleagues, research institutions, governments and other interested researchers. A report may emanate from completed research or on - going research projects.

Scholarly Journals: Scholarly journals generally contain reports of original research or experimentation written by experts in specific fields, Articles in scholarly journals usually undergo a peer review where other experts in the same field review the content of the article for accuracy, originality and relevance.

Literature Review Articles: Literature review articles assemble and review original research dealing with a specific topic. Reviews are usually written by experts in the field and may be the first written overview of a topic area. Review articles discuss and list all the relevant publications from which the information is derived.

Trade Journals: Trade journals contain articles that discuss practical information concerning various fields. These journals provide people in these fields with information pertaining to that field or trade.

Reference Books: Reference books provide secondary source material. In many cases, specific facts or a summary of a topic is all that is included. Handbooks, manuals, encyclopedias and dictionaries are considered reference books.

1.3.2 Where to find Secondary Data: There are numerous sources of secondary data and information. The first step in collecting secondary data is to determine which institutions conduct research on the topic area or country in question.

Large surveys and country wide studies are expensive and time consuming to conduct therefore, they are usually done by governments or large institutions with a research

orientation. Thus, government documents and official statistics are a good starting place for gathering secondary data, however as previously stated, the quality of the documents will vary depending on the country of study and the amount of resources dedicated to data collection.

Other major sources of international development data are the World Bank, the United States Agency for International Development (USAID), the United Nations Development Programme (UNDP), the Food and Agriculture Organization of the United Nations (FAO) the International Fund for Agriculture Development (IFAD), and the World Health Organization (WHO) to name a few.

International development institutes commonly share information sources and have libraries for archiving these materials. Thus, a data - gathering visit to one office might yield numerous sources of information on the topic area of interest.

University libraries are good sources of information and should be consulted. Also, it would be beneficial to establish contact with experts at local university departments that are dedicated to research on the topic areas that you are interested in (eg., Departments of Agricultural Sciences, Public Health, Economics, Anthropology and Sociology). These experts can be important sources of information on on-going research projects as well as for guiding you toward other sources of topic area information or individuals that can be contacted.

Local NGOs also often conduct empirical research and can be valuable sources of information. This, in particularly true when you are searching for regional or local - level information and data. In some cases, they might also have small libraries that provide additional information.

1.3.3 Evaluating The Quality of Your Information Sources: One of the advantages of secondary data review and analysis is that individuals with limited research training or technical expertise can be trained to conduct this type of analysis. Key to the process, however, is the ability to judge the quality of the data or information that has been gathered. The following tips will help you assess the quality of the data.

Determine The Original Purpose of the Data Collection: Consider the purpose of the data or publication. Is it a government document or statistic, data a collected for corporate marketing purposes, or the output of a source whose business is to publish secondary data (e.g. research institutions). Knowing the purpose of data collection will help to evaluate the quality of the data and discern the potential level of bias

Attempt to Ascertain the Credentials of the Source(s) or Author(s) of the Information:

What are the author's or source's credentials ---- educational background past works/ writings or experience --- in this area? For example, the following sources are generally considered reliable sources of data and information ; research reports documenting findings from agricultural research published by the FAO or IFAD socioeconomic data reported by the World Bank; and survey health data reported in USAID's Demographic Health Surveys.

Does it include a methods section and are the methods sound? Does the article have a section that discusses the methods used to conduct the study? If it does not, you can

assume that it is a popular audience publication and should look for additional supporting information or data. If the research methods are discussed review them to ascertain the quality of the study. If you are not a research methods expert, have someone else in your Country Office review the methods section with you.

What's the Data of Publication? When was the source published? If the source current or out of date? Topic areas of continuing or rapid development, such as the sciences, demand more current information.

Who is the Intended Audience? Is the publication aimed at a specialized or a general audience? Is the source too elementary aimed at the general public?

What is the Coverage of the Report or Document? Does the work update other sources, substantiate other materials / reports that you have read or add new information to the topic area?

Is it a Primary or Secondary Source? Primary sources are the raw material of the research process, they represent the records of research or events as first described. Secondary sources are based on primary sources. These sources analyze, describe and synthesize the primary or original source. If the source is secondary does it accurately relate information from primary sources?

Importantly, Is the Document or Report Well - Referenced? When data figures are given are they followed by a footnote, end note --- which provides a full reference for the information at the end of the page or document -- or the name and date of the source Without proper reference to the source of the information it is impossible to judge the quality and validity of the information reported.

1.3.4 Do The Numbers Do Not Make Sense? Data reporting characteristics vary according to what the data is being collected for and the stage of reporting. For example, health clinics might report quarterly the number of cases of diarrhea, upper respiratory infection or malnutrition that they have been treated at a clinic.

This information is useful for health care professionals who will later analyze the information to ascertain the percentage of the population in a municipality or province that were diagnosed with these problems over a given period of time for the purpose of secondary data analysis the aggregated percentage figure, rather than the number of "causes" reported should be used.

Another area of data analysis that requires a skeptical eye is employment related data. It is difficult to count the employed accurately, especially in developing countried. Employment data often do not take into account the number of people involved in informal or unrecorded activities, seasonal agricultural laborers women's agricultural labor or child labor.

Thus official employment statistics should be viewed in light of these inadequacies. Labor force data that provides a list of the categories used (e.g. employed, unemployed, underemployed, own - account workers, unpaid family workers) will help you determine the quality of the measure.

When you feel that the employment data is unreliable, looking at other economic indicators will help you develop a clearer understanding of the situation. For example, if your employment data state that only 25 percent of the population is economically active. However, data from a recent poverty survey state that only 5 percent of the population live below the absolute poverty line, you can conclude that the employment data is not a good measure to use.

1.4 Questions for Data Quality

- What are your source's credentials ?
- What methods were used ?
- Is the information current or out - of - date ?
- Is the intended audience other researchers or the general public ?
- Is the document's coverage of the topic area broad or too narrow ?
- Is it a primary or secondary source ? If it is a secondary source, does it accurately cover and report on the primary sources ?
- Does the author provide references for the data and information reported ?
- Do the numbers make sense ? Are they the numbers you want - cases versus percentages? When compared to related data are the measures somewhat consistent?

Secondary data is generally referred to as out come data. This is because secondary data generally describe the condition or status of phenomena or a group; however, these data alone do not tell us why the condition or status exists. This limitation can be overcome in two ways.

First, it can be overcome by using information from case studies and other research to fill in the gaps. For example, data on child malnutrition rates and women's level of education provide information relevant for understanding why some children are more likely to be malnourished than others. The child health research literature tells us that children whose mothers have low levels of education will likely exhibit higher malnutrition rates than children of mothers with higher levels of education. Thus, consulting relevant literature can help illuminate causal relationships.

Second, analysis of additional key data and indicators can help us acquire more explanation as to why a problem exists. For example, if low farm income has been identified as a problem, data on land size, types of crops, production value, cost of inputs and so on, can be compared to help identify who has this problem and possible causes and solutions.

Therefore, cross - analyzing key indicators and using additional information sources help us understand or make reasonably sound inferences about unmeasured conditions or situations ; thus allowing us to better understand not only what is happening and where it is happening but also why it is happening.

1.5 Using Secondary Information to Strengthens Primary Research

Secondary information is also valuable for generating hypotheses and identifying critical areas of interest that can be investigated during primary data gathering activities. For example, secondary data analysis conducted prior to the Tanzania Urban Food and Household Livelihood Security Assessment, identified key research areas that should be closely studied during the assessment (e.g. the influence of seasonality on urban livelihoods ; the impact of increased privatizing ; gender differentiation in urban land tenure policies ; and social cohesion and locality). Questions were then developed and included in the household, community group, and key informant interview questionnaires to allow analysis of these key areas of interest.

Without through analysis of secondary data, these key constraints to urban food and livelihood security possibly would not have been identified, thus the problem analysis - the why of the primary research exercise - was strengthened through secondary data and information analysis.

1.6 The Advantages and Disadvantages of Secondary Data Review and Analysis:

Advantages:

- Secondary data analysis can be carried out rather quickly when compared to formal primary data gathering and analysis exercises.
- Where good secondary data is available, researchers save time and money by making good use of available data rather than collecting primary data thus avoiding duplication of effort.
- Using secondary data provides a relatively low - cost means of comparing the level of well - being of different political units (e.g., states, departments, provinces, counties). However, keep in mind that data collection methods vary (between researches, countries, departments etc.,) which may impair the comparability of the data.
- Depending on the level of data disaggregation, secondary data analysis lends itself to trend analysis it offers a relatively easy way to monitor change over time.
- It informs and complements primary data collection, saving time and resources often associated with over - collecting primary data.
- Persons with limited research training or technical expertise can be trained to conduct a secondary data review.

Disadvantages:

- Secondary data helps us understand the condition or status of a group, but compared to primary data they are imperfect reflections of reality. Without proper interpretation and analysis they do not help us understand why something is happening.
- The person reviewing the secondary data can easily become overwhelmed by the volume of secondary data available, if selectivity is not exercised.

- It is often difficult to determine the quality of some of the data in question.
- Sources may conflict with each other.
- Because secondary data is usually not collected for the same purpose as the original researcher had, the goals and purposes of the original researcher can potentially bias the study.

1.7 Secondary Versus Primary Data

Secondary data complements but does not replace, primary data collection and should be the starting place for any research activity.

- Because the data were collected by other researchers, and they decide what to collect and what to omit, all of the information desired may not be available.
- Much of the data available are only indirect measures of problems that exist in countries and regions.
- Secondary data can not reveal individual or group values, beliefs or reasons that may be underlying current trends.

1.8 The Importance of Properly Referencing Your Secondary Data Review

Secondary data review and analysis is a form of research and data compilation that is demanding and time - consuming ; however, without proper citation (i.e., author, date title) of materials that you used, your work will often be disregarded as it will only have limited use by those who wish to follow in your footsteps.

- A well - documented secondary data review and analysis allows for easier use of the material by other interested parties.
- Properly citing the publication date of the sources you used will allow subsequent researchers to use your work to make comparisons over time and between countries, communities, towns, regions etc.
- Proper citation allow sub sequent researchers to use your work, thus preventing unnecessary duplication of research efforts.

1.9 Summary

Secondary data can be a valuable source of information for gaining knowledge and insight into a broad range of issues and phenomena. Review and analysis of secondary data can provide a cost - effective way of addressing issues, conducting cross - national comparisons, understanding country - specific and local conditions, determining the direction and magnitude of change trends and describing the current situation. It complements, but does not replace, primary data collection and should be the starting place for any research activity.

1.10 Exercise

1. What is a crucial part of good business?
2. What are the types of Business Decisions?
3. What types of Data and/or Information are needed?
4. What are types of data?
5. What are the sources of secondary data?
6. Where to find the secondary data?
7. How to strengthesis primary research?
8. What are the advantages and disadvantages of secondary data?

1.11 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer

Dr. K. CHANDAN

Lesson - 2

DATA COLLECTION METHODS

Objective:

After going through this lesson, you will learn

- The concept of questionnairs, interviews and observation methods.
- The concept of case-studies, diaries, critical incidents and protfolios.

Structure:

2.1 Introduction:

Data collection the foundation on which the super structure of the statistical analysis is based. Data should be collected in a systematic manner to draw the right conclusion about the characteristic behaviour of the variable. The success of any enquiry depends upon the availability of accurate and reliable data. Reliability of the method choosen for data collection. Therefore data collection in a very basic activity in decision making.

2.2 Questionnaires:

Questionnaires are a popular means of collecting data, but are difficult to design and often require many rewrites before an acceptable questionnaire is produced.

Advantages:

- Can be used as a method in its own right or as a basis for interviewing or a telephone survey.
- Can be posted, e-mailed or faxed.
- Can cover a large number of people or organizations.
- Wide geographic coverage.
- Avoids embarrassment on the part of the respondent.
- Respondent can consider responses.
- Possible anonymity of respondent.
- No interviewer bias.

Disadvantages:

- Design problems.
- Questions have to be relatively simple.
- Historically low response rate (although inducements may help)
- Time delay whilst waiting for responses to be returned.
- Require a return deadline.
- Several reminders may be required.
- Assumes no literacy problems.
- No control over who completes it.
- Not possible to give assistance if required.
- Problems with incomplete questionnaires.
- Replies not spontaneous and independent of each other.
- Respondent can read all questions beforehand and then decide whether to complete or not. For example, perhaps because it is too long, too complex, uninteresting, or too personal.

Design Postal Questionnaires:

Theme and Covering Letter:

The general theme of the questionnaire should be made explicit in a covering letter. You should state who you are; why the data is required; give, if necessary, an assurance of confidentiality and/or anonymity; and contact number and address or telephone number. This ensures that the respondents know what they are committing themselves to, and also that they understand the

context of their replies. If possible, you should offer an estimate of the completion time. Instructions for return should be included with the return date made obvious. For example it would be appreciated if you could return the completed questionnaire by.... if at all possible.

Instructions for Completion:

Unfamiliar response system is employed, you should give an example.

Appearance:

Appearance is usually the first feature of the questionnaire to which the recipient reacts. A neat and professional look will encourage further consideration of your request, increasing your response rate. In addition, careful thought to layout should help your analysis. There are a number of simple rules to help improve questionnaire appearance:

- Liberal spacing makes the reading easier.
- Photo - reduction can produce more space without reducing content.
- Consistent positioning of response boxes, usually to the right, speeds up completion and also avoids inadvertent omission of responses.
- Choose the font style to maximize legibility.
- Differentiate between instructions and questions. Either lower case and capitals can be used, or responses can be boxed.

Length:

There may be a strong temptation to include any vaguely interesting questions, but you should resist this at all costs. Excessive size can only reduce response rates. If a long questionnaire is necessary, then you must give even more thought to appearance. It is best to leave pages unnumbered, for respondents to flick to the end and can be very disconcerting.

Order:

Probably the most crucial stage in questionnaire response is the beginning. Once the respondents have started to complete the questions they will normally finish the task, unless it is very long or difficult. Consequently, you need to select the opening questions with care. Usually the best approach is to ask for biographical details first, as the respondents should know all the answers without much thought. Another benefit is that an easy start provides practice in answering questions.

Once the introduction has been achieved the subsequent order will depend on many considerations. You should be aware of the varying importance of different questions. Essential information should appear early, just in case the questionnaire is not completed. For the same reasons, relatively unimportant questions can be placed towards the end. If questions are likely to provoke the respondent and remain unanswered, these too are best left until the end, in the hope of obtaining answers to everything else.

Coding:

If analysis of the results is to be carried out using a statistical package or spreadsheet it is advisable to code non - numerical response when designing the questionnaire, rather than trying to code the responses when they are returned. An example of coding is:

Male [ ] Female [ ]

1 2

The coded responses (1 or 2) are then used for the analysis.

2.3 Interviews:

Interviewing is at technique that is primarily used to gain an understanding of the underlying reasons and motivations for people's attitudes, preferences or behavior. Interviews can be undertaken on a personal one - to - one basis or in a group. They can be conducted at work, at home, in the street or in a shopping centre, or some other agreed location.

Personal Interview:

Advantages:

- Serious approach by respondent resulting in accurate information.
- Good response rate.
- Completed and immediate.
- Possible in - depth questions.
- Interviewer in control and can give help if there is a problem.
- Can investigate motives and feelings.
- Can use recording equipment.
- Characteristics of respondent assessed - tone of voice, facial expression, hesitation etc...
- Can use props.
- If one interviewer used, uniformity of approach.
- Used to pilot other methods.

Disadvantages:

- Need to set up interviews.
- Can be expensive.
- Normally need a set of questions.
- Respondent bias - tendency to please or impress, create false personal image, or end interview quickly.

- Embarrassment possible if personal questions.
- Transcription and analysis can present problems - subjectivity.
- If many interviewers, training required.

Types of Interview:

Structured:

- Based on a carefully worded interview schedule.
- Frequently require short answers with the answers being ticked off.
- Useful when there are a lot of questions which are not particularly contentious or thought provoking.
- Respondent may become irritated by having to give over - simplified answers.

Semi - Structured:

The interview is focused by asking certain questions but with scope for the respondent to express him or herself at length.

Unstructured:

This also called an in depth interview. The interviewer begins by asking a general question. The interviewer then encourages the respondent to talk freely. The interviewer uses an unstructured format, the subsequent direction of the interview being determined by, the respondent's initial reply. The interviewer then probes for elaboration 'Why do you say that?' Or, 'That's interesting, tell me more or, would you like to add anything else? Being typical probes.

The following section is a step - by - step guide to conducting an interview. You should remember that all situations are different and therefore you may need refinements to the approach.

Planning an Interview:

- List the area in which you require information.
- Decide on type of interview.
- Transform areas into actual questions.
- Try them out on a friend or relative.
- Make an appointment with respondent - accuse details why and how

Conduction an Interview:

- Personally - Arrive on time be smart smile employ good manners find a balance between friendliness and objectivity.
- At the start - Introduce your self re - confirm the purpose assure confidentiality - if relevant specify what will happen to the data.

- The questions - Speak slowly in a soft, yet audible tone of voice control your body language know the questions and topic ask all the questions.
- Responses - Recorded as you go on questionnaire written verbatim, but slow and time consuming summarized by you taped agree beforehand have alternative method if not acceptable consider effect on respondent's answers proper equipment in good working order sufficient tapes and batteries minimum of background noise.
- At the end - Ask if the respondent would like to give further details about anything or any questions about the research thank them.

Telephone Interview:

This is an alternative form of interview to the personal, face - to face interview.

Advantages:

- Relatively cheap.
- Quick.
- Can cover reasonably large numbers of people or organizations.
- Wide georgraphic coverage.
- High response rate - keep going till the required number.
- No waiting.
- Spontaneous response.
- Help can be given to the respondent.

Disadvantages:

- Often connected with selling.
- Questionnaire required.
- Not everyone has a telephone.
- Repeat calls are inevitable average 2.5 calls to get someone.
- Time is wasted.
- Straightforward questions are required.
- Respondent has little time to think.
- Cannot use visual aids.
- Can cause irritation.

- Good telephone manner is required.
- Question of authority.

Getting Started:

- **Located the respondent:**
  - Repeat calls may be necessary especially if you are trying to contact people in originations where you may have to go through secretaries.
  - You may not know an individual's name or title - so there is the possibility of interviewing the wrong person.
  - You can send an advance letter informing the respondent that you will be telephoning. This can explain the purpose of the research.
- **Getting them to agree to take part:**
  - You need to state concisely the purpose of the call - scripted and similar to the introductory letter of a postal questionnaire.
  - Respondents will normally listen to this introduction before they decide to co - operate or refuse.
  - When contact is made respondents may have questions or raise objections about why they could not participate. You should be prepared for these.

Ensuring Quality:

- **Quality of questionnaire** - follows the principles of questionnaire design. However, it must be easy to move through as you cannot have long silences on the telephone.
- **Interview Schedule** - each interview schedule should have a cover page with number, name and address. The cover sheet should make provision to record which call it is, the date and time, the interviewer the outcome of the call and space to note down specific times at which a call - back has been arranged. Space should be provided to record the final outcome of the call - was an interview refused, contact never made number disconnected, etc.,
- **Procedure for call backs** - a system for call backs needs to be implemented. Interview schedules should be sorted according to their status weekday call back, evening call back, weekend call back, specific time call back.

Comparison of postal, telephone and personal interview surveys:

The table below compares the three common methods of postal telephone and interview surveys - it might help you to decide which one to use.

	Postal Survey	Telephone Survey	Personal Survey
Cost (assuming a good response rate)	Often lowest	Usually in between	Usually highest
Ability to probe	No personal contact or observation	Some chance for gathering additional data through elaboration on quesions, but no personal observation	Greatest opportunity for observation, building rapport, and additional probing.
Respondent ability to complete at own convenience	Yes	Perhaps, but usually no	Perhaps, if interview time is prearranged with respondent
Interview bias	No chance	Some, perhaps due to voice inflection	Greatest chance
Ability to decide who actually responds to the questions	Least	Some	Greatest
Impersonality	Greatest	Some due to lack of face - to - face contact	Least
Complex Questions	Least Suitable	Somewhat Suitable	More suitable
Visual aids	Little opportunity	No opportunity	Greatest opportunity
Potential negative respondent reaction	"Junk mail"	"Junk calls"	Invasion of privacy
Interviewer Interview environment	Least	Some in to call	Greatest
Time lag between soliciting and receiving response	Greatest	Least	May be considerable if a large area involved
Suitable types of questions	Simple, mostly dichotomous (yes/no) and multiple choice	Some opportunity for open - ended questions especially if interview is recorded	Greatest opportunity for open ended questions
Requirement for technical skills in conducting interview	Least	Medium	Greatest
Response rate	Low	Usually high	High

Table 3.1 Comparison of the three common methods of surveys

2.4 Observation:

Observation involves recording the behavioural patterns of people, objects and events in a systematic manner. Observational methods may be:

- Strucutred or unstructured
- Disguised or undisguised
- Natural or contrived
- Personal
- Mechanical
- Non-Participant
- Participant, with the participant taking a number of different roles.

Structured or Unstructured:

In structured observation the researcher specifies indetail what is to be observed and how the measuremetns are to be recorded. it is appropriate when the problem is clearly defined and the information needed is specified.

Disguised or Undisguised:

In disguised observation, respondents are unaware they are being observed and thus behave naturally. Disguise is achieved, for example, by hiding, or using hidden equipment or people disguised as shoppers.

In undisguised observation, respondents are aware they are being observed. There is a danger of the Hawthorne effect - people behave differently when being observed.

Natural or contrived:

Natural observation involves observing behaviors as it takes place in the environment, for example, eating hamburgers in a fast food outlet.

In contrived observation the respondents behavior is observed in an artificial environment, for example, a food tasting session.

Personal:

In personal observation, a researcher observes actual behaviors as it occurs. The observer may or may not normally attempt to control or manipulate the phenomenon being observed. The observer merely records what takes place.

Mechanical:

Mechanical devices (Video, closed, circuit television) record what is being observed. These devices may or may into require the respondent's direct participation. They are used for continuously recording on - going behaviors.

Non-Participant:

The observer does not normally question or communicate with the people being observed. He or she does not participate.

Participant:

In participant observation, the researcher becomes or is part of the group that is being investigated. Participant observation has its roots in ethnographic studies (study of man and races) where researchers would live in tribal villages, attempting to understand the customs and practices of that culture. It has a very extensive literature, particularly in sociology (development, nature and laws of human society) and anthropology (physiological and psychological study of man).

The role of the participant observer is not simple. There are different ways of classifying the role:

- Researcher as employee
- Researcher as an explicit role
- Interrupted involvement
- Observation alone

Researcher as Employee:

Will have implications for the extent to which he or she will be able to move around and gather information and perspectives from other sources. This role is appropriate when the researcher needs to become totally immersed and experience the work or situation at first hand.

There are a number of dilemmas. Do you tell management and the unions? Friendships may compromise the research. What are the ethics of the process? Can anonymity be maintained? Skill and competence to undertake the work may be required. The research may be over a long period of time.

Researcher as an explicit role:

The researcher is present every day over a period of time but entry is negotiated in advance with management and preferable with employees as well. The individual is quite clearly in the role of a researcher who can move around, observe, interview and participate in the work as appropriate. This type of role is the most favored, as it provides many of the insights that the complete observer would gain, whilst offering much greater flexibility without the ethical problems that deceptionentails.

Interrupted Involvement:

The researcher is present sporadically over a period of time, for example, moving in and out of the organization to deal with other work or to conduct interviews with, or observations of, different people across a number of different organizations. It rarely involves much participation in the work.

Observation alone:

The observer role is often disliked by employees since it appears to be 'eavesdropping'. The inevitable detachment prevents the degree of trust and friendship forming between the researcher and respondent, which is an important component in or their methods.

Choice of roles:

The role adopted depends on the following:

- Purpose of the research. Does the research require continued longitudinal involvement (long period of time) or will in depth interviews, for example, conducted over time give the type of insights required?
- Cost of the research. To what extent can the researcher afford to be committed for extended periods of time? Are there additional costs such as training?
- The extent to which access can be gained. Gainingaccess where the role of the researcher is either explicit or covert can be difficult and may take time.
- The extent to which the researcher would be comfortable in the role. If the researcher intends to keep his identity concealed, will he or she also feel able to develop the type of trusting relationships that are important? What are the ethical issues?
- The amount of time the researcher has at his disposal, Some methods involve a considerable amount of time. If time is a problem alternate approaches will have to be sought.

2.5 Case - Studies:

The term case - study usually refers to a fairly intensive examination of a single unit such as a person, a small group of people or a single company Case - studies involve measuring what is there and how it got there. In this sense, it is historical. It can enable the researcher to explore, unravel and understand problems, issues and relationships. It cannot, however, allow the researcher to generalis, that is to argue that from one case - study the results, findings or theory developed apply to other similar case - studies. The case looked at may be unique and therefore not representative of other instances. It is, of course, possible to look at several case - studies to represent certain features of management that we are interested in studying. The case - study approach is often done to make practical improvements. Contributions to general knowledge are incidental.

1. Determine the present situation.
2. Gather background information about the past and key variables.
3. Test hypotheses. The background information collected will have been analyzed for possible hypotheses. in this step, specific evident about each hypothesis can be gathered. This step aims to eliminate possibilities which conflict with the evidence collected and to gain confidence for the important hypotheses. The culmination of this step might be the development of an experimental design to test out more rigorously the hypotheses developed, or it might be to take action to remedy the problem.
4. Take remedial action. The aim is to check that the hypotheses tested actually work out in practice. Some action, correction or improvement is made and a re - check carried out on the situation to see what effect the change has brought about.

The case - study enables rich information to be gathered from which potentially useful hypotheses can be generated. It can be a time - consuming process. It is also inefficient in researching situations which are already well structured and where the important variables have been identified. They lack utility when attempting to reach rigorous conclusions or determining precise relationships between variables.

2.6 Diaries:

A diary is a way of gathering information about the way individuals spend their time on professional activities. They are not about records of engagements or personal journals of thought. Diaries can record either quantitative or qualitative data and in management research can provide information about work patterns and activates.

Advantages:

- Useful for collecting information from employees.
- Different writers compared and contrasted simultaneously.
- Allows the researcher freedom to move from one organization to another.
- Researcher not personally involved.
- Diaries can be used as a preliminary or basis for intensive interviewing.
- Used as an alternative to direct observation or where resources are limited.

Disadvantages:

- Subjects need to be clear about what they are being asked to do, why and what you plan to do with the data.
- Diarists need to be of a certain educational level.
- Some structure is necessary to give the diarist focus, for example a list of headings.
- Encouragement and reassurance are needed as completing a diary is time - consuming and can be irritating after a while.
- Progress needs checking from time to time.
- Confidentiality is required as content may be critical.

2.7 Critical incidents:

The critical incident technique is an attempt to identify the more 'noteworthy' aspects of job behaviour and is based on the assumption that jobs are composed of critical and non - critical tasks. For example, a critical task might be defined as one that makes the difference between success and failure in carrying out important parts of the job. The idea is to collect reports about what people do that is particularly effective in contributing to good performance. The incidents are scaled in order of difficulty, frequency and importance to the job as a whole.

The technique scores over the use of diaries as it is centrad on specific happenings and on what is judged as effective behaviour. However, it is laborious and does not lend itself to objective quantification.

2.9 Protfolios:

A measure of a manager's ability may be expressed interms of the number and duration of issues or problems being tackled at any one time. The compilation of problem portfolios is recording information about how each problem arose, methods used to solve it difficulties encountered etc. This analysis also raises questions about the person's use of time. What proportion of time is occupied in checking in handling problems given by others; on self generated problems on 'top priority' problems on minor issues, etc.? The main problem with this method and the use of diaries is getting people to agree to record everything in sufficient detail for you to analyses. It is very time consuming!

Sampling:

Collecting data is time consuming and expensive, even for relatively small amounts of data. Hence it is highly unlikely that a complete population will be investigated. Because of the time and cost elements the amount of data you collect will be limited and the number of people or organizations you contact will be small in number. You will, therefore have to take a sample and usually a small sample.

Sampling theory says a correctly taken sample of an appropriate size will yield results that can be applied to the population as a whole. There is a lot in this statement but the two fundamental questions to ensure generalization are:

1. How is a sample taken correctly?
2. How high should the sample be?

The answer to the second question is 'as large as possible given the circumstances'. It is like answering the question How long is a piece of string? It all depends on the circumstances.

Whilst we do not expect you to normally generalize your results and take a large sample, we do expect that you follow a recognized sampling procedure, such that if the sample was increased demineralization would be possible. You therefore need to know some of the basics of sampling. This will be done by reference to the following example.

Example:

The theory of sampling is based on random samples - where all items in the population have the same chance of being selected as sample units. Random samples can be drawn in a number of ways but are usually based on having some information about population members. This information is usually in the form of an alphabetical list - called the sampling frame.

Three types of random sample can be drawn - a simple random sample (SRS) a stratified sample and a systematic sample.

Simple and sampling can be carried out in two ways - the lottery method and using random numbers.

The lottery method involves:

- Transferring each person's name from the list and putting it on a piece of paper.
- The pieces of paper are placed in a container and thoroughly mixed.

- Their required number are selected by someone without looking.
- The names selected are the simple random sample.

This is basically similar to a game of bingo or the national lottery. The procedure is easy to carry out especially if both population and sample are small, but can be tedious and time consuming for large populations or large samples.

Alternatively random numbers can be used. Random numbers are strings of digits that have been generated by the lottery method and can be found in books of statistical tables. An example of these is:

03	47	43	73	86	36	96	47	36	61
97	74	24	67	62	42	81	14	57	20
16	76	62	27	66	56	50	26	71	07
12	56	85	99	26	96	96	68	27	31
55	59	56	35	64	38	04	80	46	22

Random numbers tend to be written in pairs and blocks of 5 by 5 to make reading easy. However care is needed when reading these tables. The numbers can be read in any direction but they should be read as a single string of digits i.e. left to right as 0, 3, 4, 7 etc. or top to bottom as 0, 9, 1, 1, 5, 3, 7, ... etc It is usual to read left to right.

The random number method involves:

- Allocating a number to each person the list (each number mist consist of the same number of digits so that the tables can be read consis tently).
- Find a starting point at random in the tables (close your eyes and point)
- Read off the digits.
- The names matching the numbers are the sample units.

For the example of selecting nine people at random from 90.

a) The sampling frame is the list of 90 people. Number this list 00, 01, 02, ..., 89. Note that each number has two digits and the numbering starts from 00.

b)

c) Let the next two digits be 76, then the person numbered 76 is the second sample unit.

This procedure is repeated until the nine people have been identified.

d) Any number occurring for second time is ignored as is any two digit number over 89.

Simple random number sampling is used as the basis for many other methods, but has two disadvantages.

1. A sampling frame is required. This may not be available exist or be incomplete.

2. The procedure is unbiased but the sample may be biased. For instance, if the 90 people are a mixture of men and women and all men were selected this would be a biased sample.

   To overcome this problem stratified sample can be taken. In this the population structure is reflected in the sample structure, with respect to some criterion.

   For example suppose the 90 people consist of 30 men and 60 women. If gender is the cirterion for stratification then :

   $\frac{30}{90}$ of the sample should be men

   i.e., $\frac{30}{90} \times 9 = 3$ men

   $\frac{60}{90}$ of the sample should be women

   i.e., $\frac{60}{90} \times 9 = 6$ women

   Thus the sample reflects the population structure in terms of gender.

   The three men and six women would then be selected by simple random sampling e.g., random numbers.

   The problem with this approach is the criterion for stratification (e.g., age, sex, job, description), is chosen by you - it is subjective and may not be the best or more appropriate criterion. Also a more detailed sampling frame is required.

Systematic sampling:

Whilst not truly random this is a method that is used extensively because it is easy to operate and quick, even when the population and the sample are large.

For example for the population 90 and sample of nine

i.e., - to 9

10 to 19

etc

80 to 89

Select a number between 1 and 9 using random number tables.

Suppose this number is 6.

Person numbered 6 is chosen.

Then the 16th, 26th, 36th, 46th, 56th, 66th, 76th and 86th people are there remaining sample units.

If no sampling frame is available access to the population is necessary, such as customers of a business such as a leisure centre, restaurant or museum.

Systematic sampling can be used by selecting a random number say 25.

Then the 25th person to enter is the first sample unit.

The 50th person to enter is the second sample unit.

This process is carried on until the required sample size is met.

This approach usually generates a good cross section of the population. However, you may need a team of people when no sampling frame exists to help with counting, interviewing etc.

2.9 Summary:

Data collection is a very basic activity in decision making. In this lesson, concepts of questionnairs, interviews and observation methods are explained. Compared among postel, telephone and personal invterview methods of surveys. Case-studies, diaries, critical incidents and protfolios methods are also described.

2.10 Exercises:

1. What is the questionaires methods and explain with an example.
2. Give different methods of collection of data.
3. Give briefly the characteristics of a good questionaire or a schedule.
4. Mention different kinds of statistical investigations.
5. Explainthe interview methods.
6. Compaire among the postel, telephone and personal interview methods of surveys.
7. Describe the observation methods.
8. Explain the critical incidents technique.
9. Describe the protfolios method.

2.11 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics

   Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi.,

   S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 3

DATA REPRESENTATION - TABLES - GRAPHS - DIAGRAMS

Objective:

After going through this lesson, you will learn

(i) To motivate the idea of data condensing and promote statistical tables, graphs and diagrams as simple tools for this purpose.

(ii) To explain the steps in tabulation of data and the features of a good statistical table.

(iii) To outline the construction of frequency tables and their modifications in the form of relative and cumulative frequency tables.

(iv) To explain and illustrate the construction of graphs and diagrams for quantitative and qualitative data sets.

Structure:

3.1 Introduction:

Successful use of collected data depends of a great extent, upon the manner in which it is arranged, displayed and summarised. Data would assist the decision maker in making educated guesses abou tthe causes and pattern of changes in certain variables.

In the introductory lesson we acquainted ourselves with different types of data and the need for their systematic analysis. Let us now take a progressive step by getting to know the methods of dealing with data. We begin with the idea of summarizing data.

3.2 Tabulation:

Data Condensing:

When a large amount of data is available, it is desirable to condense it for ease of comprehension and further analysis. The most common method of summarising data is to express them in the form of tables. Here is an example.

Example 3.1:

The results of degree examination for students of a college are shown below:

Summary of results of an examination

Average Marks	Number of students		
	2006	2007	2008
(1)	(2)	(3)	(4)
Less than 35	8	9	10
35 - 49	40	35	45
50 - 59	20	26	30
60 or more	5	10	15
Total	73	80	100

A statistical table is a systematic organization of data into rows and columns. Rows are horizontal arrangements and columns are vertical. Table 3.1 has 4 rows and an equal number of columns. It is a good practice to number the columns for ease of reference. Table 3.1 compactly presents the examination results, which may be further analyzed. A comparison over the years 2006 - 08 is also made easy. A specific aspect like the number of failures can also be focussed upon. Thus the purposes of tabulation are

(i) to summarize data with a view to allow further analysis,

(ii) to facilitate comparison of similar data across places, years etc.,

and

(iii) to draw the attention of the readers to the special features.

The main steps in tabulation are as follows:

(i) Deciding the objective:

Data from a census or a survey are used for many purposes. Tables should be prepared highlighting a particular characteristic as per the purpose. In the absence of a specific objective, general purpose tables may be constructed. These provide secondary data for others to analyses.

(ii) Choosing suitable characteristics for classification:

If the data are on qualitative characteristics such as sex, color or mother-tongue, then the basis of classification suggests itself. For instance, with data on human populations one can have tables with sex (male, female), marital status (single, married, widowed, divorced), educational level (illiterate, primary school, high school, college) etc., as the bases. For quantitative variables like height, weight, thickness or area, one usually prepares frequency tables as in table 3.1

(iii) Deciding about the type of table:

With several characteristics of interest, one can have tables based on one or more variables or attributes taken at a time. For instance, we use bivariate tables for compactly presenting data on two characteristics. This will be illustrated in see 3.4. A table showing data on two or more attributes is referred to as a contingency table. An illustration is in example 3.2. If more than two variables or attributes are to be accommondated in a table then one may subdivide the rows and columns, with characteristics arranged in a logical order. Usually there should be more rows than columns. In the final analysis there are only two guidelines for preparing a table. First, use of common sense in planning a table and second viewing the proposed table form the stand point of the user. The details of the mechanical arrangement must be governed by a single objective, namely to make the table as easy to read and understand as the nature of the material permits.

(iv) Giving a title, footnote etc.

An appropriate title highlighting the essence of the table is necessary part. The time reference and the units of measurement are to be indicated. Data source and meaning of symbols, if any, are to be given as foot notes.

To sum up, the requisites of a good table are that it should be (i) compact, (ii) providing a summary of relevant information, (iii) accurate, unambigous and complete, (iv) neat and well spaced with a suitable title, and (v) having characteristics strung together in a logical sequence. The following is a typical example of a contingency table.

Example 3.2

The results of a survey on the specialization of students in degree and the subsequent field of study at the PG level are summarized below:

Table 3.2 Field of study at PG level versus Degree level specializations

Specialization in Degree	Field of study in University			Total
	Biology	Medicine	Agriculture	
Biology	26	52	23	101
Physics and Mathematics	3	44	8	55
Agriculture	4	1	15	20
Humanities	6	11	10	7
Total	39	108	56	203

In table 3.2 the rows represent the specialization at degree level and the columns refer to the study field at PG level. Examining the association between the two attributes may be an objective of analysis.

Frequency Table:

Quantitative data from surveys and reports are often summarizes as a frequency table. An example is provided by Table 3.1. The data may be concerned with a population or a sample. The steps in preparing such a table are outlined below:

(i) Determine the classes or subintervals: This is to be done such that a measurement falls into one and only one subinterval called a class. The measurements are grouped according to the class into which they fall. The number of such classes and the midpoints of the classes are at the discretion of the analyst. These can be chosen to achieve effective and useful data condensing. Now - a - days, with computer facility available, the analyst can in fact try different combinations of (a) the number of class intervals and (b) their mid points. The resulting tables may be visually compared and the best looking combination may be finally used.

(ii) Obtain the class frequencies: Classify the observations into one of the classes. This is best done by using tally - marks in the case of manual tabulation. The number of tallies in a class gives the class frequency. The class intervals together with the corresponding frequencies constitute a frequency table. The next example clarifies these steps.

Example 3.3

The following data, obtained in an air pollution study, are 100 determinations of daily emission of sulphur oxides (in tons) from a large industrial plant.

9.0	15.8	26.4	17.3	11.2	23.9	24.8	18.7	13.9	9.0
23.7	22.6	9.6	6.2	14.7	17.5	26.1	12.8	28.5	17.5
26.8	22.7	18.2	20.5	11.0	20.1	15.5	19.2	16.6	10.5
19.0	15.1	22.8	26.5	20.0	21.0	19.1	21.0	16.5	19.0
18.5	23.1	24.6	20.0	16.1	18.0	7.6	13.4	23.4	14.0
14.1	29.6	19.2	17.0	20.8	24.3	22.4	24.5	18.4	18.0
8.2	22.0	12.4	22.4	13.2	11.8	19.2	20.0	25.7	31.0
25.9	10.5	15.9	27.5	18.1	17.9	9.4	24.0	20.1	28.0
11.2	25.6	17.0	15.1	18.8	24.0	20.5	22.6	18.0	18.0
24.0	20.7	17.5	20.1	10.8	20.1	15.5	19.4	16.7	10.1

Since the smallest observation is 6.2 and the largest is 31.0, we might choose the seven classes 5.0 - 8.9, 9.0 - 12.9 ..... and 29.0 to 32.9 . We note that these seven classes accommodate all of the data, they do not overlap and they are all of the same length viz. 4. We next tally the 100 observations and obtain the following frequency table.

Table 3.3 A frequency table for air pollution data

Tons of sulphur Oxides (1)	Tally marks (2)	Frequency (f) (3)
5.0 - 8.9	|||	3
9.0 - 12.9	𝍸 𝍸 ||||	14
13.0 - 16.9	𝍸 𝍸 𝍸 |	16
17.0 - 20.9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	35
21.0 - 24.9	𝍸 𝍸 𝍸 𝍸	20
25.0 - 28.9	𝍸 𝍸	10
29.0 - 32.9	||	2
	Total	100

The numbers in column 3 of the table which show how many items fall into each class, are the class frequencies. The smallest and the largest values that can go into any given class are referred to as class limits. In Table 2.3 these are 5.0 and 8.9, 9.0 and 12.9, ..., and 29.0 and 32.9. More specially 5.0, 9.0, ..., 29.0 are the lower class limits, while 8.9, 12.9, ..., 32.9 are the upper class limits.

The choice of the class limits depends on the extent to which the measurements to be grouped are rounded. If these are weights to the nearest kilogram, the class 40 - 49 will in fact contain all weights between 39.5 and 49.5 kgs. On the other hand for lengths measured to the nearest decimeter, the class 10.0 - 12.4 will actually contain all lengths between 9.95 and 12.45. These pairs of values are generally called the class boundaries or real class limits.

It must be emphasized that when we pass over from a data set to a frequency table there will be some loss of information. Given a data set, a frequency table can be constructed, while given a frequency table the original data set cannot be retrieved fully. This is so because the details regarding values in any class interval is now lost. In a sense this is the price we pay for achieving data condensation through a frequency table. It must also be noted that with another set of class intervals a different frequency table would be obtained. For instance, the classes 5.0 - 0.9, 10 - 14.9, ..., 30 - 34.9 for pollution data lead to the next table.

Table 3.4 Another frequency table for the air pollution data

Class interval (1)	Frequency (f) (2)
5.0 - 9.9	7
10.0 - 14.9	16
15.0 - 19.9	34
20.0 - 24.9	32
25.0 - 29.9	10
30.0 - 34.9	1
Total	100

The availability of a computer makes it easy to play with different sets of class intervals and choose the best looking frequency table for any data set. Relative to the original air pollution data of Example 3.3 the compactness of tables 3.3 or 3.4 is apparent. Such a table is also sometimes referred to as a frequency distribution, but we prefer to call it a frequency table. It is a convenient, commonly used data condensing tool, particularly in the context of large data sets.

For further statistical analysis the concepts of a class mark and class width are necessary. A class mark is the midpoint of a class. It is obtained by adding the lower and upper limits of a class and dividing the sum by 2. Thus

class mark = (lower class limit + upper class limit) / 2

In table 3.3 the class marks are 6.95, 10.95, ..., 30.95 while in table 3.4 these are 7.45, 12.45, ..., 32.45. The difference between two successive lower (or upper) limits is the class width. In table 3.3 the width is 4 and in the next table it is 5. If all classes are equally wide, then the class width is also given by the difference between any two successive class marks.

This is in fact the case with table 3.3 and 3.4. It is pertinent to mention that class - width may not be given by the difference between the upper and lower limits of a class (why?). We next discuss two useful variants of a frequency table.

Relative and cumulative frequency tables:

These are two modifications of a frequency table to suit particular needs. The first is to express it as a relative frequency table by dividing each class frequency by the total frequency. Table 3.5 illustrates this.

Table 3.5: Relative frequency table for data in table 3.3

Class mark (1)	Relative frequency (2)	Percentage Distribution (3)
6.95	0.03	3
10.95	0.14	14
14.95	0.16	16
18.95	0.35	35
22.95	0.20	20
26.95	0.10	10
30.95	0.02	2
Total	1.00	100

A relative frequency table is often employed to compare two or more situations; for instance, to compare the sulphur oxide emissions of two different plants. These may have unequal total frequencies. But this does not pose any problem for comparision since a relative frequency distribution shows only the proportion of observations in the different classes. Multiplication of relative frequencies by 100 gives a percentage frequency distribution. Table 3.5 provides a mental picture as to how the total frequency is "distributed" among different classes.

A second way modifying a frequency table is to express it as a 'less than' or 'more than' cummulative frequency table. This is done simply by adding the class frequencies, starting either at the top or at the bottom of the table. The next table illustrates these concepts.

Table 3.6 Cumulative Frequency table for the emission data

(a) 'Less than' type		(b) 'More than' type	
Tons of sulphur oxide 'less than' (1)	Cumulative frequency (F) (2)	Tons of sulphur oxide 'more than' (1)	Cumulative Frequency (F) (2)
5.0	0	33.0	0
9.0	3	29.0	2
13.0	17	25.0	12
17.0	33	21.0	32
21.0	68	17.0	67
25.0	88	13.0	83
29.0	98	9.0	97
33.0	100	5.0	100

A 'less than' type cumulative frequency table gives us the number of observations with values less than a specified level at a glance. A similar interpretation of 'more than' type cumulative frequencies is apparent. In Table 3.6, for example, 68 observations have value less than 21, while 67 observations have value more than 17. The cumulated tables provide an aggregated picture, indicating the position 'before' or 'beyond' a specified 'level'.

3.3 Pictoral Presentation of Data:

A picture is said to be worth one thousand words. Presenting data in the form of pictures makes them easier for comprehension. Well drawn picutes attract and hold the attention of viewers, convey the information at a glance and help to discover relations. They are also free from the language barrier, except for the accompanying descriptions. If properly prepared, such visual aids are of real practical value. These aids may be conveniently grouped into two: diagrams and graphs.

In a two dimensional set up if both the axes of a coordinate system are used to represent numerical values while drawing a picture we will call it a graph. On the other hand, if lengths in only one direction or areas or symbols are used to represent magnitudes we call the corresponding picture a diagram. We first consider a few diagrams which are often seen in newspapers, magazines and reports. These are

(i) a line diagram

(ii) a bar diagram and

(iii) a pie diagram

In a diagram the vertical axis often represents a quantity (e.g. Rupees, Tons) or percentages, while the horizontal axis indicates the categories, places or time points. This diagram does not present specific data as clearly as a table does, but this is often able to display relations more clearly. Bar diagrams show by the length of the bars the quantities to be represented. Here too one axis shows quantities or percentages, while the other is for categories or time points. The bars may be suitably subdivided, when a quantity is made up of several components. For instance, total food production may comprise paddy, wheat, jowar, ragi and other grains. In this case, a bar is subdivided into five components to reflect the relative magnitudes of these factors. Gaps are allowed between the bars to facilitate easy comparison between categories. All the bars are of equal width, though the width as such has no physical interpretation. If a comparison between items within a category is desired and there are several categories then one may juxtapose the bars for the items and allow gaps between categories. This is sometimes called a composite bar diagram. For instance, import and export may be the items to be compared and several countries may be the categories. Then the bars for import and export will be adjacent while gaps will be allowed between countries. Line and bar diagrams serve the same purpose, but subdivision or juxtapostion cannot be conveniently done with the former.

Pie diagrams are simply one or more circles divided into sectors which are proportional in area to the quantities being presented. The sectors of a circle often represent the components of a whole for instance, the break up of the household expenditure as towards food, rent, fue and light, clothing and miscellaneous items. Since the area of a sector is proportional to the angle at the centre, essentially one has to allocate the total angle of 360^0 in proportion to the magnitudes being presented. Sometimes a comparison between circles is to be allowed. Then their areas are to be

proportional to the quantities. This is simply done by keeping the square root of the radii proportional to the quantities (why?). Three illustrations now follow. In each case more than one type of diagrams may be drawn. Some of them are shown in Figs 3.1 to 3.4.

Example 3.4:

The percentage of deaths in India due to major causes during the year 2008 was as follows:

Table 3.7 Distribution of deaths by major causes in India, 2008

Cause Group (1)		Percentage of Deaths (2)
(a)	Accidents & Injuries	8.5
(b)	Complications of pregnancy & Childbirth	1.1
(c)	Fevers	7.3
(d)	Digestive disorders	6.4
(e)	Respiratory disorders	18.9
(f)	Cental nervous system disorders	4.4
(g)	Disorders of circulatory system	11.1
(h)	Other clear symptoms	8.3
(i)	Cases peculiar to infancy	10.2
(j)	Senility	23.8

Fig 3.1 A line diagram for the distribution of deaths by major causes in India (Data in Table 3.7)

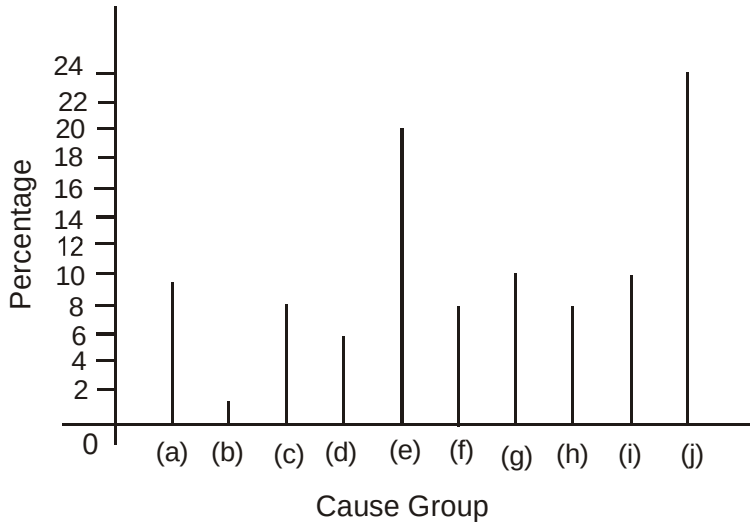


Fig 3.2 A bar diagram for the number of inhabited villages in South Indian States (Data in Table 3.8)

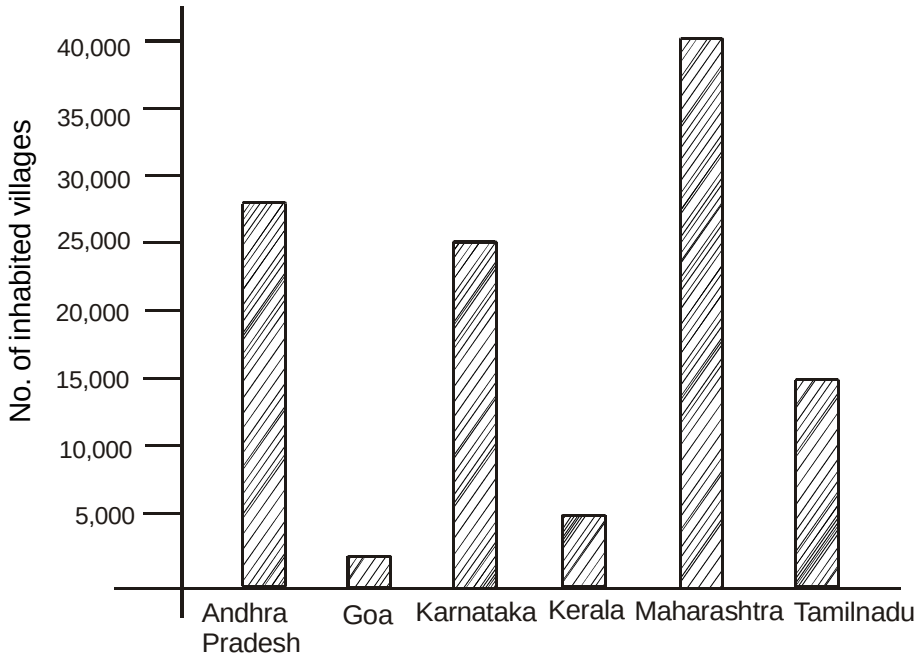
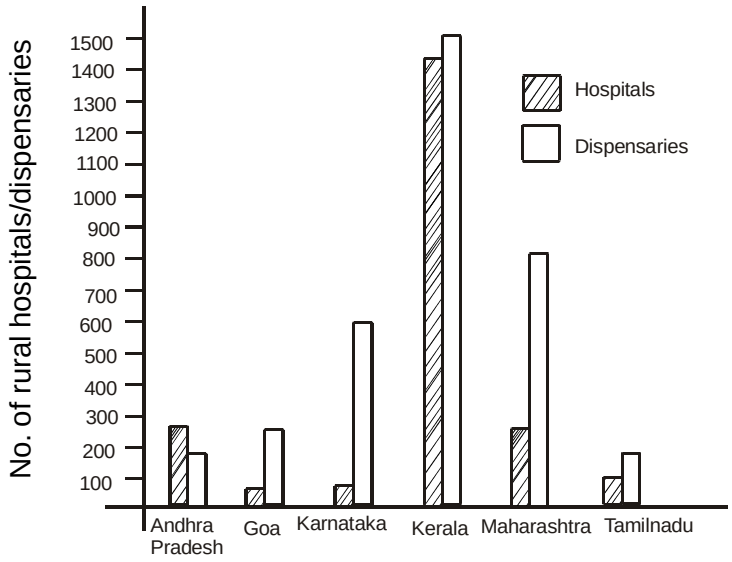


Fig 3.3 A composite bar diagram for data in Cols (3) and (4) of Table 3.8



Example 3.5

The information in the following table relates to the South indian States.

Table 3.8 Some basic information on South Indian States

State (1)	No.of inhabited Villages (2)	No.of rural Hospitals (3)	No.of rural Dispensaries (4)
Andhra Pradesh	27379	322	137
Goa	412	43	307
Karnataka	27028	25	603
Kerala	1362	1440	1442
Maharastra	39354	345	796
Tamil Nadu	15831	89	147
Total	**111366**	**2264**	**3432**

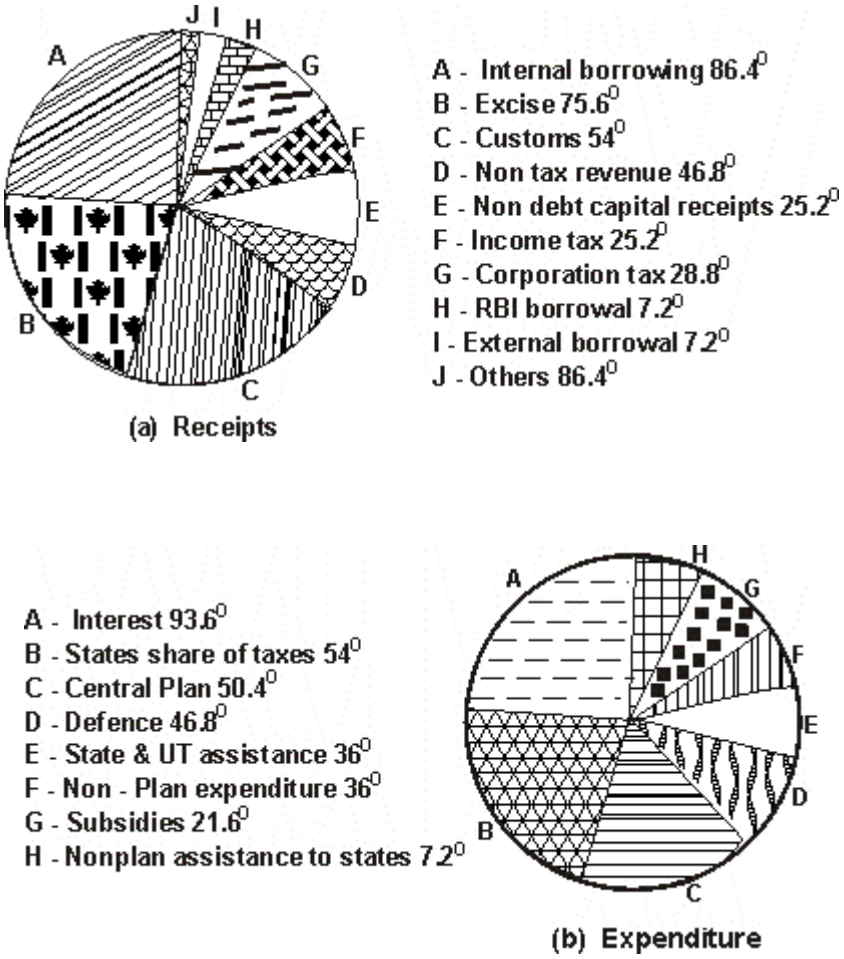
Example 3.6

The break up of Indian Union Budget, 2005 is shown at a glance in the next table.

Table 3.9 How the Indian rupee comes and goes, 2005

Receipts	Paise	Expenditure	Paise
Internal borrowings	24	Interest	26
Excise	21	State's share of taxes & duties	15
Customs	15	Central plan	14
Non-tax revenue	13	Defence	13
Non-debt capital receipts	7	State & UT plan assistance	10
Income Tax	7	Nonplan expenditure	10
Corporate Tax	8	Subsidies	6
Borrowing from RBI	2	Nonplan assistance to State &	
External borrowings	2	UT Govts	6
Other taxes	1		
Total	**100**		**100**

Fig 3.4 Pie diagrams for the budget data of Table 3.9



3.4 Graphical Presentation:

We now proceed to discuss the construction of the following graphs, in the stated order.

(i) Histogram

(ii) Frequency polygon

(iii) Frequency Curve

(iv) Cumulative frequency polygons and curves

Graphical representation of time - series data will be studied in Chapter 8.

Histogram:

This is a set of adjacent rectangles with bases along the intervals between class boundaries and with areas proportional to the frequencies in the corresponding classes. It is the most commonly used frequency graph and is illustrated in Fig 3.5. A histogram is constructed as follows:

(i) Represent the variable being measured on the horizontal axis and the class frequencies on the vertical axis in a graph paper.

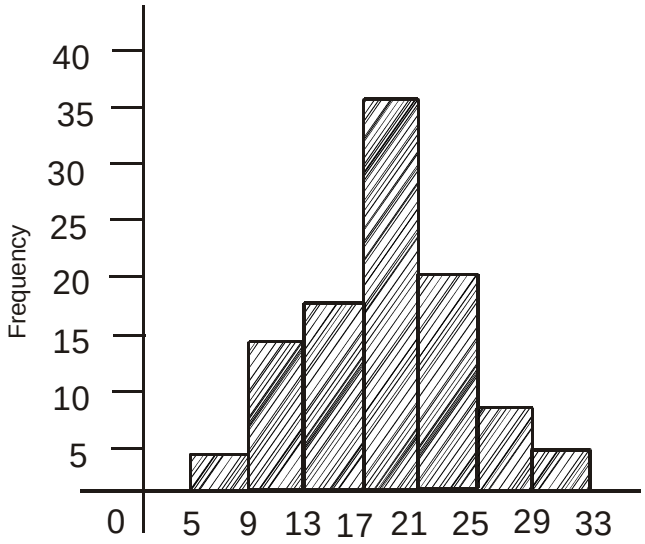
(ii) Erect rectangles on the horizontal axis with bases equal to the class interval and heights determined by the corresponding class frequencies.

(iii) The markings on the horizontal scale can be the class limits as in Gif. 3.5 the class boundaries the class marks or even arbitrary key values.

For easy reference, indicating the class limits on the graph is preferred, although the rectangles really extend from one class boundary to the next. A histogram cannot be drawn for distributions with open classes.

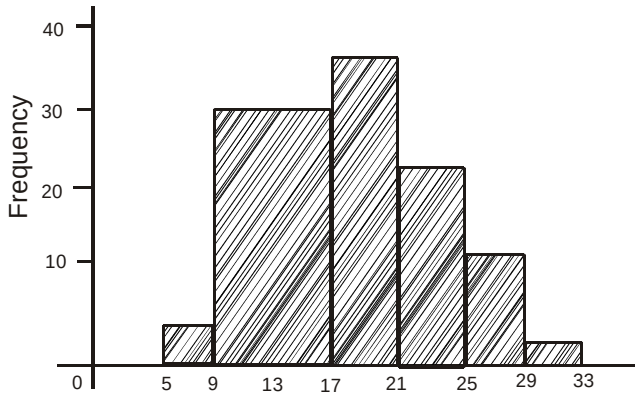
Basically a histogram is an area graph in the sense that the area of a rectangle is proportional to the corresponding frequency. With equal intervals this reduces to proportionality between heights and frequencies as in Fig 3.5. But in general, every height need not be proportional to the corresponding frequency (Refer Fig 3.7). As a consequence of the method of construction the total area under a histogram is proportioanl to the total frequency in the set.

Fig 3.5 Histogram for emission data of Table 3.3



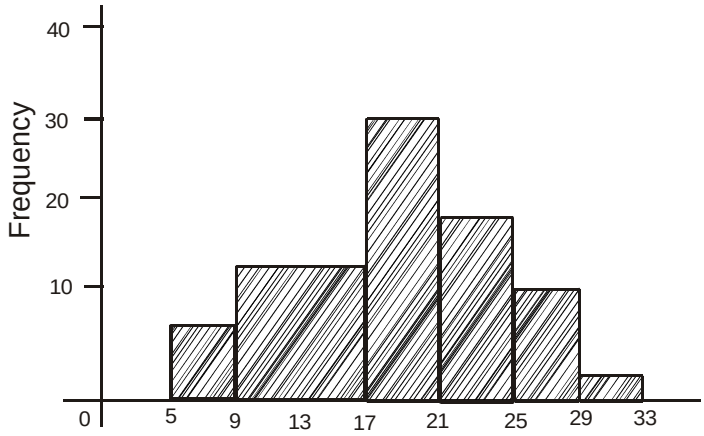
In Fig 3.6 we illustrate a trap in constructing histograms with unequal class intervals. In Table 3.3, suppose we combine the second and third intervals and have the interval 9.0 - 16.9 with frequency 14 + 16 = 30. With heights taken proportional to the frequencies we end up with the incorrect histogram as in Fig 3.6.

Fig 3.6 An incorrect histogram for emission data



Here the class 9.0 - 16.9 has twice as much width as the others. Thus to keep area proportionality, the height of the second rectangle in Fig 3.6 must be reduced by a factor of 2. The corrected histogram is shown in Fig 3.7

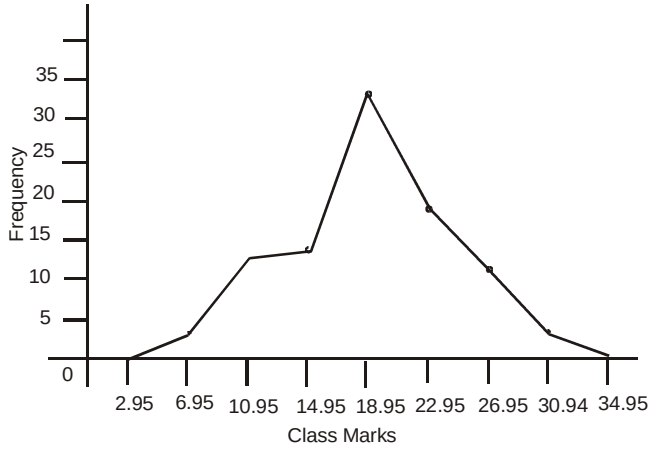
Fig 3.7 Another correct histogram for emission data



Frequency Polygon:

A less widely used graph is the frequency polygon obtained by plotting the frequencies along the y - axis against the class marks along the x - axis. The successive points are joined by straight lines (Refer Fig 3.8).

Fig 3.8 Frequency Polygon for Emission Data

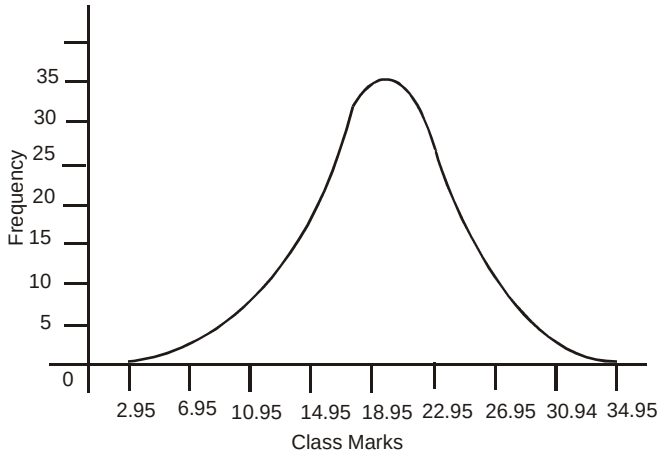


The polygon is closed at both the ends by bringing it down to zero at the class mark immediately previous to the first one and also at the classmark immediately next to the last one. The total areas under a histogram and the corresponding frequency polygon may be verified to be equal and both represent the total frequency. We suggest this as a simple exercise with a hint that both the figures may be drawn on the same graph and the areas of extra triangular pieces that are included in the polygon may be compared with those of similar piece that are excluded with reference to the histogram. However frequency polygons are used only when all the class intervals are equal.

Frequency Curve:

this is a theoretical concept which may be developed as follows. We suppose that the size of the sample can be increased indefinitely, so that even with very small class intervals there are many observations in each class. Then the outline of the histogram will approximate to a smooth curve which is called the frequency curve. In practice such a curve is used to approximate histograms by eye judgment. The total area under the curve, like the total area under the histogram or frequency polygon, represents the total frequency.

Fig 3.9 Frequency curve for Emission Data

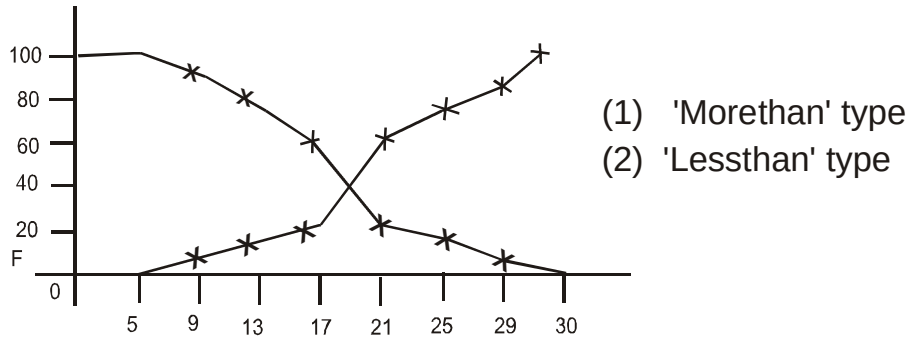


Next we consider two diagrams for cumulative frequencies.

Cumulative Frequency Polygons:

If the less than cumulative frequency F is plotted against the upper class boundary and the points are joined by straight lines we get the less than cumulative frequency polygon. It starts from zero at the lower boundary of the first interval. A plot of more than cumulative frequencies against lower class boundaries leads to the more than cumulative frequency polygon. Fig 3.6 shows the cumulative frequency polygons for the data of Table 3.10.

Fig 3.10 Cumulative Frequency Polygons For Emission Data



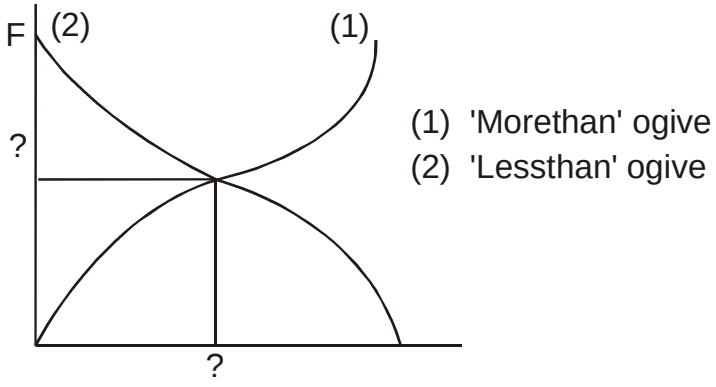
If the assumption is made that the observations in a class are evenly spread over the entire interval, the intermediate points on the polygon also represent the cumulative frequencies at the corresponding x - values.

One may use cumulated relative (or percentage) frequencies to construct cumulative frequency polygons, in which case the polygons will span from 0 to 1 (or 100) on the vertical axis.

Ogives Curves:

Smooth curves can be used to approximate the cumulative frequency polygons, just like a frequency curve approximate a histogram. They are called ogives or ogive curve. Moreover it is often simple to draw and ogive as compared to a frequency curve. Fig 3.11 shows the ogives for the emission data of Table 3.3.

Fig 3.11 Ogivs for emission data



We close the discussion in this section with two questions. What is the significance of the point of intersection of the two ogives? What are its x and y - coordinates?

3.5 Bivariate Tables:

Example 3.1 provided a good illustration of frequency tables based on a single variable, viz., scores of students at an examination. Sometimes we come across data on two related variables for which preparing a single table is desirable in order to examine the relation. Bivariate tables are handy for this purpose. Here the rows represent one variable and the columns the other variable. The following is an example.

Example 3.7

One hundred workers in a factory are classified by (i) age and (ii) number of days of overtime duty during a month.

Age (yrs) / overtime (days)	0 - 4	5 - 9	10 - 14	15 - 19	Total
20 - 29	-	10	15	3	28
30 - 39	-	9	6	8	23
40 - 49	6	10	5	1	22
50 - 59	10	8	-	-	18
60 - 69	4	5	-	-	9
Total	**20**	**42**	**26**	**12**	**100**

An examination of the above table indicates that persons with higher ages generally work less overtime in the chosen factory. This is an inverse relation. Examples of a direct relation are provided by the following pairs of variables:

(i) Yield of paddy and amount of fertilizers applied.

(ii) Annual sales and expenditure on advertisement.

(iii) Examination score and number of hours of study by a student.

Of course, the direct relation implied by (iii) is rather unfortunate !

3.6 Inter-Relationships of Graphs:

Frequency Distribution

Step 1 → Find Range of data set
Range = Max Value - Min Value

Step 2 → Let n^0 of subintervals = K

Find class interval $\left\{\frac{\text{Range}}{K}\right\}$ = Integer of.

Step 3 → Starting from Minimum value write down classes and count the number of observations in each class (i.e. freq.)

Draw rectangle over each interval → Histogram

Draw bars of height equal to frequency of that class. → Bar Diagram

(1) Calculate Relative freq $= \frac{f_i}{n}$

(2) Assign Angle, Angle = $360^0 \times$ Relative freq.

(3) Percentage = Relative freq. x 100

→ Pie Chart

Subject to:
(1) Interval should be equal in size
(2) Number of interval should be between 5 & 15
(3) Intervals should be non - overlapping

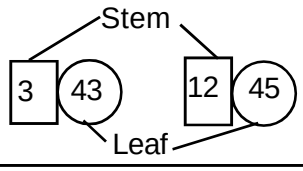
Join midpoints of tops of bars by straight line → Freq. Polygon

Calculate Cumulative Freq. Join midpoints of Tops of Cumu;ative Freq. → Cumulative Frequency Polygon or Ogive

Histogram of Cumulative Freq. → Frequency Histogram by a Smooth Curve → Cumulative Freq. Histogram

STEMS & LEAF

(1) Identify Stem & Leaves e.g.
(2) Arrange STems in increasing order in a vertical column.
(3) Place leaves in the appropriate Stem row



3.7 Summary:

(i) Statistical tables and diagrams are common tools for summarizing data and presenting them in a compact form.

(ii) The key steps in tabulation are deciding the objective, characteristics for classification, type of table and a title.

(iii) A good table must be compact, accurate, unambiguous and complete with characteristics strung together in a logical order.

(iv) Frequency tables represent quantitative data while contingency tables are for qualitative data.

(v) The class intervals together with the corresponding frequencies constitute a frequency table, sometimes called a frequency distribution.

(vi) The class limits are the smallest and largest values that can go into a class. The class boundaries are the real class limits.

(vii) A class mark is the midpoint of a class and can be obtained as the mean of the lower and upper class limits.

(viii) The class - width is the difference between two successive lower (or upper) limits.

(ix) Relative and cumulative distributions are two useful modifications of a frequency distribution. The former is obtained by dividing each class frequency by the total frequency. The latter is obtained by cumulation of frequencies from the top or bottom of a frequency table.

(x) Well drawn diagrams attract and hold the attention of people and convey the information at a glance.

(xi) The principal frequency graphs are histogram, frequency polygon and frequency curve. Cumulative frequency polygons and ogives are for presenting cumulative frequencies.

(xii) Histogram is an area - diagram. Total area under the histogram represents total frequency.

(xiii) Frequency polygon is less used. A frequency curve is a theoretical concept associated with very large samples. In practice it is drawn as an approximating smooth curve superimposed on a histogram.

(xiv) Line, bar and pie diagrams are examples for commonly used diagrams to present data falling into several categories.

(xv) A bivariate table may be used to summarize data on two related variables.

3.8 Exercises:

1. (a) What is a statistical table? Outline the main steps in tabulation.

   (b) Explain the steps in forming a frequency table. Define the terms (i) Class limits, (ii) Class boundaries, (iii) Class mark and (iv) class width.

2. (a) How are relative and cumulative frequency tables obtained from a frequency table? What is a percentage distribution?

   (b) Explain how pictorial presentation of data is helpful.

3. (a) Outline the construction of the following diagrams:

   (i) A line diagram, (ii) A bar diagram and, (iii) A pie diagram.

   What are subdivided and composite bar diagrams?

   (b) How do a bar diagram and a histogram differ? Explain clearly.

4. (a) Describe the construction and uses of the following frequency graphs:

   (i) Histogram (ii) Frequency polygon (iii) Frequency Curve

   How are these related?

   (b) Explain how the following diagrams are drawn. State their utility.

   (i) Less than cumulative frequency polygon

   (ii) More than cumulative frequency polygon

   How are the ogives obtained from these?

5. (a) Substantiate or refute the following diagrams are drawn. State their utility.

   (i) In a histogram the heights of the rectangles are always proportional to the respective class frequencies.

   (ii) In a difference between the lower and upper limits of a class is called the class width.

   (iii) The total areas under a histogram and the corresponding frequency polygon are equal.

   (b) Explain how and when a bivariate table is useful.

6. The following data relate to loan advanced to 40 farmers by a cooperative bank.

   The figures are in hundreds of rupees.

12	8	11	9	5	14	12	10	14	20
18	19	10	7	6	9	10	12	13	15
18	12	8	12	17	20	15	12	14	17
11	15	18	17	19	12	15	5	5	9

(a) Construct a frequency table with class intervals 0 - 4, 5 - 9 etc.,

(b) Draw a histogram and a frequency polygon.

(c) Obtain the less than and more than type cumulative frequency tables.

(d) Draw the two ogives. What is significance of their point of intersection?

(e) Find the percentage of farmers with a loan of (i) at least 1000 Rs., and (ii) at most Rs. 1300.

7. (a) Draw a component bar diagram for the following data related to the number of branches of a nationalised bank.

Branch type/year	2005	2006	2007	2008
Rural	240	240	250	260
Semi - urban	125	130	130	140
Urban	75	75	80	90
Metropolitan	80	100	115	125

(b) The monthly household expenditure of Mr. Rama Rao for July 2008 was as follows:

(i) Use a pie diagram to represent the inforamtion

Item	Expenditure (Rs.)
Rent	1100
Food	1400
Fuel & light	250
Clothing	300
Miscellaneous	550

(ii) Suppose that after 10 years, that is in July 2018 Mr. Rama Rao's expenditure gets doubled for each item. Draw a second pie diagram to allow comparison between the two time points.

8. The birth weight (in kgs) of 40 children is recorded as follows:

2.0	2.1	2.3	3.1	3.0	2.7	2.8	3.5	3.1	2.0
4.0	3.2	3.3	2.8	2.9	2.6	2.4	2.7	3.0	2.9
3.1	3.0	2.0	2.8	3.5	4.0	4.1	3.9	2.8	2.9
4.0	3.4	3.6	2.0	2.6	2.7	2.6	3.0	3.1	3.5

Prepare (i) a frequency table with calss interval 0.4

(ii) a relative frequency table and

(iii) a less than type cumulative frequency table.

What is the percentage of children with birth weight less than 3 kg?

9. Represent the following data in the form of a double bar chart and comment.

Year	Amount (crores of Rs)	
	Imports	Exports
1991 - 92	3,600	3,200
1992 - 93	9,000	6,600
1993 - 94	4,800	3,200
1994 - 95	11,600	7,100

3.9 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 4

MEASURES OF CENTRAL TENDENCY

Objective:

After going through this lesson, you will learn:

- The concept and significance of measures of central tendency.
- To comutevarious measures of central tendency such as arithmetic mean, weighted arithmetic mean, median, mode, geometric mean and harmonic mean.
- To compute quantiles such as quartiles, deciles and percentiles.
- The relationship among various averages.

Structure:

4.1 Introduction:

The methods for graphically presenting data discussed in the earlier chapter provide a starting point for analyzing data. However, the graphic methods do not reveal all the information contained in a data set. Managers who wants to know about the company performance and who want to take necessary decisions are most likely not satisfied with frequency distributions of data. Decision

makers need to become acquainted with some statistical methods of analyzing the data. Specifically, they need to know the methods of summarizing and describing numerical data. One of the popular statistical measure in this regard is the measures of central tendency. The objective of these measures is to find one representative value which can be used to locate and summarize the entire set of varying values. That one value should be helpful to make many decisions concerning the entire set.

4.2 Need For A Single Figure To Represent Data:

The collected information or data from a particular observation about a variable are generally classified and presented in the form of table, if one wishes to draw certain inferences about the characteristic value of such variable under study. The tabulation arranges the facts in a systematic manner and helps one to understand the facts. But in many cases the information presented in tales is so large that the characteristics of the data cannot be readily understood. Thus, there is a need for further condensation of the collected data. There is a great need for a single figure or number which could adequately describe the data. The measures of central tendency are in this direction and they would help us to find some central value around which the data tends to cluster.

4.3 Significance:

A measure of central tendency, by condensing the mass of data in one single value, enable us to get an idea of the entire data. For example, it is impossible to remember monthly wages earned by workers in an organisation. But if the average income is obtained, we get a single value which can be remembered easily.

Measures of central tendency would be helpful for comparison of two or more sets of data. For example, the average number of Bank Customers on Mondays can be compared with the number of customers on Saturdays.

Measures would be helpful to trace the precise relationship when it is desired to establish relationship between different groups in numerical terms. Instead of simply saying that average American is more than that of average. Indian, it is better to establish relationship with average per capital income.

4.4 Types of Averages:

Following are some of the important measures of central tendency or averages used in business and industry.

1. Simple Arithmetic Mean
2. Weigted Arithmetic Mean
3. Median
4. Mode
5. Geometric Mean
6. Harmonic Mean

While middle two are called positional averages as they are determined based on the position of observation, the others are called Algebraic Averages.

4.4.1 Simple Arithmetic Mean:

The Arithmetic Mean is the most commonly used and readily understood measure of central tendency. We find the arithmetic mean or mean by summing the values of all observations and dividing by the number of observations.

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots\cdots\cdots + x_n}{N}$$

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

This average is denoted as where $\bar{X}$ indicates the arithmetic mean. $\sum x$ Indicates the sum of all the observations and N, the number of observations considered.

For example, let us consider the annual consumption of edible oil by 10 families in a locality.

26, 32, 16, 21, 40, 31, 17, 20, 24, 37 Kgs.

If we calculate the arithmetic mean, then

$$\bar{X} = \frac{26 + 32 + 16 + 21 + 40 + 31 + 17 + 20 + 24 + 37}{10}$$

$$= \frac{264}{10} = 26 \cdot 4 \text{ Kgs.}$$

Average annual family consumption of edible oil is 26.4 Kgs.

In this illustration, the data is not grouped, hence we have considered observation directly. But in case of tabulated data, the observations will be classified according to certain range of the value and frequencies are determined. In such cases a mid point of the class interval would be treated as representative value of that class and it is multiplied by the frequency. The calculation of arithmetic average would be done by using the following formula.

$$\bar{X} = \frac{\sum f x}{n}$$

where,

$\bar{X}$ = Arithmetic Mean

f = Frequency (number of observations) in each class

x = Middle value of each class interval

n = Total number of observations i.e., $\sum f$

Example 1:

This method is illustrated for the following data which relate to the monthly sales of 200 firms.

Monthly Sales (Rs. Thousand)	No. of Firms	Monthly Sales (Rs. Thousand)	No.of Firms
300 - 350	5	550 - 600	25
350 - 400	14	600 - 650	22
400 - 450	23	650 - 700	7
450 - 500	50	700 - 750	2
500 - 550	52		

For computation of arithmetic mean, we need the following table:

Monthly Sales (Rs. Thousand)	Mid Point X	No. of firms f	fx
300 - 350	325	5	1625
350 - 400	375	14	5250
400 - 450	425	23	9775
450 - 500	475	50	23750
500 - 550	525	52	27300
550 - 600	575	25	14375
600 - 650	625	22	13750
650 - 700	675	7	4725
700 - 750	725	2	1450
		N = 200	$\sum fx = 102000$

$$\bar{X} = \frac{\sum fx}{N} = \frac{10200}{200} = 510$$

Hence the average monthly sales are Rs. 510

Method of Arbitrary Average:

When the calculation of all arithmetic by hand by ourselves we can further simplify our calculation of mean from grouped data. Considering the value of an observation as an arbitrary average the mean value can be arrived at scaling the values of other observations. The principle on which the above method is based on that the algebraic sum of deviations from actual average is equal to Zero. The deviations from an arbitrary average are averaged

and used for arriving at the correct average. Generally, the middle value is chosen as average and deviation of other classes are arrived from it. If the width of the class interval is same, then the deviations can further be shortened by sealing them down with the value of the width of the class interval.

When observation class - interval are scalled down, the formula for calculation of Arithmetic Mean would be

$$\bar{X} = A + \frac{\sum df}{n} \times i$$

Where:

A = Arbitary Average

d = Scalled down value for each class interval

f = Frequency of class interval

n = Total number of observations

i = Width of class interval

Example 2:

This formula makes the computations very simple and takes less time. To apply this formula, let us consider the same example discussed earlier and shown again in the following table:

Monthly Sales (Rs. Thousand)	Mid Point	No.of Firms f	(X - 525)/50 = d	fd
300 - 350	325	5	- 4	- 20
350 - 400	375	14	- 3	- 42
400 - 450	425	23	- 2	- 46
450 - 500	475	50	- 1	- 50
500 - 550	525	52	0	0
550 - 600	575	25	+ 1	+ 25
600 - 650	625	22	+ 2	+ 44
650 - 700	675	7	+ 3	+ 21
700 - 750	725	2	+ 4	+ 8
		N = 200		$\sum$ fd = 60

$$\bar{X} = A + \frac{\sum fd}{N} \times i = 525 \quad \frac{60}{200} \times 50$$

= 525 - 15 = 510 or Rs. 510

If may be observed that this formula is much faster than the previous one and the value of arithmetic mean remains the same.

Merits and demerits of Arithmetic Mean:

The arithmetic mean's calculation is simple and the procedure is familiar to most people.

As the arithmetic value considers all the values in a distribution, it is considered to be more representative of this distribution.

Arithmetic mean is unique to a given data and it is rigidly defined so that different interpretation by different persons are not possible. Arithmetic mean lends support for further mathematical treatment. It is used in the computation of other measures such as standard deviation, coefficient of skewness.

Although means is reliable and it takes into consideration all the observations. However, it is affected by extreme values in the data which sometimes distort the representative character of the mean.

It is not possible to calculate arithmetic mean in case of open ended class intervals.

4.4.2 Weighted Arithmetic Mean:

Simple average assume equal importance to all the values or size of items in a given distribution. But in practice sometimes we may have to give more importance to some compared to others. For example, in construction of commodity price index the prices of all the commodities are not of equal importance, as their consumption differes. Similarly when a firm wishes to ascertain the average cost of producing one unit, the weightages are to be considered for number of skilled labour hours used and unskilled labour hours used.

In weighted average the component items are being multiplied by certain values known as "Weights" and the aggregrate multiplied product is being divided by the total sum of their weights instead of the total number of observations.

$$\bar{X} = \frac{w_1x_1 + w_2x_2 + \cdots\cdots\cdots + w_nx_n}{w_1 + w_2 + \cdots\cdots\cdots + w_n}$$

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

Example 3:

Suppose you got 80 in the final examination, 95 in the first mid term assignment, as 85 in the second mid term assignment then

$$\bar{X}_w = \frac{50(80) + 25(95) + 25(85)}{100}$$

$$= \frac{4000 + 2375 + 2125}{100} = \frac{8500}{100} = 85$$

The following table shows this computation in a tabular form which is easy to employ for calculation of weighted arithmetic mean.

	Grade X	Weight W	WX
Final Examination	80	50	4000
First assignment	95	25	2375
Second assignment	85	25	2125
		$\sum W = 100$	$\sum WX = 8500$

$$\bar{X}_w = \frac{\sum WX}{\sum W} = \frac{8500}{100} = 85$$

The concept of weighted arithmetic mean is important because the computation is the same as used for averaging ratios and determining the mean of grouped data. Weighted mean is specially useful in problems relating to the construction of index numbers.

4.4.3 Median:

A second important measure of central tendency is the Median. Median differs from other means as it is determined based on the position of an time. It is a single value from the data set that measures the central item in the data. Median is that value which divides the distribution into two equal parts. Fifty percent of the observations in the distribution would be above the value of median and fifty percent of the observations would be below Median. So Median calculation requires the data to be sorted and arranged in order of size or magnitude.

When the data is arranged either in ascending or descending order, the Median can be determined based on the following formula.

$$\text{Median} = \left[\frac{N+1}{2}\right] \text{th item's value.}$$

If the data set contains an odd number of times, the middle item of the arranged data would become the Median. If there are even number of items, the median would be the average of the two middle items.

In case of grouped data the identification of Median is not possible with observation. Generally, in grouped data the value of variable under our consideration is presented in ranges called class intervals. The number of observations falling within such range is indicated as frequency. Of course one greatest advantage i grouped data is that the values of the variable are automatically arranged in order.

Since the sum total of frequencies indicate the total observations in the data set $\left(\frac{N}{2}\right)$th item gives the class interval in which Median falls. in a continuous distribution $\left(\frac{N}{2}\right)$th item

gives the middle value. $(N+1)/2$ is generally not considered for the reason that in a continuous distribution N + 1 class limits can make N class divisions.

Having identified the class in which the median value fall, the exact value is interpolated within that class by using the following formula.

$$\text{Median} = L_1 + \frac{\frac{N}{2} - CF}{F} \times I$$

Where,

L_1 = Lower limit of the class interval where Median is likely to fail.

$\frac{N}{2}$ = Item whose value is to be interpolated.

CF = Cumulative Frequency upto the Median Class.

F = Frequency of the Median Class.

I = Width of the class interval.

Example 4:

As an illustration, consider the following data which relate to the age distribution 1000 workers in an industrial establishment.

Age (Years)	No.of Workers	Age (Years)	No.of Workers
Below 25	120	40 - 45	150
25 - 30	125	45 - 50	140
30 - 35	180	50 - 55	100
35 - 40	160	55 and above	25

Determine the median age.

Calculation of Median value is facilitated by the use of a cumulative frequency division as shown below in the table.

Age (Years)	No.of Workers (f)	Cumulative Frequency (c f)
Below 25	120	120
25 - 30	125	245
30 - 35	180	425
35 - 40	160	585
40 - 45	150	735
45 - 50	140	875
50 - 55	100	975
55 and above	25	1000
	N = 1000	

Median = size of $\frac{N}{2}$ th observation $= \frac{1000}{2} = 500$ th observation which lies in the class 35 - 40.

$$\text{Median} = L + \frac{N/2 - cf}{F} \times I = 35 \ \frac{500 - 425}{160} \times 5$$

$$= 35 + \frac{375}{160} = 35 + 2 \cdot 34 = 37 \cdot 34 \text{ years.}$$

Hence the median age is approximately 37 years. This value of median suggests that half of the workers are below the age of 37 years and other half of the workers are above the age of 37 years.

Mathematical Property of Median:

The important mathematical property of the median is that the sum of the absolute deviations about the median is a minimum. In symbols $\sum |X - \text{Med} \cdot| =$ a minimum.

Although the median is not as popular as the arithmetic mean, it does have the advantage of being both easy to determine and easy to explain.

As illustrate earlier, the median is affected by the number of observations rather than the values of the observations, hence it will be less distroted as a representative value than the arithmetic mean.

An additional advantage of the median is that it may be computed for an open - end distribution.

The major disadvantage of median is that it is a less familiar measure than the arithmetic mean. However, since median is a positional average, its value is not determined by each and every observation. Also median is not capable of algebraic treatment.

Quantiles:

Quantiles are the related positional measures of central tendency. These are useful to frequently employed measures of non - central location. The most familiar quantities are the Quartiles, deciles and percentiles.

Quartiles: Quartilesare those values which divide the total data into four equal parts. Since three points divide the distribution into four equal parts, we shall have three qyartiles. Let us call them Q_1, Q_2 and Q_3. The first quartile, Q_1 is the value such that 25% of the observations are smaller and 75% of the observations are larger. The second quartile Q_2, is the median i.e., 50% of the observations are smaller and 50% are larger. The third quartile Q_3, is the value such that 75% of the observations are smaller and 25% of the observations are larger.

For grouped data, the following formulas are used for quartiles.

$$Q_j = L + \frac{jN/4 - pcf}{f} \times i \quad \text{for } j = 1, 2, 3$$

where L is lower limit of the quartile class, pcf is the preceding cumulative frequency to the quartile class, f is the frequency ofthe quartile class, and i is the size of the quartile class.

Deciles: Deciles are those values which divide the total data into ten equal parts. Since nine points divide the distribution into ten equal parts, we shall have nine deciles denoted by

$D_1, D_2, \cdots\cdots\cdots, D_9$

For grouped data, the following formulas are used for deciles:

$$D_k = L + \frac{KN/10 - pcf}{f} \times i \quad \text{for } K = 1,2,.....,9.$$

where the symbols have usual meaning and interpretation.

Percentiles: Percentiles are those values which divide the total data into hundred equal parts. Since ninety nine points divide the distribution into hundred equal parts we shall have ninety nine percentiles denoted by

$$P_1, P_2, P_3, \cdots\cdots\cdots, P_{99}$$

For grouped data, the following formulas are used for percentiles.

$$P_1 = L \ \frac{IN/100 - pcf}{f} \times i \quad \text{for } I = 1, 2, ..., 99$$

To illustrate the computations of quartiles, deciles and percentiles, consider the following grouped data which relate to the profits of 100 companies durign the year 1987 - 88.

Profits (Rs. Lakhs)	No.of Companies	Profits (Rs. Lakhs)	No.of Companies
20 - 30	4	60 - 70	15
30 - 40	8	70 - 80	10
40 - 50	18	80 - 90	8
50 - 60	30	90 - 100	7

Calculate Q_1, Q_2 (median), D_6, and P_{90}, from the given data and interpret these values. To compute Q_1, Q_2, D_6 and P_{90} we need the following table.

Profits (Rs. Lakhs)	No.of Companies (f)	c.f
20 - 30	4	4
30 - 40	8	12
40 - 50	18	30
50 - 60	30	60
60 - 70	15	75
70 - 80	10	85
80 - 90	8	93
90 - 100	7	100

Size of $N/4$ th observation = $\frac{100}{4} = 25$ th observation, which lies in the class 40 - 50.

$$Q_1 = L + \frac{N/4 - pcf}{f} \times i = 40 + \frac{25 - 12}{18} \times 10 = 40 + 7 \cdot 22 = 47 \cdot 22$$

This value of Q_1 suggests that 25% of the companies earn an annual profit of Rs. 47.22 lakh or less.

Median or Q_2 = Size of $\frac{N}{2}$ th observation = $\frac{100}{2} = 50$ th observation.

which lies in the class 50 - 60.

$$Q_2 = L + \frac{N/2 - pcf}{f} \times i = 50 + \frac{50 - 30}{30} \times 10 = 50 + 6 \cdot 67 = 56 \cdot 67$$

This value of Q_2, (or median) suggests that 50% of the companies earn an annual profit of Rs. 56.67 lakh or less and the remaining 50% of the companies earn an annul profits of Rs. 56.67 lakh or more.

D_6 = size of $\frac{6N}{10}$ th observation = $\frac{6 \times 100}{10} = 60$ th observation which lies in the class 50 - 60.

$$D_6 = L + \frac{6N/10 - pcf}{f} \times i = 50 + \frac{60 - 30}{30} \times 10 = 50 + 10 = 60$$

Thus 60% of companies earn an annual profit of Rs. 60 lakh or less and 40% of the companies earn Rs. 60 lakh or more.

P_{90} = size of $\frac{90\,N}{100}$ th observation = $\frac{90 \times 100}{100} = 90$ th observation, which lies in the class 80 - 90.

$$P_{90} = L + \frac{90N/100 - pcf}{f} \times i = 80 + \frac{90 - 85}{10} \times 10 = 80 + 5 = 85$$

This value of 90th percentile suggests that 90% of the comapnies earn an annual profit of Rs. 85 lakh or less and 20% of the companies earn more than Rs. 85 lakh or more.

Locating The Quantiles Graphically:

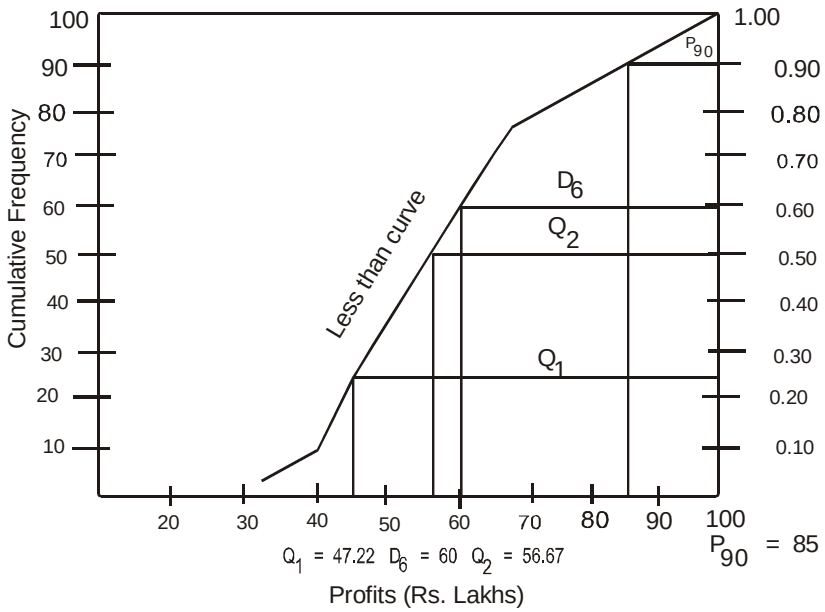
To locate the median graphically, draw less than cumulative frequency cureve cless than ogive). Take the variable on the X - axis and frequency on the Y - axis Determine the Median value by locating $N/2$ th observation on the Y - axis. Draw a horizontal line from this on the cumulative frequency curve and from where it meets the curve draw a perpendicular on the X - axis. The point where it meets the X - axis is the value of median.

Similarly we can locate graphically the other quantiles such as quartiles, deciles and percentiles.

For the data of previous illustration, locate graphically the value of Q_1, Q_2, D_{60} and Q_{90}.

The first step is to make a less than cumulative frequency curve as shown in figure 1.

Fig 1: Cumulative Frequency Curve



To determine different quantiles graphically, horizontal lines are drawn from the cumulative relative frequency values. For example if we want to determinethe value of median (or Q_2), a horizontal line can be drawn from the cumulative frequency value of 0.50 to the less than curve and then extending the vertical line to the horizontal axis. In a similar way, other values can be deternined as shown in the graph. From the graph, we observe

$$Q_1 = 47 \cdot 22,\ Q_2 = 57 \cdot 67,\ D_{60} = 60 \cdot 0,\ P_{90} = 75.$$

It may be noted that these graphical values of quantiles are the same as obtained by the formulas.

Merits and Demerits:

Median has superiority over Mean in respect of computational advantage. The extreme values do not affect the median so strongly. Median can be calculated even if open ended class intervals exist at either ends of a distribution.

As Median is an average based on position, we have to arrange the data before the calculation of it. Most statistical interference cannot be drawn with the help of median.

4.4.4 Mode:

Mode is a positional average. It refers to that value in a data set which occurs most of the times. It can be designed as a point around which most items tend to concentrate. It has greater use in Business situations. Suppose a marketing manager wants to know about he fast moving size of a garment or a shoe, mode helps him.

In a continuous data distribution, mode presents the a symmetry of concentration of frequencies around a particular value. One can easily locate mode based on the point of a distribution curve.

In case of individual items mode is the value of the item which repeats most of the times. Similarly in a grouped data, mode can be located in class interval which is having highest frequency. But the determination of mode value requires us to take the relative forces of the neighbouring frequencies on mode class. It can be ascertained by the following formula.

$$\text{Mode} = L_1 + \frac{d_1}{d_1 + d_2} \times I$$

Where,

L_1 = Lower limit of the Model Class

d_1 = Differences in Frequency between Mode Class and its preceding class

d_2 = Difference in the frequency between Mode Class and its Later class

Example 5:

Daily Sales (Rs. Thousand)	No.of Firms	Daily Sales (Rs. Thousand)	No.of Firms
20 - 30	15	60 - 70	35
30 - 40	23	70 - 80	25
40 - 50	27	80 - 90	5
50 - 60	20		

Since the maximum frequency 35 is in the class 60 - 70, therefore 60 - 70 is the modal class. Applying the formula, we get

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2} \times i = 60 + \frac{35 - 20}{(35-20) + (35-20)} \times 10$$

$$= 60 + \frac{150}{25}$$

= 60 + 6 = Rs. 66

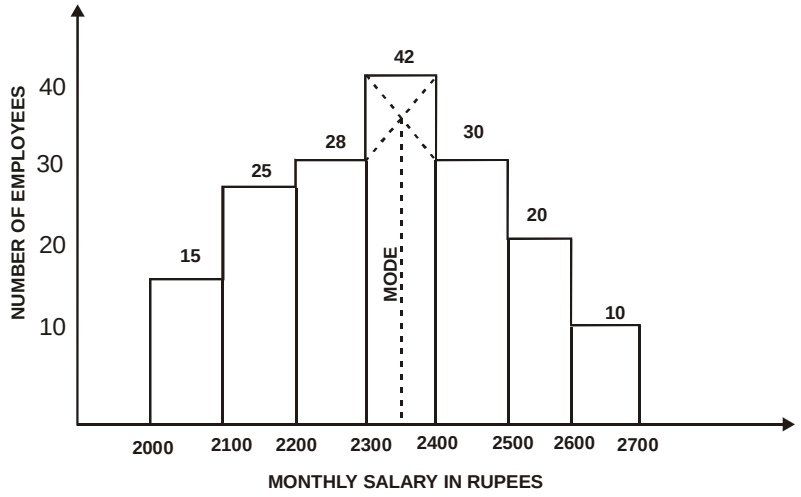
Hence modal daily sales are Rs. 66

To better understand the formula, let us try to locate mode in the following figure of frequency bar graph.

Consider the following data to do cate the value of made graphically.

Monthly Salares Rs.	No.of Employees	Monthly Salary Rs.	No.of Employees
2000 - 2100	15	2400 - 2500	30
2100 - 2200	25	2500 - 2600	20
2200 - 2300	28	2600 - 2700	10
2300 - 2400	42		

Fig II. Histogram of Monthly Salaries



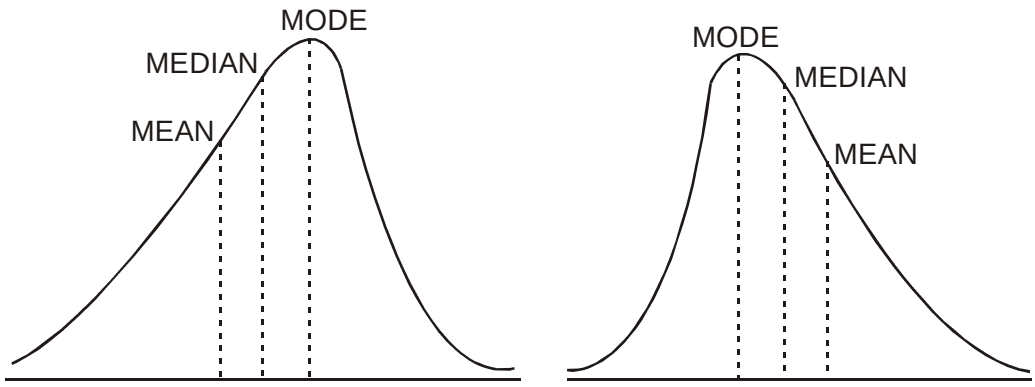
The two straight lines are drawn diagonally in the inside of the modal class bars and then finally a vertical line from the intersection of the two diagonal lines is drawn on the X - axis. Thus the modal value is approximately Rs. 2353. It may be noted that the value of mode would be approximately the same if we used the algebraic method.

The chief advantage of the mode is that it is by definition the most representative value of the distribution. For example, when we talk of modal size of shoe or garment, we have this average in mind. Like median, the value of mode is not affected by extreme values and its value can be determined in open - end distributions.

The main disadvantage of the mode is its indeterminate value i.e., we cannot calculate its value precisely in a grouped data, but merely estimate it. When a given set of data have two or more than two values as maximum frequency, it is a case of bimodal or multi modal distribution and the value of mode cannot be determined. The mode has no useful mathematical properties. Hence, in actual practice the mode is more important as a conceptual idea than as a working average.

Relationship Among Mean, Median and Mode:

In case of perfectly symmetrical distribution the arithmetic mean, mode and median should coincide each other. When distribution is positively skewed the distribution curve bends to left hand side, with Mean having the highest value, followed by Median and then by Mode. On the other hand, when the distribution is negatively skewed the curve bends right hand side and Mode having highest value, followed by Median and Mean.



But in a moderately skewed distribution, an interesting relationship is observed by the statisticians between Mean, Median and Mode. In such distributions it is proved that the distance between Mean and Median is approximately one - third of the distance between Mean Mode.

$$\text{Mean - Median} = \frac{1}{3}\text{(Mean - Mode)}$$

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

This formula would be helpful to ascertain Mode from Mean and Median at times of indeterminateness.

4.4.5 Geometric Mean:

Geometric Mean would be helpful in averaging ratios of changes. In order to ascertain average growth rate over several years in various types of business activities, the geometric mean is more appropriate one compared to simple arithmetic mean. It is said to be appropriate in computing the average rate of growth of population or average increase in the rate of sales, profits, production, gross national product.

It is calculated as follows:

$$\text{Geometric Mean} = \sqrt[n]{\text{Producto fall X values}}$$

$$G \cdot M \cdot = \sqrt[n]{X_1, X_2, X_3, \cdots\cdots\cdots, X_n}$$

Example 6:

A machine was purchased for Rs. 50,000 in 1984. Depreciation on the diminishing balance was charged @ 40% in the first year, 25% in the second year and 15% per annum during the next three years. What is the average depreciation charged during the whole period?

Since we are interested in finding the average rate of depreciation, geometric mean will be the most appropriate average.

Year	Diminishing value (for a value of Rs. 100) X	Log X
1984	100 - 40 = 60	1.77815
1985	100 - 25 = 75	1.87506
1986	100 - 15 = 85	1.92941
1987	100 - 15 = 85	1.92941
1988	100 - 15 = 85	1.92941
		Log X = 9.44144

$$GM = \text{Anti log}\left(\frac{\sum \log X}{N}\right)$$

$$= \text{Anti log}\left(\frac{9 \cdot 44144}{5}\right) = \text{Antilog } 1.8883 = 77.32$$

The diminishing value being Rs. 77.32, the depreciation will be 100 - 77.32 = 22.68%

The geometric mean is very useful in averaging ratios and percentages. It also helps in determining the rates of increase and decrease. It is also capable of further algebraic treatment, so that a combined geometric mean can easily be computed.

However, compared to arithmetic mean, the geometric mean is more difficult to compute and interpret. Further, geometric mean cannot be computed if any observation has either a value zero or negative.

4.4.6 Harmonic Mean:

The Harmonic Mean is a measure of central tendency for the data expressed as rates such as kilometers per hour, tonnes per day, kilometers per litre etc. Interestingly, this average is more suitable for finding out an average speed of a vehicle, average speed with which a work is done. The Harmonic Mean is defined as the reciprocal of the arithmetic mean of the reciprocal of the individual observations.

The arithmetic mean of the reciprocals of N values.

$X_1, X_2, X_3, \ldots\ldots\ldots, X_n$ is

$$\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}{N}$$

The Harmonic Mean is the reciprocal of the arithmetic mean of the above.

$$\text{Harmonic Mean} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Example 7:

The explain the computational procedure, let us consider the following example.

In a factory, a unit of work is completed by A in 4 minutes, By B in 5 minutes, by C in 6 minutes, By D in 10 minutes and By E in 12 minutes. Find the average number of units of work completed per minute.

The calculations for computing harmonic mean are given below:

X	1/X
4	0.250
5	0.200
6	0.167
10	0.100
12	0.083
	$\sum 1/x$ = 0.8

Hence the average number of units computed per minute is 6.25

The harmonic mean like arithmetic mean and geometric mean is computed from each and every observation. It is specially useful for averaging rates.

However, harmonic mean cannot be computed when one or more observations have zero value or when there are both positive or negative observations. In dealing with business problems, harmonic mean is rarely used.

Although, the Geometric Mean and Harmonic Means have equal and even more advantage in computation to certain areas of study compared to simple average, care should be taken to verify whether any observation is having a value equal to zero. In such a case, if totally distorts the result we are likely to arrive at.

This averages in general try to arrive a representative figure to understand the their characteristic of the variable under consideration.

4.5 Summary:

The arithmetic mean is widely used and under stood as a measure of central tendency. The concepts of weighted arithmetic mean, geometric mean, and harmonic mean are useful for specific type of applications. The median is generally a more representative measure for open-end distribution and highly skeueed distribution. The made should be used when the most demanded.

4.6 Exercises:

1. List the various measures of central tendency studied in this Lesson and explain the difference between them.
2. Discuss the mathematical properties of arithmetic men and median.
3. Review for each of the measure of central tendency, their advantages and disadvantages.
4. Explain how you will decide which average to use in a particular problem.
5. How do you account for the predominant choice of arithmetic mean of statistical data as a measure of central tendency? Under what circumstances would it be appropriate to use mode or median?
6. Define median and mode with examples. Show how can they be calculated in case of discrete value?
7. Under what circumstances are Geometric Mean and Median considered to be most suitable measures for describing the central tendency of a frequency distribution.
8. Explain when will you use the following:
   a. Mode in place of Arithmetic Mean
   b. Geometric Mean in place of Arithmetic Mean
   c. Arithmetic Mean in place of Median

9. What are quantiles? Explain and illustrate the concepts of quartiles, deciles and percentiles.

10. Following is the cumulative frequency distribution of preferred length of study table obtained from the preference study of 50 students.

Length	No.of Students	Length	No. of Students
more than 50 cms	50	more than 90 cms	25
more than 60 cms	46	more than 100 cms	18
more than 70 cms	40	more than 110 cms	7
more than 80 cms	32		

A manufacturer has to take decision on the length of study - table to manufacture. What length would you recommend and why?

11. A three month study of the phone calls received by Small Company yielded the following information:

Number of Calls per day	No.of.Days	Number of Calls per day	No.of Days
100 - 200	3	600 - 700	10
200 - 300	7	700 - 800	9
300 - 400	11	800 - 900	8
400 - 500	13	900 - 1000	4
500 - 600	27		

Compute the arithmetic mean median and mode.

12. In the following distribution of travel time of 213 days to work of a firm's employee, find the modal travel time.

Travel Time (in minutes)	No.of Days	Travel Time (in minutes)	No.of Days
Less than 80	213	Less than 40	85
Less than 70	210	Less than 30	50
Less than 60	195	Less than 20	18
Less than 50	156	Less than 10	2

13. The mean monthly salary paid to all employees in a company is Rs. 1600. The mean monthly salaries paid to technical employees are Rs. 1800 and Rs. 1200 respectively. Determine the percentage of technical and non - technical employees of the company.

14. The following distribution is with regard to weight (in grams) of apples of a given variety. If an apple of less than 122 grams is to be considered unsuitable for export. What is the percentage of total apples suitable for the export?

Weight (in grams)	No.of Apples	Weight (in grams)	No.of Apples
100 - 110	10	140 - 150	35
110 - 120	20	150 - 160	15
120 - 130	40	160 - 170	5
130 - 140			

Draw an ogive of more than one type and deduce how many apples will be more than 122 grams.

15. The geometric mean of 10 observations on a certain variable was calculated to be 16.2. It was later discovered that one of the observations was wrongly recorded as 10.9. When in fact it was 21.9. Apply appropriate correction and calculate the correct geometric mean.

16. An incomplete distribution of daily sales (Rs. Thousand) is given below. The data relate to 299 days.

Daily Sales (Rs. Thousand)	No.of Days	Daily Sales (Rs. Thousands)	No.of Days
10 - 20	12	50 - 60	?
20 - 30	30	60 - 70	25
30 - 40	?	70 - 80	18
40 - 50			

You are told that the median value is 46. Using the median formula, fill up the missing frequencies and calculate the arithmetic mean of the complete data.

17. The following table shows the income distribution of a company.

Income (Rs.)	No.of Employees	Income (Rs.)	No.of Employees
1200 - 1400	8	2200 - 2400	35
1400 - 1600	12	2400 - 2600	18
1600 - 1800	20	2600 - 2800	7
1800 - 2000	30	2800 - 3000	6
2000 - 2200	40	3000 - 3200	4

Determine: (i) the mean income (ii) the median income (iii) the mean (iv) the income limits for the middle 50% of the employees (v) D_7, the seventh decile and (vi) P_{80} the eightieth percentile.

4.7 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 5

MEASURES OF VARIATION AND SKEWNESS

Objectives:

After going through this lesson, you will learn the concept of variation in its absolute and relative sense to compute range, quartile deviation, mean deviation and standard deviation the computation of skewness.

Structure:

5.1 Introduction:

In this lesson, we shall discuss the concept of variation and skewness. A measure of variation (or dispersion) describes the spread or scattering valus. Different sets of data may have thesame measure of central tendency but differ greatly in terms of variation. The skewness shows the lack of symmetry in distribution.

5.2 Averages Have A Limitation:

Average or Measures of Central Tendency, viz., Mean, Median and Mode identify a single figure from an array of data points or in a tabulated data distribution. That single figure gives us the idea of the concentration of observations around a central value of the distribution. However, these measures do not tell us how the individual items have been distributed around the value and whether such deviations of individual items from the average are large - or small.

Let us consider the following example:

Details of daily sales of three Retail Shops During a Week

Retail Shop A Daily Sales (Rs.)	Retail Shop B Daily Sales (Rs.)	Retail Shop C Daily Sales (Rs.)
3,500	3,000	2,000
3,500	4,000	3,800
3,500	3,500	2,200
3,500	3,750	4,100
3,500	3,250	3,600
3,500	3,500	5,300
$\bar{X}$ =3,500	$\bar{X}$ =3,500	$\bar{X}$ = 3,500

In this example, all the three shops show an average daily sales of Rs. 3,500. But if we carefully observe the daily sales figures individually, Shop A has similar sales everyday while in other two, there are wide fluctuations. The fluctuations are larger in Shop C compared to Shop B. Managers do have some implications for these fluctuations. But the averages are insufficient measures to present such characteristics. Therefore new measures are to be supplemented. On such measure is 'Dispersion'.

Need For Measures of Dispersion:

Dispersion indicates the extent to which an individual value of a data falls away from the average or central value. The measure of dispersion helps in identifying the homogenity, heterogenity of the data in a distribution.

5.3 Significance of Measures of Dispersion:

Measuring the dispersion or variation in a given data distribution would help the following situations:

(a) It provides additional information to determine the reliability of an average and explains how far an average is representative of the entire data.

(b) The measurement of variation would help in analyzing the nature and causes of variation in the basic characteristic of the variable under study. It would help in considering steps to control the variability.

(c) Measure of variance is helpful in comparing two or more distributions with regard to their variability.

(d) Measuring variability is of great importance to advanced statistical analysis. It has a greater role in sampling and statistical inferences.

Business Applications:

The measurement of variability has greater significance in certain functional areas of Business Management. For example, a Portfolio or Investment Manager of a large mutual fund company is interested in dispersion of a firm's earnings.

5.4 Measures of Variation:

Some of the well known measures of variations are as follows:

(i) Range

(ii) Quartile Deviation

(iii) Mean Deviation

(iv) Standard Deviation

Absolute and Relative:

Generally, the dispersion in a data is presented in absolute or in relative terms. The absolute measures of dispersion are expressed in the same statistical units in which the original data are given, such as Rupees, Kilograms, Tonnes etc. These values could be used to compare the variation in two distributions provided the variables are expressed in the same units and of the same average size. In case the two sets of data are expressed in different units of measurements, the absolute measure is not comparable. In such a situation a relative measure of dispersion would be useful. It is also called the co - efficient of dispersion / variation.

$$\text{Coefficient of Dispersion} = \frac{\text{Absolute Dispersion}}{\text{Measure of Central Value from which deviations are taken}}$$

5.4.1 Range:

One of the simplest methods of measures of dispersion is to find the difference between the highest and the lowest values in the data series. The difference between two extreme value is called 'Range'.

$$\text{Range} = X_{max} - X_{min}$$

where X_{max} : Maximum Value

X_{min} : Minimum Value

A relative measure to range is called "Coefficient of Range" and it is obtained by the following formula.

$$\text{Coefficient of Range} = \frac{X_{max} - X_{min}}{X_{max} - X_{min}}$$

If the "Means" of the distributions are same, a distribution with smallest range indicates less dispersion and the Mean of such a distribution is more typical of that group.

For the earlier example, the absolute and relative ranges of dispersion can be shown as follows:

Calculated Range Values

	Shop A	Shop B	Shop C
Average	3500	3500	3500
Range	3500 - 3500	4000 - 3000	5300 - 2000
Max. - Min.	= 0	= 1000	= 3300
Coefficient of Range = $\frac{X_{max} - X_{min}}{X_{max} - X_{min}}$	= 0	= 0.14	= 0.45

The example shows that there is no variation in the retail shop A's daily sales, the variation is small in case of shop B and variation is very large in shop C.

Merits and Demerits:

The range is very easy to understand, but its usefulness as a measure of dispersion is limited. It is mostly used in preparing Control Charts by Quality Control Divisions. For Metereological departments, the range is a good indicator for Weather Forecast.

However, the Range is not based on the entire set of data and depends on only two extreme observations. Range is also much affected by fluctuations in extreme observations. In case of grouped data range can be approximated to the difference between upper limit of the largest class and lower limit of the lowest class.

5.3.2 Quartile Deviation:

Inter - Quartile Range:

The range as a measure of dispersion is based on only two extreme items and it fails to take into account the scatters within the range. To tackle this problem, 'Inter - quartile range' is in wide use. It measures approximately how far from the median we must go on either side before we can include one - half of the values of the given data set.

Inter-quartile range = value of quartile 3 - value of quartile 1

$$= Q_3 - Q_1$$

Quartiles divide the data into four equal parts each containing 25 percent of the data. Inter - quartile range considers the middle 50 percent of the data and thereby avoids the calculation of spread based on extreme values.

The quartile deviation, expressed from inter - quartile range, is the average of the differences between third quartile value and the first quartile.

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

The quartile deviation is more representative of the spread of the data since it considers the middle 50 percent of the data. For comparision of variability of two distributions, a relative measure known a as Coefficient of Quartile Deviation is generally worked out using the following formula.

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

Calculation of Quartiles:

Median divides the data into two equal parts. Similar to Median, there are other measures which divide the data into certain equal parts. Among them, 'Quartiles' divide the data into four equal parts, 'Deciles' divide the data into 10 equal parts and 'Percentiles' divide the data into 100 equal parts.

Quartiles, Deciles, Percentiles:

The computation of Quartiles, Deciles and Percentiles uses the similar formulae used in the calculation of Median. For example, the Quartiles are calculated as per the following formula:

In Discrete data series:

$$Q_1 = \text{size of the } \frac{N+1}{4}\text{th item of the data series}$$

$$Q_3 = \text{size of the } \frac{3(N+1)}{4}\text{ th item of the series}$$

Derivation of Quartitles:

In case of tabulated / grouped data

Q_1 = Size of the N/4th item. As the data is arranged against different class - intervals, the size of the item to be ascertained from

$$Q_1 = L_1 + \frac{\frac{N}{4} - CF}{F} \times i$$

where L_1 = Lower limit of the quartile class

N = Size of the sample or total frequency

CF = Cumulative Frequency upto the quartile class

F = Frequency of the quartile class

i = Width of the class - interval

Similarly,

$$Q_3 = L_1 + \frac{\frac{3N}{4} - CF}{F} \times i$$

Example 1:

The Export Promotion Council is considering the variations Chilli exports from India so as to suggest a policy to stabilise to the Chilli Production. The Council shows the following figures.

1991 - 92	92 - 93	93 - 94	94 - 95	95 - 96	96 - 97	97 - 98	98 - 99	99 - 2000
239	201	163	241	266	237	225	239	284

Let us calculate Quartile Deviation for the data. So as to, identify the quartiles, we have to arrange the data in either ascending or decending order:

Sl. No.	1	2	3	4	5	6	7	8	9
Ordered Data:	163	201	225	237	239	239	241	266	284

$\uparrow Q_1$ $\uparrow Q_3$

Quartile 1 = $Q_1 = \frac{N+1}{4}$ th item's value

$= \frac{9+1}{4} = \frac{10}{4} = 2 \cdot 5$th item's value

$= 201 + 1/2\,(225 - 201) = 213$

Quartile 3 = $Q_3 = \frac{3(N+1)}{4}$ th item's value

$= \frac{3(9+1)}{4} = \frac{10}{4} = 7 \cdot 5$th item's value

$= 241 + 1/2\,(266 - 241) = 253 \cdot 5$

Quartile deviation

$$= \frac{Q_3 - Q_1}{2} = \frac{253 \cdot 5 - 213}{2} = 20 \cdot 25$$

On an average the year - to - year exports are varying to the extent of Rs. 20.25 crores. The Coefficient of Quartile Deviation is:

$$\text{Coefficient of } Q\cdot D = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{253\cdot 5 - 213}{253\cdot 5 + 213}$$

$$= \frac{40\cdot 5}{466\cdot 5} = 0\cdot 08681$$

It indicates that, on an average, the variation works out to the extent of 8.7 percent of the average exports.

Example 2:

In a medium sized construction enterprise, the Personnel Manager is interested in knowing the dispersion in wages received by the unskilled laboures. He has considered the daily wages paid to the workers and tabulated the data as follows:

Weekly Wage Distribution

Wage in Rs. Per day	No. of Workers
Less than Rs. 35	14
35 - 37	62
37 - 40	99
41 - 43	18
43 above	7

One of the appropriate measures of dispersion in case of open ended class - intervals is the Quartile Deviation. Let us calculate the quartile deviation to the above example.

Calculation of Quartiles for Weekly Wages

Wages per day (Rs.)	Adjusted Class intervals (Rs.)	No.of Workers (f)	Cumulative Frequency (CF)
Less than 35	< 34.5	14	14
35 - 37	34.5 - 37.5	62	76 Q_1 Class
37 - 40	37.5 - 40.5	99	175 Q_3 Class
41 - 43	40.5 - 43.5	18	193
Above 43	43.5 above	7	200

Q_1 = value of $\frac{200}{4}$th item i.e., 50th item

50th item falss in 34.5 - 37.5 class interval

$$Q_1 = L_1 + \frac{\frac{N}{4} - CF}{F} \times I$$

$$= 34 \cdot 5 + \frac{50 - 14}{62} \times 3$$

= 36.24

$$Q_3 = \text{Value of } 3\left[\frac{200}{4}\right]\text{th item i.e., 150th item}$$

150th item falls in 37.5 - 40.5 class interval

$$Q_3 = 37 \cdot 5 + \frac{3(50) - 76}{99} \times 3$$

= 39.74

$$\frac{Q_3 - Q_1}{2} = \frac{39 \cdot 74 - 36 \cdot 24}{2}$$

= 1.75

It indicates that, on an average the dispersion of wages among middle 50 workers is Rs. 1.75

5.4.3 Mean Deviation:

The measure of mean (average) deviation is an improvement over the previous measures and this measure considers all the observations in the data. Mean deviation is the arithmetic average of the d eviations of all the items from an average (either it may be Mean, Median or Mode). It reflects the average amount of scatter of the items in a distribution. With a purpose to measure the scatter, this measure ignores the signs of deviations while calculating the deviations from average.

The procedure involved in calculation of the Mean Deviation is simple. First we must ascertain any one of the averages for the given data. Then deviation of all the individual items from the said average - Mean or Median.

Mode are to considered in their absolute size. The mean deviation is the average of such absolute deviations calculated from a Mean.

$$\text{Mean Deviation} = \frac{\sum \left| X_i - \bar{X} \right|}{N}$$

where X = individual item in a data set

$\bar{X}$ = Mean of the data

N = Total number of items

| | = Absolute sizes

When mean deviation is calculated from other measures of central tendency.

$$\text{Mean Deviation} = \frac{\sum |X_i - \text{Median}|}{N} \quad \text{or}$$

$$\text{Mean Deviation} = \sum \frac{|X - \text{Mode}|}{N}$$

In case of grouped data, the Mean deviation is ascertained by the following formula:

$$\text{Mean Deviation} = \frac{\sum f |X_1 - \text{Mean}|}{\sum f}$$

Relative Measure:

Mean Deviation is an absolute measure of dispersion. To express it in relative terms, the absolute mean deviation is to be divided by the measure of central value used for estimating such dispersion.

$$\text{Coefficient of Mean Deviation} = \frac{\text{Mean Deviation}}{\text{Mean}}$$

$$\text{or} \quad = \frac{\text{Mean Deviation}}{\text{Median}}$$

$$\text{or} \quad = \frac{\text{Mean Deviation}}{\text{Mode}}$$

As mean deviation can be calculated from any measure of average, it is also called an Average Deviation.

Example 3:

The following is the sample data of production rate of Plastic Overhead Tanks from M/s Gummadi Polymers (P) Lts.

16	20	17	26	18	22	21	23	19	24

the company's production manager feels that any average absolute deviation of more than 3 tanks a day indicates unacceptable production rate variation. So, should he be worried now? To answer, let us calculate Mean Deviation form Arithmetic Mean.

Calculation of Mean Deviation

Daily production rate of Tanks X	Deviation from Arithmetic Mean $X-\bar{X}$	Absolute Deviation in Production rate $\lvert X-\bar{X}\rvert$
16	16 - 21 = - 5	5
20	20 - 21 = - 1	1
17	17 - 21 = - 4	4
26	26 - 21 == 5	5
18	18 - 21 = - 3	3
22	22 - 21 = 1	1
25	25 - 21 = 4	4
23	23 - 21 = 2	2
19	19 - 21 = - 2	2
24	24 - 21 = 3	3
$\sum x = 210$		$\sum\lvert X-\bar{X}\rvert = 50$

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

$$= \frac{210}{10} = 2$$

$$\text{Mean Deviation} = \frac{\sum\lvert X-\bar{X}\rvert}{N}\bar{X}$$

$$= \frac{50}{10} = 5$$

The sample shows a mean absolute deviation in production rate of 5 which is larger than the specified norm of 3, demanding necessary action.

Example 4: The Administrator of Appolo Hospitals conducted a survey on number of days that 200 randomly chosen patients stayed in their Madras based hospital following a major Operation.

Hospital Stay (in days)	1 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
Frequecny (f)	45	60	35	20	20	20

The Administrator is considering to plan for additional beds if the absolute deyiation of "Post - Operation Hospital Stay" is more than 10 days. In order to help the administrator, let us calculate the Mean Deviation for the data.

Calculation of Mean Deviation

Hospital Stay (in days)	Frequency f (f)	Mid - Value of class interval (X)	f x	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$
1 - 5	45	2.5	112.5	9.25	416.25
5 - 10	60	7.5	450.0	4.25	255.00
10 - 15	35	12.5	437.5	0.75	26.25
15 - 20	20	- 17.5	350.0	5.75	115.00
20 - 25	20	22.5	450.0	10.75	215.00
25 - 30	20	27.5	550.0	15.75	315.00

$$\sum fX = 2350, \quad \sum|X-\bar{X}| = 13425$$

Average No.of days of stay $(\bar{X}) = \frac{\sum fx}{\sum f} = \frac{2350}{200} = 11\cdot 75 \text{ days}$

Mean Deviation $= \frac{\sum f\left(X-\bar{X}\right)}{\sum f} = \frac{1342\cdot 5}{200} = 6\cdot 7 \text{ days}$

As the Mean Deviation is less than 10 days, the administrator need not initiate steps to increase the number of beds.

Merits and Demerits:

An outstanding advantage of the average deviation is its relative simplicity. Any one who is familiar with the concept of simple average can readily appreciate the procedure in calculating the average deviation. The main defect of this method is that it is not suitable for further mathematical treatment.

5.4.4 Standard Deviation:

The Standard Deviation is always computed around "Mean". Denoted by 'σ' (read as sigma). It is calculated by using the following formula.

$$\sigma = \sqrt{\frac{\sum(X-\bar{X})^2}{N}}$$

The square of the Standard Deviation is called Variance. Therefore, Variance = σ^2.

Formula for Grouped Data:

In case of grouped data, the Standard Deviation can be calculated by using the following formula:

$$\sigma = \sqrt{\frac{\sum f (X - \bar{X})^2}{N}}$$

Sometimes, when the mean contains decimal places the calculation becomes cumbersome. In such case, the Standard Deviation can be calculated even without finding the deviations from mean. The formula to use directly on original values is:

$$\sigma = \sqrt{\frac{\sum f X^2}{N} - \left[\frac{\sum f X}{N}\right]^2}$$

The standard deviation is regarded as the best measure of dispersion as it possess most of the qualities of an ideal measure of dispersion. It is based on all observations and least affected by sampling fluctuations. Therefore, it would help in testing the hypothesis and other tests of significance.

Example 5:

Shown below:

58 55 62 56 54 62 58 60 78 67

and calculate Standard Deviation to the data.

Average Speed of the Vehicles

Speed in KM PH **X**	**Deviation from Mean Speed** $X - \bar{X}$	**Squared Value of Deviations** $(X - \bar{X})^2$
58	58 - 61 = - 3	9
55	55 - 61 = - 6	36
62	62 - 61 = 1	1
56	56 - 61 = - 5	25
54	54 - 61 = - 7	49
62	62 - 61 = 1	1
58	58 - 61 = - 3	9
60	60 - 61 = - 1	1
78	78 - 61 = 17	289
67	67 - 61 = 6	36
$\sum X = 610$		$\sum (X - \bar{X})^2 = 456$

$$\bar{X} = \frac{\sum X}{N} = \frac{610}{10} = 61$$

Calculation of Standard Deviation

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

$$= \sqrt{\frac{456}{10}} = \sqrt{45 \cdot 6}$$

$$= 6.75$$

Example 6:

A manufacturer of Ready - Made Collars supplies a Tailor the following data regarding the neck - circumference of PG Level Students.

Neck Circumference (in inches)	12.0	12.5	13.0	13.5	14.0	14.5	15.0	15.5	16.0
Frequency	5	20	30	43	60	56	37	16	3

Considering the variations in neck circumference, the Tailor wishes to order those sizes to meet the needs of most of his customers. Let us help the Tailor to work out the Standard Deviation of neck circumference of PG Level Students. Further, as it is statistically proved that $X \pm 3\sigma$ covers 99% of the entire sample units, let us calculate the highest and lowest size for Neck Circumference so as to advise the tailor to acquire those Ready Made Collars.

Calculation of SD for Collar Size

Nick Circumferences (X)	**Frequency (f)**	**fX**	fX^2
12.0	5	60.0	720.00
12.5	20	250.0	3125.00
13.0	30	390.0	5070.00
13.5	43	580.0	7830.75
14.0	60	840.0	11760.00
14.5	56	821.0	11774.00
15.0	37	555.0	8325.00
15.5	16	248.0	3844.00
16.0	3	48.0	768.00

$\sum f = 270 \qquad \sum fX = 3783.55 \qquad \sum fX^2 = 53222 \cdot 75$

Standard Deviation with direct formula:

$$\sigma = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$

$$= \sqrt{\frac{53222 \cdot 75}{270} - \left(\frac{3783 \cdot 5}{270}\right)^2}$$

$$= \sqrt{0 \cdot 7581} = 0 \cdot 8707$$

$$\bar{X} = \frac{\sum fX}{N} = \frac{37835}{270} = 14 \cdot 01$$

Area Under Normal Curve:

If the data is normally distributed, the entire area under a frequency distribution curve can be measured in terms of standard deviation.

$\bar{X} \pm 1\sigma = 67\%$ of the area (approximately)

$\bar{X} \pm 2\sigma = 95\%$ of the area (approximately)

$\bar{X} \pm 3\sigma = 99\%$ of the area (approximately)

Theoretically the lowest size if

$$= \bar{X} - 3\sigma$$

$$\bar{X} - 3\sigma = 14 \cdot 01 - 3(0 \cdot 8707)$$

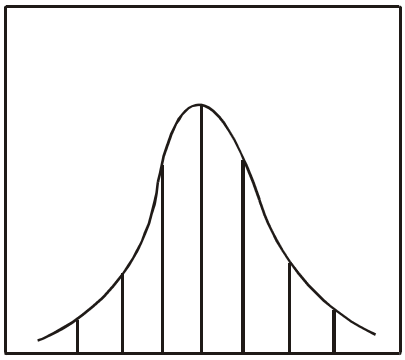
$$= 11 \cdot 39$$

Highest Size $= \bar{X} + 3\sigma$

$$= 14 \cdot 01 + 3(0 \cdot 8707)$$

= 16.62

Area Under Normal Curve



$-3\sigma \ -2\sigma \ -1\sigma \ \bar{X} \ +1\sigma \ +2\sigma \ +3\sigma$

To meet the majority needs, Tailor has to place orders for $\bar{X} \pm 1\sigma$ size, i.e.

Lowest size to be ordered

$= \bar{X} - 1\sigma$

$= 14 \cdot 01 - (0 \cdot 8707)$

= 13.14

Highest size to be ordered

$= \bar{X} + 1\sigma$

= 14.88

These sizes would suit to at least 67 percent of the customers of the Tailor.

5.5 Co-Efficient of Variation:

Co-efficient of variation is the relative measure of Standard Deviation. It is simply the ratio of Standard Deviation to Mean and expressed as percentage.

$$\text{Co-efficient of Variation} = \frac{\sigma}{\bar{X}} \times 100$$

When co-efficient of variation is less, then the variance is said to be less and the data is denoted to be consistent. This measure is of much help in making comparison between different distributions.

Example 7:

A company is considering to replace electric bulbs wherever the bulbs have burnt away. While making a decision, the manager wants to reconsider the purchase of different brands and switch over to the one whose life is more consistent. The past history concerning the burning life of the bulbs of two makes is as follows:

Burning Life of Bulbs

Length of Life (in hours)	Brand A No.of Bulbs	Brand B No. of Bulbs
500 - 700	5	4
700 - 900	11	30
900 - 1100	26	12
1100 - 1300	10	8
1300 - 1500	8	6

Let us calculate the variability in life of the bulbs of two makes.

Calculation of Variability in Life of Bulbs

	Brand A	Brand B
Sample Size N :	60	60
Average Life $\bar{X}$:	1016.67	940.0
Standard Deviation σ :	219.97	220.0

Coefficient of Variation:

Brand A $= \frac{\sigma}{\bar{X}} \times 100$ Brand B $= \frac{\sigma}{\bar{X}} \times 100$

$= \frac{219 \cdot 97}{1016 \cdot 67} \times 100$ $= \frac{220}{940} \times 100$

= 21.63% = 23.40%

Among the two brands, Brand A bulbs are found relatively consistent.

5.6 Skewness:

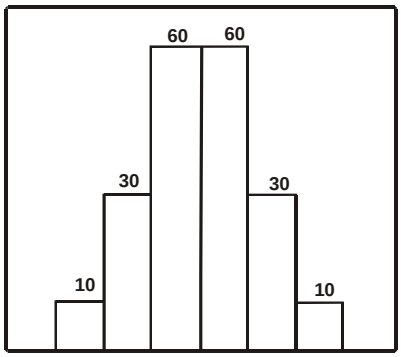
Asymmetry Measure:

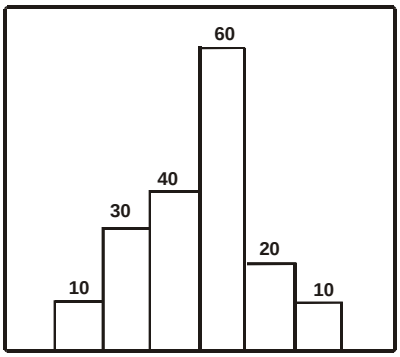
Measures of Central Location help us to estimate the representative value of a series. The Measures of Dispersion help us to indicate the extent to which individual items have scattered away from the Central Value. Skewness is another measure that refers to the amount of symmetry or asymmetry of a distribution. It describes the shape of distribution.

There may be two distributions having the same Mean and Standard Deviation, yet their shapes may be quite different, one may be symmetrical, i.e., larger frequencies are concentrated in the central part of the distribution; and another may not be symmetrical i.e., much of the frequencies might have concentrated on either side. Observe the following figures:

Skewness refers to the asymmetry or lack of symmetry in the shape of frequency distribution. This characteristic is of particular importance in connection with judging typically of certain measures of central tendency.

A. Symmetrical **B. Asymmetrical**



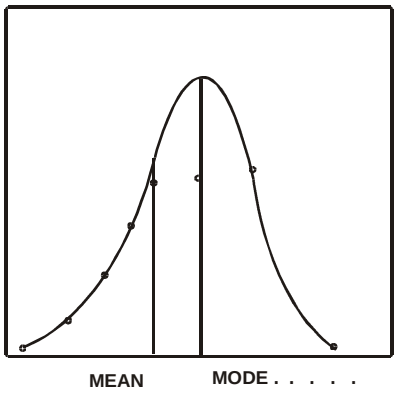


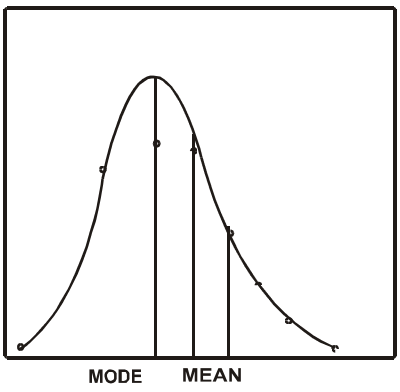
Symmetrical Vs. Asymmetrical:

If in a distribution the values of a variable are distributed at equal distances on either side of the central value, the distribution is said to be symmetrical and its two tails are of equal length.

If the tail is larger than the right tail, distribution is said to be negatively skewed. In such a distribution, the Mode is larger than the Mean. On the other hand, if the Right tail is larger than the left one, i.e., more frequencies are concentrated in early class intervals, the distribution is said to be positively skewed. Here the mode will be smaller than the Mean. The shapes of these distributions can be seen in Figures.

A. Negatively Skewed **B.Positively Skewed**





How to measure it?

In order to make comparisons between the skewness in two or more distributions, a coefficient of Skewness is generally calculated based on the Mean and Mode Values or based on Quartiles:

Popular Measures:

1. Karl Pearson's Coefficient of Skewness
2. Bowley's Coefficient of Skewness

Based on Mean and Standard Deviation:

Karl Pearson's measure considers the distance between Mode and Median for measuring the Skewnes in a given distribution. If the distribution is symmetrical both Mean and Mode should fall at the same value. In any asymmetric distribution generally they certainly do differ.

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

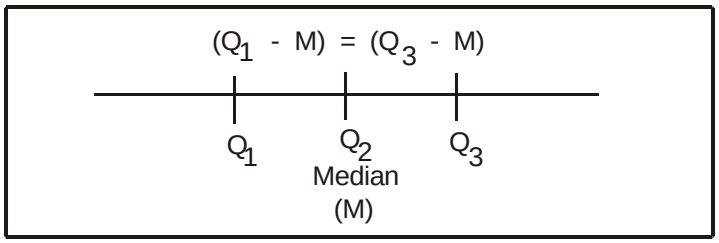
If the Mode can not be determined exactly in a distribution then using the approximate relationship, mode = 3 median - 2 Mean. The above formula can be converted as follows:

$$\text{Skewness} = \frac{3(\text{Mean} - \text{Mode})}{\text{Standard Deviation}}$$

If the value of the coefficient is zero, the distribution is symmetrical. If the value of the coefficient is positive, it is a positively skewed distribution or if the value of the coefficient is negative, it is said to be a negatively skewed distribution.

Based on Quartiles:

Bowley has propounded another measure of Skewness based on the relative position of the Median and Quartile 1 and Quartile 3. If the distribution is symmetrical, then the distance between Median and Quartile 1 should be the same to the Median and Quartile 3.



$$(Q_3 - M) = (M - Q_1)$$

$$(Q_3 - M) - (M - Q_1) = 0$$

$$Q_3 - Q_1 - 2M = 0$$

The coefficient of Skewness can be found by the following formula:

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

To understand these concepts let us consider the following example.

Example 8:

A Trade Union Leader wants to know the wage differentials in two similar manufacturing firms operating in the same vicinity. He has collected the following details of weekly wage distribution.

Weekly Wages in Two Firms

	Firm A	Firm B
Mean	80	85
Median	55	55
Mode	50	50
Quartiles	45	40
	70	75
S.D.	15	20

Let us calculate the variability in wages.

Coefficient of variation is an indicator of variability.

Firm A: $CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$ Firm B: $CV = \frac{20}{85} \times 100$

$= \frac{15}{80} \times 100 = 18 \cdot 75$ $= 23 \cdot 53$

There exists grater variation in payment of wages in Firm B.

Karl Pearson's Coefficient of Skewness:

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

Firm A: $Sk = \frac{80 - 50}{15} = 2 \cdot 00$ Firm B: $SK = \frac{80 - 50}{20} = 1 \cdot 75$

Bowley's Coefficeint of Skewness:

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

Firm A : $SK = \frac{70 + 45 - 2(55)}{70 - 45} = 0 \cdot 20$

Firm B: $SK = \frac{70 + 40 - 2(55)}{70 - 40} = 0 \cdot 14$

Since positive Skewness exists in both the firms, large number of workers are receiving lower wages compared to the average. Further the Skewness is larger in Firm A indicating that the wage payments are positively skewed and that the firm is relatively paying less to a large majority of its workers.

Example 9:

The labour commissioner of an industrial area is interested in enquiring into the gains and losses that have accrued to workers subsequent to a labour dispute. The details are as follows:

Benefits of a Dispute Settled

	Before the dispute	After the dispute
No.of Workers	1,000	950
Average Wages (Rs.)	1,100	1,200
Standard Deviation (Rs.)	250	200
Median Wages (Rs.)	1,200	1,000

Let us calculate the effect of increased average wages on the Wage Bill of the organisation.

Before: Total wage bill $= \sum X = N\,\bar{X}$

$$= 1000 \times 1100 = 11,00,000$$

After: Total wage bill $= 950 \times 1200 = 11,40,000$

The wage bill increases by Rs. 40,000 though the number of workers have gone down from 1000 to 950.

Variability in wage receipts:

Coefficient of variation $= \frac{\sigma}{\bar{X}} \times 100$

Before: $CV = \frac{250}{1100} \times 100 = 22 \cdot 73\%$

After: $CV = \frac{200}{1200} \times 100 = 16 \cdot 67\%$

The variability in individual wages has gone down after the dispute.

Skewness:

Change in the Mode

Mode = 3 Median - 2 Mean

Before: Mode = 3 (1200) - 2(1100)

= 1400

After: Mode = 3 (1000) - 2 (1200)

= 600

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

Before: $SK = \frac{1100 - 1400}{250} = -1 \cdot 20$

After: $SK = \frac{1200 - 600}{200} = +3 \cdot 00$

The negative skewness has become positive after the dispute. It indicates that the relative concentration of wages after the dispute is in the lower class intervals. In other words, while many are getting lower wages compared to average, a very few are able to gain higher wages after the dispute.

5.7 Summary:

In this lesson, we have shown how the concepts of measures of variation and skewness are important. Measures of variation considered were the range, average deviation, quartile deviationand standard deviation. The skewnes was used in relation to lack of symmetry.

5.8 Exercises:

(1) Discuss the importance of Measuring Variability for managerial decision making.

(2) What are different Measures of Dispersion? How would you calculate them from a given frequency distribution? Briefly discuss the relative merits of different measures of dispersion.

(3) Why is Standard Deviation regarded as superior to other measures of dispersion? Explain with examples.

(4) What is coefficient of variation? What are the relative advantages of a relative measure to that of an absolute measure?

(5) How does Skewness differ from Dispersion? What is the objective of measuring Skewness?

(6) The following are the Rents being Collected for Apartments in a large Metropolitan city in India. Compute the range of these data and comment on its suitability.

Rs.	1,500	1,700	2,500	1,800	1,600
	3,500	1,000	1,500	1,200	1,400

(7) A Statistical Consultant of Larsen and Tourbo Switch Gear Division observed that the monthly rupee value of total production of thedivision is fluctuating seriously. Since too many fluctuations may need the recommendation of changes in process outlay, he wishes to analyse the said aspect from the following data:

(RS. '000)	1,224	1,040	6,330	5,020
	4,300	1,620	2,220	1,000
	960	1,200	3,000	3,500
	1,250	4,000	2,000	1,800
	3,000	2,500	4,000	3,000

Calculate inter - quartile range and mean deviation.

(8) Citibank of Madras in observing growth in customer's arrivals and over crowding at certain counters. In order to open a new counter, the management is interested in analysing the average variation in customers visit to bank. The number of customers visitng during recent six months are as follows:

Range of Customers	No.of Days
0 - 50	20
50 - 100	60
100 - 150	50
150 - 200	30
above 200	20
	180

Calculate the variance and standard deviation.

5.9 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.

2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics

   Charles E. Merill Publishing Company.

3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi.,
   S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 6

INDEX NUMBERS

Objective:

After going through this lesson you will learn

- The concept of Index Number
- To compute the Index Number, Chain Base Index Number and Cost for living Index Number.

Structure:

6.1 Introduction:

Statistical methods and theories could be useful in

1. Identifying the pattern of data
2. Describing relationships between variables and
3. Providing a cenceptual foundation to analyses and cope with the change

In this unit, you will be more interested in the measurement and prediction of change. You will follow a forecasting approach which consists of -

1. Look at procedures which have been developed to measure relative changes in economic conditions over time.
2. Analyzing past empericel data to detect reasonably dependable patterns, and
3. Projecting these patterns into future to arrive at future expertations.

Index Numbers are generally used to measure changes occuring over some period of time. They are the number which express the value of a variable at any 'given data called the given period' as a percentage of the value of that variable at some standard date called the 'base period'. When compared with the raw data Index Number many have the following advantages:

(a) They many simplify data and aid in communication.

(b) They many facilitate coparism.

(c) They may be used to show typical seasonal variations.

6.2 Some Importatnt Index Numbers:

Index Numbers may be classified into following three categories.

1. Price Indexes
2. Quantity Indexes
3. Value Indexes

They could be calculated by

1. Simple Aggregate Method and
2. Weighted Aggregate Method

Several notations are necessary before describing these formula.

p_{ij} - Price of the j^{th} commodity in the i^{th} year.

q_{ij} - Quantity of the j^{th} commodity in the i^{th} year.

V_{ij} - Value of the j^{th} commodity in the i^{th} year.

Where $V_{ij} = p_{ij} \times q_{ij}$ commodities (j) varies from 1 to n and period (i) varies from 1 to k years with base year = 0 as Σp_{oj} and Σq_{oj} refer to base year price and quantity respectively.

1. Simple Aggregate Method:

This method expresses aggregate of prices in any year as a percentage of their aggregate in the base year. They price (or quantity) index for the i^{th} year 9i = 1, 2, ...,k) as compared to the base year (i = 0) is

$$P_{oi} = \frac{\Sigma p_{ij}}{\Sigma p_{oj}} \times 100$$

$$Q_{oi} = \frac{\Sigma q_{ij}}{\Sigma q_{oj}} \times 100$$

2. Weighted Aggregate Method:

This method applies different weights to different commodities. Usually weights are proportional to the qunatity consumed in some predefined year.

$$P_{oi} = \frac{\sum W_j P_{ij}}{\sum W_j P_{oj}} \times 100$$

where W_j = weight associated with the j^{th} commodity.

a. Base year method (Lasperpre's Price Index)

If $W_j = P_{oj}$ = base year quantity

we get

$$P_{oi} = \frac{\sum p_{ij} q_{oj}}{\sum p_{oj} q_{oj}} \times 100$$

b. Given year method (Pasche's Price Index)

If $W_j = q_{ij}$ = base year qunatity.

we get $p_{oi} = \frac{\sum p_{ij} q_{ij}}{\sum p_{oj} q_{ij}} \times 100$

c. Use a given year method (Marshal - Edgeworth price index)

If $W_j = \left(q_{oj} + q_{ij}\right)/2$

= Arithmetic means of the base year and the current year quantities.

We get

$$p_{oi} = \frac{\sum p_{ij} \left(q_{oj} + q_{ij}\right)}{\sum p_{oj} \left(q_{oj} + q_{ij}\right)} \times 100$$

d. Fishers Indeal Index. If we take geometric mean of Laspeyre's Price Index and Pasche's Price Index,

We get

$$P_{oi} = \left[\frac{\sum p_{ij} q_{oj}}{\sum p_{oj} q_{oj}} \times \frac{\sum p_i}{\sum p_o} \cdot \frac{q_{ij}}{q_{ij}}\right]^{1/2} \times 100$$

We will discuss in later seetions why it is called "ideal" price index.

3. Qunatity index numbers study the changes in the volume of goods concumed. They can by obtained by interchanging $\left(q_{ij}\right)$ quantity with $\left(p_{ij}\right)$ price.

4. Value Index Numbers are givenby the aggregate expenditure for any given year expressed as a percentage of the same in the base year.

$$V_{oi} = \frac{\sum p_{ij}q_{ij}}{\sum p_{oj}q_{oj}} \times 100$$

6.3 Chain Base Index Number:

Till now we have assumed base period as fixed at some previous period. If however, we are interested in making the data for the two periods homogenous, which could be best done by talking two adjacent period. This could be done by using Chain Base Method. WE calculate a series of index numbers for each year with the proceeding year as base viz., $p_{01}, p_{02}, p_{03} \cdots\cdots\cdots, p_{k-1}, p_k$ and so on where p_{xy} is the price index number for x as base year and y as given year and the index numbers is obtained by the successive multiplications of the index number so obtained to give,

P_{01} = given link

$p_{02} = p_{01} \times p_{12}$

$p_{03} = p_{01} \times p_{12} \times p_{23} = p_{02} \times p_{23}$

:

:

:

$p_{ok} = p_{0(k-1)} \times p_{(k-1)k}$

A part of from homogenuity the chain base method includes newer items and delete older ones in order to make the index more representative with-out effecting comparability and without doing recalculation.

Algorithm for construction

Step 1: For each period find link relatives link relative for

$$i^{th} \text{ period} = \frac{\text{figure of } i^{th} \text{ period}}{\text{figure of preceeding period}} \times 100$$

Step 2: Chain link relatives together by successive multiplication to get chain indices.

$$\text{Chain Index} = \frac{\text{Current Years links RElatives} \times \text{Preceding year chain index}}{100}$$

Example 1:

The following table gives the averages whole some process of four fruits for the year 1990 to 1994. Compute chain base index number.

Fruit	1990	1991	1992	1993	1994
Banana	4	6	8	4	14
Oranges	6	12	18	8	6
Mango	4	12	20	8	16
Apple	5	7	18	1	22

Solution: Step 1:

Link Relatives Based on Proceeding Years					
Fruit	**1990**	**1991**	**1992**	**1993**	**1994**
Banana	100	150	133	50	350
Oranges	100	200	150	44	75
Mango	100	300	167	40	200
Apple	100	140	257	61	200
Total of link Relatives	400	790	707	195	825
Averages	100	197.5	176.8	48.8	206.3

Step 2:

Chain Base Index	100	$\frac{197\cdot5\times100}{100}$ = 197.5	$\frac{176\cdot8\times197\cdot5}{100}$ = 349.2	$\frac{48\cdot8\times349\cdot2}{100}$ = 170.4	$\frac{206\cdot3\times170\cdot4}{100}$ = 351.5

6.4 Cost For Living Index:

Cost of living index numbers indicate whether the reel wages are rising of falling, money wages remaining unchanged cost of living index number is the reciprocal of purchasing power of money. We havethe following methods of to construct the cost of living Index Numbers.

(i) Weighted Aggregates Method:

In this method "quantity" consumed in the base year is taken as weights assigned to various commodities. Thus,

$$\text{Cost of Living Index} = \frac{\text{Total expenditure in current year}}{\text{Total expenditure in base year}} \times 100$$

$$= \frac{\sum p_{ij} q_{oj}}{\sum p_{oj} q_{oj}} \times 100$$

(ii) Method of weighted relatives:

Weighted average of price relatives is taken as cost of living index. The weights being the values of quantities consumed in the base year. In usual notations,

$$\text{Price Relative } p_j = \frac{p_{ij}}{p_{oj}} \times 100$$

Relative and weights $w_j = p_{oj}, q_{oj}$

$$\text{then, cost of living index} = \frac{\sum_i w_j p_j}{\sum_j w_j}$$

However, if we put values of p_j and w_j in the formula we will get the formula in method (1) provided $p_{ij} = p_{oj}$ where p_t = typical price during the time periods considered. Present formula, therefore is only a different way of construction.

Example 2: A university professor is intrested in measuring how the prices of certain commodities has changed during his past 5 years of teaching statistics.

Commodity	Unit of purchase	Typical annual Consumption	price	
			1990	1995
Pen	12 piece	4	60	75
Board Pen	6 piece	1/2	72	92
Registers	1 box	1	950	1100
Aspirin	1 pack	6	59	75
			1141	1342

Calculate Index Numbers for 1995 by taking 1990 as base year by

1. Simple Aggregate Method
2. Weighted Aggregate Method
3. Weighted Average of Relatives Indexes or Method of Weighted Relatives

Simple Aggregate Method:

Sol **(i)** $I_{1990} = \frac{\sum Pij}{Poj} \times 100 = \frac{1141}{1141} \times 100 = 100$

$$I_{1995} = \frac{1342}{1141} \times 100 = 117 \cdot 6$$

or Price Index for Professor's Commodity.

Year	Price Index (1990 = 100)
1990	100.00
1995	117.00

(ii) Weighted Aggregate Method

Cost of Living Index, $I_{1995} = \frac{\sum Pij\ qt}{\sum Poj\ qt} \times 100$

where qt = Typical Annual Consumption.

Commodity	Price		qt	$P_{1990} \times qt$	$P_{1995} \times qt$
	1990	1995			
Pen	60	75	4	240	300
Board Pen	72	92	1/2	36	46
Registers	950	1100	1	950	1100
Aspirin	59	75	6	354	450
			Total	1580	1896

So,

$$I_{1995} = \frac{1896}{1580} \times 100 = 120$$

(iii) Method of Weighted Relatives

We make the following table to do the computations.

Commodity	Price		Price Relatives		q_1	Weights	Weighted Price	
	1990		$\frac{P_{1990}}{P_{1990}}\times 100$	$\frac{P_{1995}}{P_{1990}}\times 100$		$Poj \times qt$	1990	1995
						$A \times E$	$C \times F$	$D \times F$
	A	B	C	D	E	F	G	H
Pen	60	75	100	125	4	240	24000	30000
Board Pen	72	92	100	127.8	0.5	36	3600	4600
Registers	950	1100	100	115.8	1	950	95000	110010
Asprirn	59	75	100	127.1	6	354	35400	44990
					Total	1580	15800	189600

Therefore, Cost of Living Index

$$I_{1990} = \frac{158000}{15800} = 100$$

$$I_{1995} = \frac{189600}{15800} = 120$$

Remark: It may be roted method (2) and (3) gives the some result then why one should go for (3) as it is a lengthy method. The answer of course is that the two methods are not identical $Pt \neq Poj$. As it is often desirable to use prices other than those in the base period to represent what is typical during the time periods studied.

6.5 Criteria for a Good Index Numbers:

There are basically 3 errors in the contruction of Index Numbers.

(a) Formula Error: May be caused due to usage of different formulae none of which measure change precisely.

(b) Sampling Error: May be caused due to sampling of commodities to be included in Index.

(c) Homogeneity Error: May be caused due to change in composition of commodities in the two periods.

Some mathematical tests are designed for error(a)

1. Unit Test: Index number should be independent of the units in which the prices and quantities of various commodities are quoted.

2. Fisher's Time Reversal Test: Index Number should maintain the time consistency by working both forward and backward with respect to time, or

$$P_{ij'} = \frac{1}{P_{i'i}} \left(i = i' = 0, 1, 2, \cdot \cdot \cdot, K\right)$$

3. Fisher's Factor Reversal Test: Index Number should maintain consistency if prices and quantities are interchanged. The multiplication of two results should give value ratio except for a constant symbolically,

$$P_{oi} \times Q_{oi} = \frac{\sum v_{ij}}{\sum v_{oj}} = \frac{\sum P_{ij}\, q_{ij}}{P_{oj} q_{oj}}$$

4. Circular Test: Index Number satisfy the following relation if base is shifted.

$$P_{ab} \times P_{bc} \times P_{ca} = 1,\ a \neq b \neq c = 0, 1, 2, \cdot \cdot \cdot, K$$

Example 3: Show that Fisher's Ideal Index Number satisfies Time and Factor reversal test.

Solution: Fisher's Index Number is

$$P_{oi} = \left[\frac{\sum P_{ij}\, q_{oj}}{\sum P_{oj}\, q_{oj}} \times \frac{\sum P_{ij}\, q_{ij}}{\sum P_{oj}\, q_{ij}}\right]^{1/2} \times 100$$

and $$P_{io} = \left[\frac{\sum P_{oj}\, q_{ij}}{\sum P_{ij}\, q_{ij}} \times \frac{\sum P_{oj}\, q_{oj}}{\sum P_{ij} q_{oj}}\right]^{1/2} \times 100$$

clearly, $P_{oi} \times P_{io} = 1$, so it sulisfies time reversal test.

Further, $$P_{oi} \times Q_{oi} = \left(\frac{\sum P_{ij}\, q_{oj}}{\sum P_{oj}\, q_{oj}} \times \frac{\sum p_{ij}\, q_{ij}}{\sum P_{oj}\, q_{ij}}\right)^{1/2} \times \left(\frac{\sum p_{oj}\, q_{ij}}{\sum q_{oj}\, P_{oj}} \times \frac{\sum q_{ij}\, P_{ij}}{\sum P_{ij}\, q_{oj}}\right)^{1/2}$$

$$= \frac{\sum P_{ij}\, q_{ij}}{\sum P_{oj}\, q_{oj}} = v_{oi}$$

so it satisfies Fotor Reversal Test also.

4. Compute the Laspeyre's index number, Paasche's index number, Fisher's index number for the following data:

Commadity	Quantity		Prices	
	1978 (Q_0) - 1980 (Q_1)		1978 (P_0) - 1980 (P_1)	
A	40	42	1.72	1.92
B	30	35	1.8	1.85
C	10	17	1.90	2.10
D	20	14	1.33	1.66
E	16	12	2.12	3.15
F	15	10	5	6

Commadity	Quantity		Prices					
	1978 - 1980		P_0	P_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	40	42	1.72	1.92	68.8	72.24	80.64	76.8
B	30	35	1.8	1.85	54	63	64.95	55.5
C	10	17	1.9	2.1	19	32.3	35.7	21
D	20	14	1.33	1.66	26.6	18.62	23.24	33.2
E	16	12	2.12	3.15	33.92	25.44	37.8	50.4
F	15	10	5	6	75	50	60	90
					276.52	261.89	302.13	326.9

Laspeyres index number $P_{oi}^{LA} = \frac{\sum p_{ij}\, q_{oj}}{\sum p_{oj} q_{oj}} \times 100$

$= \frac{326 \cdot 9}{276 \cdot 52} \times 100$

$= 118 \cdot 219$

Paasche's index number $p_{oi}^{PA} = \frac{\sum p_{ij} q_{ij}}{\sum p_{oj} q_{ij}} \times 100$

$$= \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{302 \cdot 13}{261 \cdot 59} \times 100$$

$$= 115.4975$$

Fisher's index number $P_{oi}^{F} = \sqrt{\frac{\sum p_{ij} q_{oj}}{\sum p_{oj} q_{oj}} \times \frac{\sum p_{ij} q_{ij}}{\sum p_{oj} q_{ij}}} \times 100$

$$= 116.85047$$

$$Q_{oi}^{F} = \sqrt{p_{oi}^{FA} \cdot p_{oi}^{PA}}$$

Laspeyres's index number $Q_{oi}^{LA} = \frac{\sum q_{ij} p_{oj}}{\sum q_{oj} p_{oj}} \times 100$

$$= \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

$$= \frac{261 \cdot 59}{276 \cdot 52} \times 100$$

$$= 94.6007$$

Paasche's Quantity Index numbers $= \frac{\sum q_{ij} p_{ij}}{\sum q_{oj} p_{ij}} \times 100 = 92 \cdot 422$

Fishers Quantity Index Numbers $Q_{oi}^{F} = \sqrt{q_{oi}^{LA} \cdot q_{oi}^{PA}}$

$$= 93 \cdot 505$$

Fishers Quantity Index Number = 93.505

5. From the given data calculate Fisher's Index Number, Laspeyreb, Pascher's Browley's Index Numbers from the given data.

Item	1995		1996	
	p_0	q_0	p_1	q_1
A	4	6	2	8
B	6	5	5	10
C	5	10	4	14
D	2	13	2	19

Item	1995		1996					
	p_0	q_0	p_1	q_1	p_0q_0	p_1q_0	p_0q_1	p_1q_1
A	4	6	2	8	24	12	16	32
B	6	5	5	10	30	25	50	60
C	5	10	4	14	50	50	56	70
D	2	13	2	19	20	26	38	38
					130	103	160	200

Laspeyres price index number $P_{oi}^{LA} = \frac{\sum p_1q_0}{\sum p_0q_0} \times 100$

$= \frac{103}{130} \times 100 = 79 \cdot 2308$

Paaschels Index Number $P_{oi}^{PA} = \frac{\sum p_1q_1}{\sum p_0q_1} \times 100 = \frac{160}{62} \times 100 = 80$

Probish Bowley's Price Index Number $p_{oi}^{PB} = \frac{1}{2}\left[p_{oi}^{PA} + p_{oi}^{LA}\right]$

$= \frac{1}{2}\left[79 \cdot 2328 + 80\right] = 79 \cdot 6154$

Lasseyre's Qunatity Index Number $Q_{oi}^{LA} = Q_{oi}^{La} = \frac{\sum q_1q_0}{\sum q_0p_0} \times 100 = \frac{200}{130} \times 100$

$= 153.8462$

Paasulre's Quantity Index Number

$$Q_{oi}^{PA} = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100 = \frac{160}{103} \times 100$$

$$= 155.3398$$

Fisher's Price Index Number $\left(P_{oi}^{F}\right) = \sqrt{P_{oi}^{ka} \times P_{oi}^{pa}}$

$$= \sqrt{6338 \cdot 464}$$

$$= 79.6144$$

Fisher's Quantity Index Number

$$\left(Q_{oi}^{F}\right) = \sqrt{Q_{oi}^{kQ} \times Q_{oi}^{pa}}$$

$$= \sqrt{23898 \cdot 43794}$$

$$= 154.59119$$

Drdosh Bowley's Quantity Index Number

$$Q_{oi}^{DB} = \frac{1}{2}\left(Q_{oi}^{la} + Q_{oi}^{pa}\right)$$

$$= \frac{1}{2}\left(153 \cdot 8462 + 155 \cdot 8398\right)$$

$$= {}^{309 \cdot 186}\!/_{2} = 154 \cdot 593$$

Example 6: **Simple aggregate method:**

Commadity	Price in 1995 (p_0)	Price in 2000 (p_1)
A	20	25
B	30	30
C	10	35
D	25	35
E	40	45
F	50	55
	175	205

Simple aggregrate method $= \frac{\sum p_1}{\sum p_0} \times 100 = \frac{205}{155} \times 100 = 117 \cdot 1428$

Example 7: Simple average of relatioves method using A.M.

Commiadity	p_0	p_1	**Price relative** $= \frac{p_1}{p_0} \times 100$
A	20	25	125
B	30	30	100
C	10	15	150
D	25	35	140
E	40	45	112.5
F	50	55	110
			737.5

$$\text{A.M.} = \frac{\sum p}{n} = \frac{737 \cdot 5}{6} = 122 \cdot 9167$$

Example 8: Simple average of relatives method using G.M.

Commiadity	p_0	p_1	**Price relative** $= \frac{p_1}{p_0} \times 100$	**log p**
A	20	25	125	2.0169
B	30	30	100	2.0000
C	10	15	150	2.1761
D	25	35	140	2.1461
E	40	45	112.5	2.0511
F	50	55	110	2.0414
				12.5116

$$G \cdot M \cdot = \text{anti}\log\left(\frac{\sum \log p}{n}\right)$$

$$= \text{anti}\log\left(\frac{12 \cdot 5116}{6}\right) = 121 \cdot 7026$$

6.6 Summary:

Index Number is a measure of a charateristic of a group of items. First we compute the relative change in the characteristic of a single item. Then such relative are combined using some form of averaging. The characteristic may be price, quantily, value or cost. The average may be simple A.M, geometric mean or weighted geometric mean of prices or price relatives.

The main steps in the construction of an index number are (i) stating the purpose and objectives (ii) choosing the base period (iii) seleding the items to be included (iv) deciding the appropriate formula and (v) choosing suitable weights.

6.7 Exercise:

1. What is the rationale behind index numbers?
2. Define index numbers.
3. What are the uses of index numbers?
4. What type of index numbers are usually calculated?
5. Briefly discus the unweighted quantity index.
6. Explicate Laspeyre's price and quantity index numbers.
7. Discuss Paasche's price and quantity index numbers.
8. Describe chain base index number.
9. Explain cost for living index.
10. Discuss criteria for a good index numbers.
11. Construct a suitable price index from the following data:

Commodity	Base Year		Current Year
	Price	Quantity	Price
A	64	100	60
B	60	75	50
C	60	75	36

12. Given the following data on oil production (i) calculate Laspeyres index (ii) Passche's index (iii) Fisher's index for prices and quantities.

Item	Base Year		Current Year	
	Price	Quantity	Price	Value
Coconut	35	100	40	140
Ground nut	25	250	30	320
Mustard	20	100	28	120
Sunflower	10	120	20	200

Interpret the results verify whether time and factor reversal tests are satisfied by these indices.

13. Compute Fisher's quantity index number for the following group.

Commodity	Price		Quantity	
	Base Year	Current Year	Base Year	Currnt Year
A	175	200	2	3
B	210	230	3	4
C	475	525	1	2
D	100	120	4	4

14. Given the following data, compute Fisher's price index

Commodity	Base Year		Current Year	
	Price per unit	Volume (in Rs)	Price per unit	Volume (in Rs)
A	2	40	5	50
B	4	15	8	40
C	1	10	2	30
D	5	25	10	60

15. Monthly household expenditure of textile mill workers in 1980 and the percentage increase during 1980 - 90 are given below. Compute over-all percentage of increase in the cost of living.

Group in 1980	Expenditure (Rs.) during 1980 - 90	% increase
Food	1200	90
Fuel & light	120	75
House Rent	140	150
Clothing & Footwear	200	100
Miscellaneous	250	140

16. Calculate the CPI for 1990 with 1980 as base given the following group indices for 1982 and 1990 with 1970 as base year.

Group	Group index 1982	with base 1970 1990	weight
Food	240	400	30
Clothing	280	300	10
Fuel & light	320	700	15
Misc	250	400	13

17. The following table gives the index of earning and CPI with 1952 = 100 for a certain class of workers. Calculate the earnings by deflating therupee.

Year	Earnings	C.P.I.
1945	94	91
1950	89	92
1955	94	103
1960	110	115
1965	132	152
1970	180	200
1975	195	240
1980	250	300

6.8 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 7

CORRELATION

Objective:

After reading this lesson, you would be able to

Measure the stength of linear relationship between two variables.

Structure:

7.1 Introduction:

For example the relationship between the age of husband and age of wife, price of a commodity and the amount demanded, heights and weights of a group of persons, income and expenditure of a group of persons etc., are studied with the help of correlation analysis. The measure of correlation called correlation coefficient expresses the degree and direction of the relationship, thus the term correlation (or covariation) indicates the relationship between two such variables in which with change in the values of one variable, the values of the other variable also changes. Correlation indicates the relationship between two variables so that movements in one variable tend to be accomplished by movement in the second variable.

7.2 Uses of Correlation:

The study of correlation is useful in practical life because of the reasons as follows:

1. With the help of correlation analysis one can measure the degree of relationship that exists between the variables in one figure.
2. From one variable we can estimate the other variable by the help of regression analysis only when we establish the variables are related.
3. Correlation analysis is very helpful in understanding the economic behaviour. It helps us in locating such variables on which the other variables depend. It aids in locating the critically important variables on which others depend; for example we can find out

the factors responsible for price rise or low productivity.

4. Correlation study helps us in identifying such factors which can stabilize a disturbed economic situation.

5. Interrelationship studies between different variables are helpful tools in promoting research.

6. The effect of correlationis to reduce the range of uncertainity in thedecision making effort. In social sciences, particularly in business world, forecasting is an important phe nomenon and correlation studies help us to make relatively more dependable forecast.

Thus correlationstudies are very widely used as the basic tool for analysis and interpretation of statistical data relative to two or more variables. The correlation may be any one or a combination of the following reasons:

1. The correlation may be due to pure chance especially in a small sample. We may get a high degree of correlation between two variables in a sample but in the universe there may not be any relationship between the variables at all.

2. Both the correlated variables may be influenced by one or more other variables. For example when prices of rice and jute are increasing we may find high degree of correlation. In reality they are not related or neither of them are cause or effect. It may be due to their production which is affected by rainfall.

3. Both the variables may be mutually influencing each other so that neither can be designed as the cause and the other effect. For example, thedemand of a commodity may go down as a result of rise in price. It may also be so that demand ofthe commodity has gone up due to anticipated shortage in future and has resulted in price rise.

4. There might be a situation of non sense or sputious correlation between the two vaiables under study. One may find high degree of correlation between number of divorces per year and exports of television sets. Obviously there cannot be any relationship between number of divorces and television export. It should be understood that correlation is a relationship between related variables only.

The above points make it clear that correlation is only a mathematical relationship and it does not necessarly signify a cause and effect relationship between the variables.

7.3 Types of Correlation:

Correlation may be classified into three categories, namely

1. Positive or Negative
2. Simple, Multiple and Partial
3. Linear or Non-Linear

1. Positive or Negative Correlation:

It is otherwise called as direct (positive) or inverse (negative) correlation. The correlation between two variables is said to be direct or positive when they move in the same direction so that an increase of decrease in the value of one variable is associated with increase or

decrease in the value of the other. On the other hand if the variables move in the opposite direction i.e., increase in one variable is associated with decrease in the other and vice versa the correlation is said to be inverse or negative. The following example would example the difference between positive and negative correlation.

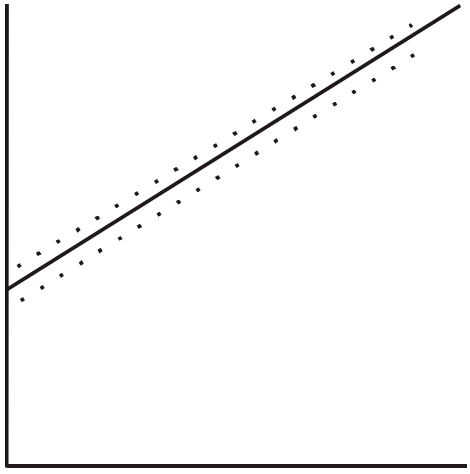
Positive Correlation:

		(1)					(2)		
X	10	15	20	25	X	17	13	7	5
Y	29	35	40	49	Y	34	22	15	7

Negative Correlation:

		(1)					(2)		
X	20	30	40	50	X	60	40	30	20
Y	50	40	35	20	Y	100	120	125	137

Figure 1 Positive Linear Correlation



7.4 Methods of Studying Correlation:

The various methods by which correlation studies are made are as follows:

1. Scatter diagram method
2. Karl Pearson's Coefficient of correlation method.
3. Rank correlation method
4. Concurrent deviation method

These methods are discussed below:

1. Scatter Diagram Method:

The scatter diagram is the simplest method of styding relationship between two

variables. The simplest device for ascertaining whether two variables are related is to prepare a dot chart, harizontal axis representing one variable and vertical axis representing the other. The diagram of dots so obtained is known as scatter diagram. From the scatter diagram we can form a fairly good, though rough idea about the relationship between two variables.

The following diagrams of the scattered data depict different types of correlation.

Fig 2: Perfect Positive Linear Correlation

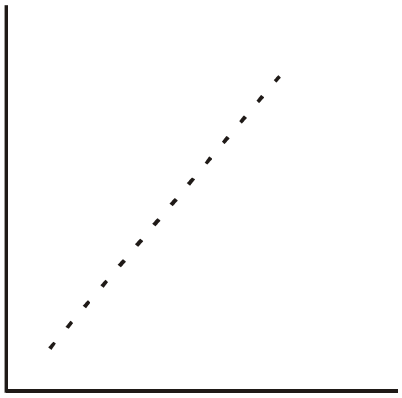
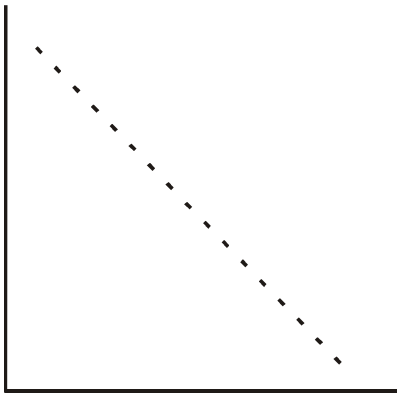


Fig 3: Perfect Negative Linear Correlation



Merits and Demerits of Scatter Diagram:

Merits:

This method is sample, easy and non methematical method of studying correlation between the variables. Even a mere inspection enables us to form a rough idea of the nature of relationship. Moreover, this method is not affected by extreme observations unlike the mathematical formula of ascertaining correlation. Drawing a scatter diagram usually is the first step in investigating the relationship between the variables.

Demerits:

The major limitaion of the methods is that it gives a visual picture of the relationship. It may tell about the nature of relationship but it fails to tell about the extent or exact degree of relationship. This method also is not amenable to any further mathematical treatment.

2. Karl Pearson's Coefficient of Correlation:

It is a mathematical method for measuring the intensity or magnitude of relationship between two series of variables. Of the several mathematical methods of measuring correlation, Karl Pearson's coefficient of correlation is most widely used in practice. The Pearson's coefficient of correlation is denoted by 'r'. The formula for computing 'r' is

$$r = \frac{\sum xy}{N\sigma x\, \sigma y}$$

where $X = (X - X')$

$Y = (Y - Y')$

σx = Standard Deviation of x series

σy = Standard Deviation of y series

N = Number of pair observations

r = Correlation Coefficient

It should be noted that this method is to be applied only where the deviations are taken from the actual mean. The value of coefficeint of correlation as obtained by the above method shall always lie between ± 1. When $r = +1$ it means there is a perfect positive correlation, when r = -1 it means there is a perfect negative correlation and when r = 0 there is no relationship between thevariables. The coefficient of correlationnot only describes the magnitude of correlation but also its direction. The above formula can be transformed into another formula were standard deviations of the two series need not be calculated and which is easire to apply.

Symbolically $$r = \frac{\sum x y}{N \sigma x \sigma y}$$

or $$r = \frac{\sum x y}{N\sqrt{\frac{\sum x^2}{N}}\sqrt{\frac{\sum y^2}{N}}}$$

where $X = (X - X')$

$Y = (Y - Y')$

$$r = \frac{\sum x y}{\sqrt{\sum x^2 \times \sum y^2}}$$

Example 1:

Calculate Karl Pearson's Coefficient of correlation from the following data.

Roll No.of The Students:	1	2	3	4	5	6	7	8
Marks in Statistics:	65	66	67	67	68	69	70	72
Marks in Accountancy :	67	68	65	68	72	72	69	71

Solution:

Let marks in statistics be denoted by x and accountancy by y.

R.N.	X	$X - X'$	X^2	Y	Y - Y	Y^2	XY
1	65	- 3	9	67	- 2	4	6
2	66	- 2	4	68	- 1	1	2
3	67	- 1	1	65	- 4	16	4
4	67	- 1	1	68	- 1	1	1
5	68	0	0	72	+ 3	9	0
6	69	1	1	72	+ 3	9	3
7	70	2	4	69	0	0	0
8	72	4	16	71	+ 2	4	8
	$\sum x = 544$	$\sum x = 0$	$\sum x^2 = 36$	$\sum y = 552$	$\sum y = 0$	$\sum y^2 = 44$	$\sum xy = 24$

$$r = \frac{\sum xy}{N \sigma x \, \sigma y}$$ (1st formula)

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$ (2nd formula)

$$x' = (X - X') \qquad y' = (Y - Y')$$

$$x = \frac{\sum x}{N} = \frac{552}{8} = 68 \qquad y = \frac{\sum y}{N} = \frac{552}{8} = 69$$

Calculation of r by first formula:

$$\sqrt{x} = \frac{\sqrt{\sum x^2}}{N} = \frac{\sqrt{36}}{8} = \sqrt{4 \cdot 5} = 2 \cdot 121$$

$$\sqrt{y} = \frac{\sqrt{\sum y^2}}{N} = \frac{\sqrt{44}}{8} = \sqrt{5 \cdot 5} = 2 \cdot 345$$

$$r = \frac{24}{8 \times 2 \cdot 121 \times 2 \cdot 345} = +0 \cdot 603$$

Calculation of r by second formula:

$$r = \frac{24}{\sqrt{36 \times 44}} = +0 \cdot 603$$

Direct Method of Calculating Correlation:

The coefficient of correlationcan be calculated directly without finding out the deviation of various values from the mean of the series with the help of following formual:

$$r = \frac{\sum xy - \sum x \sum y/N}{\left[\sum X^2 - (\sum X)^2/N\right]\left[\sum Y^2 - (\sum Y)^2/N\right]}$$

or

$$= \frac{N \sum xy - \sum x \sum y}{\sqrt{N\sum x^2 - (\sum x)^2}\ \sqrt{N\sum y^2 - (\sum y)^2/N}}$$

The above problem can be solved as follows:

R.N.	X	X^2	Y	Y^2	XY
1	65	4225	67	4489	4355
2	66	4356	68	4624	4488
3	67	4489	65	4225	4355
4	67	4489	68	4624	4556
5	68	4624	72	5184	4896
6	69	4761	72	5187	4868
7	70	4900	69	4761	4830
8	72	5184	71	5041	5112
	$X = 544$	$\sum X^2 = 37028$	$\sum Y = 552$	$\sum Y^2 = 38132$	$\sum XY = 37560$

$$r = \frac{8 \times 37560 - (544 \times 552)}{\sqrt{8(37028) - (544)^2} \times \sqrt{8\,(3812) - (552)^2}}$$

$$r = \frac{300480 - 300288}{\sqrt{296224 - 295936} \times \sqrt{305056 - 304704}}$$

$$r = \frac{192}{16 \cdot 97 \times 18 \cdot 76}$$

$$r = \frac{192}{318 \cdot 35} = +0 \cdot 603$$

When deviations are taken from assumed mean:

When the figures of bothe the series of variablesare big as well as their actual means are in fraction, the calculation of correlation by the methods discussed above would invole too many calculations and the calculation will be tedious. In such case, to make the calculation easier we can make use of assumed mean and modify the formula sccordingly for finding out correlation.

The modified formula for calculation of correlation coefficeint is as follows:

$$r = \frac{\sum dxdy - \frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}} \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

where $\sum dx$ and $\sum dy$ are the sum of deviations of the x and y series respectively from their assumed means. $\sum dx^2$ and $\sum dy^2$ are the sum of the squares of deviation of x and y series respectively from their assumed means. $\sum dxdy$ is the sum of the product of corresponding deviations of x and y series from their assumed means.

Example 2:

Calculate the coeffient of correlation for the ages of husband and wife.

Age of Husband	24	28	29	30	31	32	34	36	37	40
Age of Wife	19	23	24	25	26	27	29	30	31	33

x	y	dx (x - 32)	dy (y - 26)	dx^2	dy^2	dxdy
24	19	- 8	- 7	64	49	56
28	23	- 4	- 3	16	9	12
29	24	- 3	- 2	9	4	6
30	25	- 2	- 1	4	1	2
31	26	- 1	0	1	0	0
32	27	0	1	0	1	0
34	29	2	3	4	9	6
36	30	4	4	16	16	16
37	31	5	5	25	25	25
40	34	8	7	64	49	56
$\sum x = 321$	$\sum y = 267$	$\sum dx = 1$	$\sum dy = 7$	$\sum dx^2 = 203$	$\sum dy^2 = 163$	$\sum dxdy = 179$

$$r = \sum dxdy - \frac{\frac{(\sum dx)(\sum dy)}{n}}{\sqrt{\sum dx^2 - \frac{(\sum dx)^2}{N}}\ \sqrt{\sum dy^2 - \frac{(\sum dy)^2}{N}}}$$

$$r = \frac{179 - \frac{1 \times 7}{10}}{\sqrt{203 - \frac{(1)^2}{10}}\ \sqrt{163 - \frac{(7)^2}{10}}}$$

$$r = \frac{179 - 0 \cdot 7}{14 \cdot 24 \times 12 \cdot 57}$$

$$= \frac{178 \cdot 3}{178 \cdot 99}$$

$$= + \ 0 \cdot 996$$

Correlation of Grouped Data:

When the number of observations is large the data are often classified into two any frequency distribution, otherwise called as bivariate frequency distribution since it shows the frequency distribution of two related variables. The class intervals of y are listed in the column headings and those of x are listed in rows of the table. The frequencies for each cell of the table are determined by tallying the frequency of the distribution of a single variable.

The formula for calculating the coefficeint of correlation in such case

$$r = \frac{N\sum fdxdy - \sum fdx \sum fdy}{\sqrt{N\sum fdx^2 - \left(\sum fdx\right)^2}\ \sqrt{N\sum fdy^2 - \left(\sum fdy\right)^2}}$$

(or)

$$r = \sum fdxdy - \frac{\frac{\sum fdx \sum fdy}{N}}{\sqrt{\sum fdx^2 - \frac{\left(\sum fdx\right)^2}{N}} \times \sqrt{\sum fdy^2 - \frac{\left(\sum fdy\right)^2}{N}}}$$

Note:

This formula is the same as the one discussed above for assumed mean. The only difference is that here the deviations are also multiplied by the frequencies.

Steps:

(i) Take the step deviations of x and y series from assumed means and denote them by dx and dy.

(ii) Multiply dx and dy and the frequency of the cell. The value $fdxdy$ may be put at the top on right or left hand side of the cell.

(iii) Add all values of $fdxdy$ as calculated in step (ii) and thus obtain the value of $\sum fdxdy$.

(iv) Multiply the frequency of x series with the step deviation and total all such product to get $\sum fdx$. Similarly get $\sum fdy$.

(v) Take the square of the s tep deviations of x series and multiply them with the frequency and total the products to get $\sum fdx^2$, similarly get $\sum fdy^2$.

(vi) Substitute the values so obtained in the formula given above to get the value of r.

The following example would clarify the above points:

Example 3:

From the following table given below calculate the coefficient of correlation between family income and food expenditure of 100 families.

Bivariate Table

Food Expenditure in %	Family Income				
	2000 - 3000	3000 - 4000	4000 - 5000	5000 - 6000	6000 - 7000
10 - 15	-	-	-	03	07
15 - 20	-	04	09	04	03
20 - 25	07	06	12	05	-
25 - 30	03	10	19	08	-

$$r = \sum fdxdy - \frac{\frac{\sum fdx \sum fdy}{N}}{\sqrt{\sum fdx^2 - \frac{(\sum fdx)^2}{N}} \sqrt{\sum fdy^2 - \frac{(\sum fdy)^2}{N}}}$$

$$r = \frac{-48 - 0 \times 100}{\sqrt{120 - \frac{(0)^2}{100}} \sqrt{200 - \frac{(100)^2}{100}}}$$

$$r = \frac{-48}{10 \cdot 95 \times 10} = \frac{-48}{109 \cdot 5}$$

= - 0.4383

Merits and Demerits of Karl Pearson's Coefficient of Correlation:

Karl Pearson's method is the most popular method for measuring the degree of relationshop among all the mathematical methods used for the same. The merit of this coefficient is that it gives the degree of the relationship among the variables as well as the direction of the correlation.

However it suffers from some limitations also. These are:

1. The correlation coefficient always assumes linear relationship even though it may not be there.

2. It is liable to be misinterpreted as a high degree of correlation does not necessarily mean very close relationship. Thus great care should be exercised in interpreting the values.

3. The values of the coefficient is unduly affected by the extreme items.

4. As compared to other methods it is tedious to calculate.

3. Rank Correlation Method:

The Karl Pearson's coefficient of correlation as discussed above can not be used incases where the direct quantitative measurement of the phenomenon under study is not possible, for example efficiency, honestly intelligence etc. In such cases it may be possible to arrange various items of a series in serial order but the quantitative measurement of their value in difficult. There are many such attributes which are incapable of quantitative measurement. If it is desired to have a study of association between two such attributes say intelligence and beauty, the Karl Pearson's coefficient of correlation cannot be calculated as these attributes cannot be assigneddefinite values. In such cases one can rank or array the different items and apply Spearman's rank correlation method for finding out the degree of correlation. The formula for computing Spearman's rank correlation is

$$r_s = 1 - \frac{6\sum d^2}{N\left(N^2 - 1\right)} \quad \text{or} \quad r_2 = 1 - \frac{6\sum d^2}{N^3 - N}$$

where r_s denotes the Spearman's rank correlation.

d denotes the difference of the ranks of the paired attributes of a single item.

N stands for the number of pairs.

The values of this coefficient, interpreted in the same way as Karl Pearson's coefficient of correlationranges between + 1 to - 1 when r_s is + 1 it indicates complete agreement in the order of ranks between the two attributes. When r_s is - 1 it indicates complete disagreement in the order of the ranks as they are in opposite direction. This shall be clear from the following example:

R_1	R_2	d $\left(R_1 - R_2\right)$	d^2
1	1	0	0
2	2	0	0
3	3	0	0
			$\sum d^2 = 0$

$$r_s = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$r_s = 1 - \frac{6\times 0}{3^3 - 3} = 1 - 0 = 1$$

R_1	R_2	d $(R_1 - R_2)$	d^2
1	3	- 2	4
2	2	0	0
3	1	2	4
			$\sum d^2 = 8$

$$r = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$r = 1 - \frac{6 \times 8}{3^3 - 3} = 1 - 2 = 1$$

There are two types of problems of calculating the coefficient:

(a) When actual ranks are given

(b) When actual ranks are not given

In cach of these two types of problem a difficulty arises when ranks of two individuals are the same. Such problems need a modification in the formula given above.

(a) When actual ranks are given:

In this situation the following steps are involved.

(i) Compute the difference of ranks $(R_1 - R_2)$ and denote them by d

(ii) Compute d and total them to get $\sum d^2$

(iii) Use the formula given below to get the coefficient

$$r_s = 1 - \frac{6\sum d^2}{N^3 - N}$$

The following example will illustrate the above method.

Example 4:

The ranking of ten students in Statistics and Economics are as follows:

Statistics:	3	5	8	4	7	10	2	1	6	9
Economics:	6	4	9	8	1	2	3	10	5	7

Use Spearman's formula to find out the rank correlation coefficient.

Solution:

R_1 Statistics	R_2 Economics	d Rank difference	d^2 Square of rank difference
3	6	- 3	9
5	4	+ 1	1
8	9	- 1	1
4	8	- 4	16
7	1	+ 6	36
10	2	+ 8	64
2	3	- 1	1
1	10	- 9	81
6	5	+ 1	1
9	7	+ 2	4
			$\sum d^2 = 214$

$$r_s = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$r_s = 1 - \frac{6(214)}{10^3 - 10}$$

$$r_s = 1 - \frac{1284}{10(99)} = 1 - 1 \cdot 3 = -0 \cdot 3$$

Example 5:

Ten competitors in a beauty contest were ranked by three judges in the following order.

First judge	1	6	5	9	2	3	4	10	7	8
Second Judge	3	5	8	7	4	9	2	1	6	10
Third Judge	6	4	8	9	1	3	2	10	7	5

Use the method of rank correlation to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution:

With a view to find out which pair of judges have the nearest approach to common taste in beauty, we will compare the rank correlation between the judgements of

(i) 1st and 2nd judge

(ii) 1st and 3rd judge

(iii) 2nd and 3rd judge

The ranks given by the three judges would be denoted by R_1, R_2 and R_3

R_1	R_2	R_3	$R_1 - R_2$ $d_{1,2}$	$R_1 - R_3$ $d_{1,3}$	$R_2 - R_3$ $d_{2,3}$	d_{12}^2	d_{13}^2	d_{23}^2
1	3	- 6	- 2	- 5	- 3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	8	- 3	- 3	0	9	9	0
9	7	9	2	0	- 2	4	0	4
2	4	1	- 2	1	3	4	1	9
3	9	3	- 6	0	6	36	0	36
4	2	2	2	2	0	4	4	0
10	1	10	9	0	- 9	81	0	81
7	6	7	1	0	- 1	1	0	1
8	10	5	- 2	3	5	4	9	25
						148	54	166

We have n = 10

Spearman's rank correlation coefficient are given by

$$r_{1,2} = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 148}{10^3 - 10}$$

$$= 1 - \frac{888}{990} = 1 - 0 \cdot 896 = 0 \cdot 106$$

$$r_{1,3} = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 54}{10^3 - 10}$$

$$= 1 - \frac{324}{990} = 1 - 0 \cdot 327 = 0 \cdot 673$$

$$r_{2,3} = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 166}{10^3 - 10}$$

$$= 1 - \frac{996}{990} = 1 - 1 \cdot 006 = -0 \cdot 006$$

Since $r_{1,3}$ is the highest the pair of 1st and 3rd judges has the nearest approach to common taste in beauty.

(b) When ranks are not given:

Rank formula can be used even if we are dealing with variables which can be measured quantitatively. If we are given the actual data (not the ranks) we have to convert the data into ranks. The highest (or smallest) value is given rank 1 i.e., either ranking is to be done by discending or ascending order. Whatever may be the order of ranking it has to be uniformly followed in case of bothe the variables.

Example 6:

Calculate Spearnman's rank correlation coefficient between advertisment cost and sales from the following data:

Advertisement Cost (in '000 Rs)	39	68	65	92	85	76	26	98	36	79
Sales (in lakhs)	47	59	59	86	63	68	60	91	51	84

Solution:

Let x denote the advertisement cost ('000 Rs) and y denote the sales (lakhs).

x	y	R_x	R_y	d	d^2
39	47	8	10	- 2	4
68	54	6	8	- 2	4
65	59	7	7	0	0
92	86	2	2	0	0
85	63	3	5	- 2	4
76	68	5	4	1	1
27	60	10	6	4	16
98	91	1	1	0	0
36	51	9	9	0	0
79	84	4	3	1	1
				$\sum d = 0$	$\sum d^2 = 30$

Here n = 10

Therefore

$$r_s = 1 - \frac{6\sum d^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 30}{10^3 - 10}$$

$$= 1 - \frac{180}{990} = 1 - 0 \cdot 181 = 0 \cdot 819$$

Equal Ranks:

In some cases it may be found necessary to rank two or more individuals or entries as equal. In such cases common ranks are assigned to therepeated items. This common ranks are assigned to the repeated items. This common ranks are the arithmetic mean of ranks which these items would have got if they were different from each other and the next will get the rank next to rank used in computing the common rank. Thus if two individuals are ranked equal at fifth place, they are each given the rank $5 + 6/2$, that is 5.5, if 4 are ranked equal at 6th place they are given the rank $6 + 7 + 8 + 9/4 = 30/4 = 7 \cdot 5$ and the next rank would be 10th where equal ranks are assigned to some entries, an adjustment in the above formula is made calculating rank correlation coefficeint.

The adjustment consists of adding $1/12\left(m^3 - m\right)$ to the value of $\sum d^2$ where m stands for the number of items whose ranks are common. If there are more than one such group of items in common, this value is added as many times as the number of such groups. The formula can thus be written as

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \cdots\cdots\right\}}{N^3 - N}$$

Example 7:

A psychologist wanted to compare two methods A and B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair here approximately equal scores on an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below:

Pair	1	2	3	4	5	6	7	8	9	10	11
A	24	29	19	14	30	19	27	30	20	28	11
B	37	35	16	26	23	27	19	20	16	11	21

Find the rank correlation coefficient.

Solution:

A	**B**	**Rank of A**	**Rank of B**	**d**	d^2
24	37	6	1	5	25
29	35	3	2	1	1
19	16	8.5	9.5	- 1	1
14	26	10	4	6	36
30	23	1.5	5	- 3.5	12.25
19	27	8.5	3	5.5	30.25
27	19	5	8	- 3	9.00
30	20	1.5	7	- 5.5	30.25
20	16	7	9.5	- 2.5	6.25
28	11	4	11	- 7	49.00
11	21	11	6	5	25.00
				$\sum d = 0$	$\sum d^2 = 225$

In the A series the value 30 occurs twice. The common rank assigned to each of those values is 1.5 the arithmetic mean of 1 and 2, the ranks which these observations would have taken if they were different. The next value 29 gets the next rank i.e., 3. Again the value 19 occurs twice. The common rank assigned to it is 8.5. The arithmetic mean of 8 and 9 and the next value 14 gets the rank 10. Similarly in B series the value 16 occurs twice and the common rank assigned to each is 9.5, the arithmetic mean of 9 and 10. The next value 11 gets the rank 11.

Hence we see that in the A series item 19, 30 are repeated each occurs twice and in B series the item 16 is repeated. Thus in each of the three cases m = 2. Hence on applying the correction factor $m^3 - m/12$ for each repeated items we get

$$r_s = 1 - \frac{6\left[\sum d^2 + \frac{2^3-2}{12} + \frac{2^3-2^3-2}{12} + \frac{2^3-2}{12}\right]}{11^3 - 11}$$

$$= 1 - \frac{6 \times 226 \cdot 5}{11 \times 120} = 1 - 1 \cdot 0225 = -0 \cdot 0225$$

Merits and Demerits of the Rank Method:

Merits:

1. This method is very simple to calculate and to understand.
2. Where the data are of a qualitative nature like beauty, honestly, intelligence etc. this method can be employed usefully.
3. When the ranks of different item values in the variables only are given this is the only method for finding out degree of correlation.
4. If it is desired to use this formula when actual values are given ranks can be ascertained and correlation can be found out.
5. Since in this method $\sum d$ or sum of the differences provides a check on calculation.
6. It can be interpreted in the same way like Karl Pearson's coefficient.

Demerits:

1. This method cannot be used for finding out correlation in a grouped frequency distribution.
2. It can be conveniently used only when n is small say 30 or less. If it exceeds 30 the calculations becomes quite tedious and require a lot of time.
3. As all the information concerning the variables is not utilised this method lacks percision a compared to Pearson's Method.

4. Concurrent Deviation Method:

This method is one of the wys of ascertaining the coefficeint of correlation by an extremely simple calculation. It is based on the direction of change or variation in the two parired variables. In this method correlation is calculated between the direction of deviation and not their magnitude. In majority of Pearson's coefficient with much less calculations.

To calculate the coefficient of concurrent deviations the deviationsare not calculated from an average or assumed or moving average but only their direction from the preceding item and only the direction of the deviation (i.e., positive or negative) and not the extent of deviation are considered. The formula for calculation of coefficient of concurrent deviation is given below:

Coefficient of concurrent deviation or

$$r_c = \pm \sqrt{\pm (2c - n)/2}$$

Where r_c stands for coefficeint orconcurrent deviation.

c stands for the number of pairs of concurrent deviation and n for number of pair observations. The value of this coefficient of correlation also varies between + 1 to - 1. The plus and minus signs given in the formula should be carefully noted. If the value of $(2c - n)/n$ is negative its squae cannot be calculated and so a minus sign is placed before the sign of the root so that the square root may be calculated and the minus sign may be kept before the value of the coefficient of correlation.

The steps in the calculation of this coefficient are:

1. Examine the fluctuation of each series and find whether each item increases or diminishes in comparison with the item just preceding. All increses are noted as plus and all decreases as minus. If there is neither increase or decrease the direction is zero. The direction fo change on both series is denoted by dx and dy respectively.
2. Multiply dx and dy and determine the value of C which would be the number of positive product of duty i.e., $(-X-)$ or $(+X+)$
3. Count number of paired 'observation n'.
4. Use the formula given below to obtain the value of the coefficient or r_c

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

We will apply the formula in the following example to illustrate the steps.

Example 8:

The following are the marks obtained by a group of 10 students in Economics and Statistics.

Students	1	2	3	4	5	6	7	8	9	10
Economics	10	36	98	25	75	85	91	65	68	34
Statistics	49	39	92	60	68	62	86	58	53	47

Calculate r_c by the method of concurrent deviation.

Solution:

Students	**Marks in Economics**	**Deviation from Preceding item (dx)**	**Marks in Statistics**	**Deviation From Preceding item (dy)**	**dxdy**
1	10	-	49	-	-
2	36	+	39	-	-
3	98	+	92	+	+
4	25	-	60	-	+
5	75	+	62	+	+
6	85	+	62	-	-
7	91	+	86	-	+
8	65	-	58	-	-
9	68	+	53	-	-
10	34	-	47	-	+
	N = 9		N = 9		C = 6

$$r_c = \pm\sqrt{\pm\frac{2c-n}{n}}$$

$$r_c = \pm\sqrt{\pm\frac{2(6)-9}{9}}$$

$$r_c = \sqrt{\frac{12-9}{9}}$$

$$= \sqrt{3/9}$$

$$= 0.58$$

Merits and Demerits of Concurrent Deviation Method:

Merits:

1. As compared to other methods it is the simplest and easiest.
2. When the number of items is very large, this method may be used to form a quick idea about the degree of relationship before making use of complicated methods.

Demerits:

1. The chief demerits is that it is only a very rough indicator of correlation.
2. It does not differentiate between small and big deviation i.e., an increase from 10 to 11 is given the same weight as to 10 to 10,000. i.e., it takes into consideration the direction of change and not he magnitude. It should be however, remembered that the results obtained by this method are not very different from those obtained by the use of Karl Pearson's coefficient in the case of short term oscillation only.

7.5 Summary:

In this lesson, we have analyzed the relationship between two variables. Concept of correlation coefficient was discussed which tell us the extent to which two variables are related.

7.6 Exercise:

1. Explain what is meant by Correlation between two variables. What are the methods of finding existence of Correlation? How can it be meausred?
2. What is Rank Correlation? State the merits and demerits of Spearman's Rank Correlation.
3. Distinguish the following with suitable example:
   (i) Positive and negative correlation
   (ii) Linear and non-linear correlation
   (iii) Simple, partial and multiple correlation

4. (a) Define the Pearson coefficient of correlation. Interprret r , when r = 1, - 1 and 0.

   (b) Mention the uses of Correlation

5. Apply Spearman's Rank difference method and calculate coefficient of correlation between X and Y from the data given below:

X	22	28	31	23	29	31	27	22	31	18
Y	38	25	25	37	31	35	31	29	28	30

6. From the following data calculate Karl Pearson's Coefficient of Correlation.

Year	1977	1978	1979	1980	1981	1982	1983
Exports (in crores)	115	118	122	120	126	127	125
Imports (in crores)	140	138	142	146	145	148	146

7. Calculate Karl Pearson's coefficeint of correlation and interpret its value given the following:

	X series	Y series
Sum of deviation from assumed mean	- 14	18
Sum of squares of deviations from assumed mean	4304	6308
Sum of the products of deviations from their respective assumed mean values	1510	
No.of pairs of observations	12	

Calculate Karl Pearson's coefficient of correlation between X and Y.

8. In order to find out the correlation coefficient between two variables X and Y from 12 pairs of observations, the following calculations were made:

$$\sum x = 30, \sum y = 5, \sum x^2 = 670, \sum y^2 = 285, \sum xy = 334.$$

On subsequent verification it was found that the pair (x = 11, y = 4) was copied wrongly, the correct value being x = 10, y = 14. Find the current value of r.

9. The rank fo the same student in a test in Accoutnancy and Statistics are givenbelow. Calculate the correlation coefficient and comment the value.

Rank in Accountancy	1	2	3	4	5	6	7	8	9	10
Rank in Statistics	1	3	5	6	7	4	8	10	9	2

10. The corresponding values of two variables are given below:

X	2	3	5	8	9
Y	4	6	19	16	18

Karl Pearson's between X and Y is - 1, 0, + 1. None of these.

7.7 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 8

REGRESSION ANALYSIS

Objectives:

After reading this lesson you would be able to

- Calculate a regression line that allows to predict the value of one of the variables if the value of other variable is know.
- Analyse the data by method of least squares to determine the estimated regression line to be used in prediction.

Structure:

8.1 Introduction:

Regression Analysis, in general sense, means the estimation or prediciton of the unknown value of one variable from the known value of the other variable. It is one of the very important statistical tools which is extensively used in almost all sciences - natural, social and physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for estimation of demand and sypply curves, cost functions, production and consumption function.

Meaning:

Prediction or estimation is one of the major problems in almost all spheres of human activity. The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc are of parmount importance to a businessman or economist. Population estimates and population projections are indispendable for efficient planning of an economy. Regression analysis is one of the very scientific techniques for making such predictions.

The variable which is used to predict the variable of interent is called as the independent variable or "explanatory variable" and the variable we are trying to predict is called as "explanatory variable" or "dependent ariable". The independent variable is called X and dependent variable Y. The regression study which confines itself to a study of only two variables is called simple regression. The regression analysis which studies more than two variables at a time is called multiple regression.

It should be noted that the terms dependent and independent refer to the mathematical or functional meaning of dependence - they do not imply that there is necessarily any cause and effect relationship. What it meant is simply that estimates of values of dependent variable Y may be obtained for given values of independent variable X from a mathematical function involving X and Y are dependent on value of X. The X variable may or may not be causing change in the Y variable.

8.2 Difference Between Correlation and Regression Analysis:

Both correlation and regression analysis hekp us in studying the relationship between two variables; however they differ in their approach and objective.

1. The correlation coefficient measures the degree on covariability between the variables whereasthe objective of regression analysis is to study the nature of relationship between the variables which will be abale to help us in prediction of value of one variable on the basis of the other.
2. Correlation between two series is not necessarily a cause and effect relationship. A high degree of positive correlation between two variables does not mean that one is the effect of the other. There may be no cause and effect relationship and yet they may be correlated. Regression, however, presumes one variable as a cause and the other as its effect. It should be noted that the presence of association does not imply causation, but existence of causation always implies association.
3. The coefficient of correlation varies between $\pm$ 1 the regression coefficient have the same high as the correlation coefficient.
4. Correlation coefficient cannot exceed un ity whereas one of the regression coefficients can have a value higher than unity but the product of the two regression coefficients cannot exceed unity because r is the square root of the product of the two regression coefficients.
5. There may be nonsense correlation between two variables which is purely due to chance and has no practical relevance. However there is nothing like nonsense regression.
6. Correlation coefficient is independent of scale and origin Regression coefficients are independent of change or origin but not scale.

8.3 Methods of Studying Regression:

Method of least square:

A mathematical relationship is established between the movement of the variables and algebric equations are obtained to represent the relative movements of X and Y series. One such method is the method of least square. In this method we minimise the sum of squares of the deviations between the given values of a variable and its estimated value given by the line of best fit.

Line or regression of Y on X is theline which gives the best estimate for the value of Y for a specified value of X and similarly the line of regressions of X and Y gives the best estimate for the value of X for specified value of Y. In the method of least square the line of best fit is obatined by the equation of straight line $Y = a + b\,X$ and that in the method of least squares this line is obtained with the help of the following normal equation:

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X \; + b\sum X^2$$

If the values of X and Y variables are substituted in the above equation we get the values of a and b and thus get the regression line of Y on X. Here Y is the dependent variable and X is the dependent variable. To get the regression line of X and Y we will have to assume X as the dependent variable and Y as the independent variable.

Very often the question comes to mind as to why should there be two regression lines to obtain the values of Y and X and why one regression line does not serve the purpose. The answer is simple and it is that one regression line cannot minimize the sum of square of deviation for both X and Y series unless therelationship between them indicates perfect positive or negative correlation.

8.4 Regression Line, Regression Coefficients:

Regression Lines:

If we take the case of two variables X and Y, we shall have two regression lines as regression of X on Y and regression of Y on X. The regression line of Y on X gives the most probable values for Y for given values of X and regression line of X on Y gives the most probable values of X for given value of Y. However when there is a perfect correlation $(r = \pm 1)$ the regression lines will coincide i.e. we will have only one line. The nearer the lines, higher the degree of correlation and the father the two regression lines from each other, the laser is the degree of correlation.

It should be noted that the regression lines cut each other at the point of average of X and Y i.e. if from the point of intersection a perpendicular is drawn on X axis we got mean value of X and if perpendicular is drawn Y axis we get the mean value of Y.

it is important to note that the regression lines are drawn on least squares assumption which stipulates that the sum of squares of the deviations of the observed Y value from the fitted line shall be minimum the total of the squares of the deviations of various points is minimum only from the line of best fit. The deviation from the points from the line of best fit can be measured in two ways vertical i.e. parallel to Y axis and horizontal i.e. parallel to X axis for minimising the total of the squares separately it in essential to have two regression line. The regression line Y on X is drawn in such a way that it minimises total of squares of the vertical deviation and regressionline of X on Y minimises the total squares the horizontal deviations. This can be best appreciated with the help of the following example.

Example 1:

From the following data obtain the two regression equations using the method of least square:

X	1	6	5	2	1	1	7	3
Y	6	1	0	6	1	2	1	5

Solution:

X	Y	X^2	Y^2	XY
1	6	1	36	6
6	1	36	1	6
5	0	25	0	0
2	6	4	36	12
1	1	1	1	1
1	2	1	4	2
7	1	49	1	7
3	5	9	25	15
$\sum X = 26$	$\sum Y = 22$	$\sum X^2 = 126$	$\sum Y^2 = 104$	$\sum XY = 49$

Regression equation Y on X , Y = a + bX

To get the values of a and b the following two normal equations are used:

$$\sum Y = na + b\sum X$$

$$\sum XY = \sum X + b\sum X^2$$

Substituting the values

$$22 = 8a + b(26) \cdots\cdots\cdots (1)$$

$$49 = a(26) + b(126) \cdots\cdots\cdots (2)$$

Multiplying equation (1) by 13 and equation (2) by 4 deducting (4) from (3) we get

$$286 = 104a + 338b \cdots\cdots\cdots (3)$$

$$196 = -104a + 504b \cdots\cdots\cdots (4)$$

$$90 = -166b$$

or b = - 0.54 (approx)

Substituting 'b' in equation (1) we get

$$22 = 8a + 26(-0 \cdot 54)$$

$$= 8a - 14 \cdot 04$$

$$8a = 22 + 14 \cdot 04$$

$$a = 36 \cdot 04/8$$

= 4.51 (approx)

Y = 4.51 - (0.54) . This is the regression equation of Y on X.

Similarly the regression equation of X on Y is X = a + by. The two normal equations are

$$\sum X^2 = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Substituting the values in the above equations

$$26 = 8a + b(22) \quad \cdots\cdots\cdots\cdots (1)$$

$$4a = 22a + b(104) \quad \cdots\cdots\cdots\cdots\cdots (2)$$

Multiplying equation (1) by 11 and (2) by 4 we get

$$286 = 88a + 242\, b \quad \cdots\cdots\cdots\cdots (3)$$

$$196 = 88a + 416 b \quad \cdots\cdots\cdots (4)$$

Deducting (4) from (3) we get

$$90 = -174 b \quad \text{or}$$

b = - 0.52 (approx)

Substituting b value in equation (1) we get

26 = 8a + 22 (- 0.52)

26 = 8a + (- 11.44) or

a = 26 + 11.44/8

= 4.68

Therefore X = 4.68 - (0.52) Y. This is the regression equation of X on Y.

Deviations Taken From Arithmetic Means of X and Y:

The method of obtaining regression equations by the method of least square that we have seen above is very tedious. The work can be simplified to a larger extent if instead of obtaining the regression equations with the help of original values we take the deviations of X and Y series from their respective means. It simplifies the calculation and gives us the same result as given by the method of least square.

The regression equations in such a case are written as follows:

Regression equations of X on Y: $(X - X') = r \frac{\sigma_x}{\sigma_y} (Y - Y')$

X' is the meanof X series and Y' is the mean of Y series.

r is thecoefficient of correlation between X and Y series and σx and σy are the standard deviations of X and Y series respectively.

r $\sigma x/\sigma y$ is called the regression coefficient of X on Y and denoted by bxy. Thus

$$bxy = r \frac{\sigma x}{\sigma y} = \frac{\sum xy}{N \sigma x \sigma y} \times \frac{\sigma x}{\sigma y} = \frac{\sum xy}{N \sigma y^2}$$

$$= \frac{\sum xy}{n \sum y^2 / n} = \frac{\sum xy}{\sum y^2}$$

Thus instead of finding out the values of r, σ_x, σ_y we can directly find out the value of bxy dividing the product of the deviations of X and Y series from their respective means by the sum of the squares of the deviations of the Y series from its mean.

Similarly regression equation of Y on X.

$$(Y - Y') = r \frac{\sigma_y}{\sigma_x} (X - X')$$

r $\sigma y/\sigma x$ is called the regression coefficient of Y on X or 'byx'.

r $\sigma y/\sigma x$ can be calculated in the same way as above and its value will be $\sum xy / \sum x^2$

Thus the two regression equations can be rewritten as follows:

(i) Regression equation of X on Y

$$X - X' = \frac{\sum xy}{\sum y^2} (y - y')$$

(ii) Regression equation of Y on X

$$Y - Y' = \frac{\sum xy}{\sum x^2}(X - X')$$

Regression Coefficient:

As discussed above bxy and byx are called as the regression coefficient of regression equation X on Y and Y and Y on X. The regression coefficients bxy and byx possess some important properties. They are:

1. The under root of the product of two regression coefficeints gives us thealue of correlation coefficients. Symbolically

   $r\sqrt{bxy \times byx}$

   Proof:

   $bxy = r\ \sigma x/\sigma y$ and $byx = r\ \sigma y/\sigma x$

   $bxy\ x\ byx = r\ \sigma x/\sigma y\ x\ r\ \sigma y/\sigma x = r^2$

   $r = \sqrt{bxy \times byx}$

   Then the geometric mean of bxy and byx gives the value of coefficient of correlation.

2. Both the regression coefficients will have signs i.e., either they will be positive or negative. The reason is that so far as standard deviation is concerned they are always positive. Only coefficient of correlation can be either positive or negative. If regression coefficient are negative r will also be negative. For example if bxy = - 1.3 and byx = - 0.65

   $$r = \sqrt{-1\cdot3 \times -0\cdot65} = -0\cdot92 \text{ not } 0\cdot92$$

3. Since the correlation coefficient cannot exceed one, one of the regression coefficients must be less than one or more in other words both the coefficients cannot be greater than one.

The following example would illustrate the use of the above method of obtaining regression equation.

Example 2:

From the following data obtain the two regression equations and their correlation coefficient.

Sales:	46	42	44	40	43	41	45
Purchase:	40	38	36	35	39	37	41

Solution:

Let us denote the sales by the variable X and purchase by the variable Y.

X	X'	x^2	Y	Y'	y^2	XY
46	+ 3	9	40	+ 2	4	+ 6
42	- 1	1	38	0	0	0
44	+ 1	1	36	- 2	4	- 2
40	- 3	9	35	- 3	9	9
43	0	0	39	+ 1	1	0
41	- 2	4	37	- 1	1	2
45	+ 2	4	41	+ 3	9	6
$\sum x = 301$		$\sum x^2 = 28$	$\sum y = 266$		$\sum y^2 = 28$	$\sum xy = 21$

$$X' = \sum x/n = 301/7 = 43$$

$$Y' = 266/7 = 38$$

Regression equation of Y on X

$$Y - Y' = byx\,(X - X') \text{ where } byx = \sum xy / \sum x^2$$

Putting the values

$$Y - 38 = 21/28\,(X - 43)$$

$$Y = 0 \cdot 75\,(X - 43) + 38$$

$$= 0 \cdot 75X + 5 \cdot 73$$

Regression equation of X on Y

$$X - X' = bxy\,(Y - Y') \text{ where } bxy = \sum xy / \sum y^2$$

$$X - 43 = 21/28$$

$$= 0 \cdot 75\,(Y - 38)$$

$$= 0 \cdot 75\,Y + 14 \cdot 50$$

We have

$$r^2 = byx \cdot bxy = 0 \cdot 75 \times 0 \cdot 75$$

$$r = 0 \cdot 75$$

Since both the regression coefficients are positive r must be positive.

Hence r = 0.75

Example 3:

A panel of judges A and B graded seven debators and independently awarded the following marks:

Debator	Marks by A	Marks by B
1	41	33
2	35	40
3	29	29
4	31	32
5	45	39
6	39	36
7	32	29

An eighth debator was awarded 37 marks by judge A while judge B was not present. If judge B had also been present, how many marks do you expect him to award to the eighth debator for assuming that the same degree of relationship exists in their judgement.

Solution:

Let the marks awarded by judge 'A' be denoted by the variable X and marks awarded by judge 'B' be variable Y.

Debator	X	Y	X	Y	X^2	Y^2	xy
				$x - x'$	$y - y'$		
1	41	33	5	- 1	25	1	- 5
2	35	40	- 1	6	1	36	- 6
3	29	29	- 7	- 5	49	25	35
4	31	32	- 5	- 2	25	4	10
5	45	39	9	5	81	25	45
6	39	36	3	2	9	4	6
7	32	29	- 4	- 5	16	25	20
	$\sum X = 252$	$\sum Y = 238$	$\sum x = 0$	$\sum y = 0$	$\sum x^2 = 206$	$\sum y^2 = 120$	$\sum xy = 105$

$$X' = \sum X/n = 252/7 = 36$$

$$Y' = \sum Y/n = 238/7 = 34$$

The equation for regression line of Y on X is given as

$$Y - Y' = byx\ (X - X') \text{ where } byx = \sum xy / \sum x^2$$

$$Y - 34 = 105/206\ (X - 36)$$

$$Y = 0 \cdot 51\ (X - 36) + 34$$

$$Y = 0 \cdot 51\ X - 18 \cdot 36 + 34$$

$$Y = 0 \cdot 51\ X + 15 \cdot 64$$

When X is 37

$$Y = 0 \cdot 51 \times 37 + 15 \cdot 64$$

$$= 18.87 + 15.64$$

$$= 34.51$$

Hence if the judge B had been present he would have given 35 marks to the eighth debator.

Deviation from Assumed Mean:

Where the actual means of the two series X and Y are in fraction, the method of 'A' finding out regression equations discussed above becomes very tedious. In such cases the deviations are taken from the assumed mean. When deviations are taken from assumed mean the entire procedure of finding regression equations remains the same the only difference is we take deviation from the assumed mean. However some simplification is possible. The regression equation of X on Y is

$$X - X' = r\ \sigma x / \sigma y\ (Y - Y')$$

Now the value of $r\ \sigma x / \sigma y$ or bxy will be obtained as follows:

$$bxy = r\frac{\sigma x}{\sigma y} = \frac{\sum dxdy - \frac{\sum d \sum dy}{N}}{\sigma x\, \sigma y} \times \frac{\sigma x}{\sigma y}$$

$$\text{or} \quad bxy = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sigma y^2}$$

$$\text{or} \quad bxy = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sum dy^2 - \frac{\left(\sum dy^2\right)^2}{N}}$$

Similarly the regression equation of Y on X is

$$Y - Y' = r\,\sigma y/\sigma x\,(X - X')$$

$$\text{where } r\frac{\sigma y}{\sigma x} = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sigma y\,\sigma x} \times \frac{\sigma y}{\sigma x}$$

$$= \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sigma x^2}$$

$$bxy = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sum dx^2 - \frac{\left(\sum dx^2\right)}{N}}$$

Once the value of bxy and byx are found out without calculating r, $\sigma x, \sigma y$ they can be inserted in the formula of this regression equation.

It should be noted that in both the cases the numerator is the same, the only difference is in the denominator.

However in case of grouped data given in a two way frequency table the formula is slightly changed. In a correlation table the given frequencies are also taken, therefore the values are to be multiplies by the frequencies. Anothe important change is that if step deviation has to be taken the values given by the formula has to be multiplied by i_x/i_y in case of regression equation of X on Y and by i_y/i_x in case of regression equation Y on X where i_x and i_y are the class intervals of X and Y series. This is necessary because unlike coefficient of correlation, regression coefficience are affected by the change of scale, though they are not affected by change of origin when the regression coefficient are calculated from correlation, table their values are obtained as follows:

$$bxy = r\frac{\sigma x}{\sigma y} = \frac{\sum fdxdy - \frac{\sum fdx \sum fdy}{N}}{\sum fdx^2 - \left(\sum fd\right)^2/N} \times \frac{i_x}{i_y}$$

$$byx = r\frac{\sigma y}{\sigma x} = \frac{\sum fdxdy - \frac{\sum fdx \sum fdy}{N}}{\sum fdx^2 - \left(\sum fdx\right)^2/N} \times \frac{i_y}{i_x}$$

The following example would illustrate the application of the formula given below:

Example 4:

From the following data obtain regression equation taking the deviation from assumed mean.

X	78	89	97	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

Solution:

X	dx (X - A) A = 69	dx^2	Y	dy (Y - A) A = 112	dy^2	dxdy
78	9	81	125	13	169	117
89	20	400	137	25	625	500
97	28	784	156	44	1936	1232
69	0	0	112	0	0	0
59	- 10	100	107	- 5	25	50
79	10	100	136	24	576	240
68	- 1	1	123	11	121	- 11
61	- 8	64	108	- 4	16	32
$\sum x = 600$	$\sum dx = 48$	$\sum dx^2 = 1530$	$\sum y = 1004$	$\sum dy = 108$	$\sum dy^2 = 3468$	$\sum dxdy = 2160$

Regression equation of Y on X

$$Y - Y' = byx\,(X - X')$$

$$bxy = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}}$$

Substituting the values

$$bxy = \frac{2160 - \frac{48(108)}{8}}{1530 - \frac{(48)^2}{8}}$$

$$= \frac{2160 - 648}{1530 - 288} = \frac{1512}{1242} = 1 \cdot 22\ (\text{approx})$$

Therefore the equation of line of regression of Y on X

$$Y - Y' = byx\,(X - X')$$

$$Y - 125 \cdot 5 = 1 \cdot 22\,(X - 75)$$

$$Y = 1 \cdot 22\ X - 91 \cdot 5 + 125 \cdot 5$$

$$= 34 + 1 \cdot 22\ X$$

Regression equation X on Y

$$X - X' = bxy\,(Y - Y')$$

$$bxy = \frac{\sum dxdy - \frac{\sum dx \sum dy}{N}}{\sum dy^2 - \frac{(\sum dy)^2}{N}}$$

$$bxy = \frac{2160 - \frac{48(108)}{8}}{3468 - \frac{(108)^2}{8}}$$

$$= \frac{2160 - 648}{3468 - 1458} = \frac{1512}{2010} = 0 \cdot 75 \ (\text{approx})$$

Therefore the equation of line of regression of X on Y

$X - X' = bxy\,(Y - Y')$ where $X' = 75,\ Y' = 125$ and bxy = 0.75

$$X - 75 = 0 \cdot 75\ (Y - 125)$$

$$X - 75 = 0 \cdot 75\ Y - 93 \cdot 75$$

$$X = 0 \cdot 75\ Y - 93 \cdot 75 + 75$$

$$= 0 \cdot 75\ Y - 18 \cdot 75$$

Example 5:

Following is the distribution of students according to the heights and weights.

Height	**Weights (in lbs)**			
(in inches)	**90 - 100**	**100 - 110**	**110 - 120**	**120 - 130**
50 - 55	4	7	5	2
55 - 60	6	10	7	4
60 - 65	6	12	10	7
65 - 70	3	8	6	3

Calculate the two coefficients of regression and obtain the two regression equations.

Hight in Inches \ Weight 2 lbs		n	90 - 100	100 - 110	110 - 120	120 - 130	f	fdx	fdx^2	fdxdy
			95	105	115	125				
		dy/dx	- 2	- 1	0	1				
50 - 55	52.5	- 1	4	7	5	2	18	- 18	18	13
55 - 60	57.5	0	6	16	7	4	27	0	0	0
60 - 65	62.5	1	6	12	10	7	35	35	35	- 17
65 - 70	67.5	2	3	8	6	3	20	40	80	22
		f	19	37	28	16	100	$\sum fdx = 57$	$\sum fdx^2 = 133$	$\sum fdxdy = 26$
		fdy	- 38	- 37	0	16	$\sum fdy = -59$			
		fdy^2	75	37	0	16	$\sum fdy^2 = 129$			
		fdxdy	- 16	- 21	0	- 11	$\sum fdxdy = -26$			

$$X' = A + \frac{\sum fdy}{N} \times C$$

$$X' = 115 - \frac{59}{100} \times 10 = 109 \cdot 1$$

$$Y' = A + \frac{\sum fdx}{N} \times C$$

$$Y' = 57 \cdot 5 + \frac{57}{100} \times 5 = 60 \cdot 35$$

Regression equation of Y on X

$$Y - Y' = byx\,(X - X')$$

$$byx = \frac{\sum fdxdy - \frac{\sum fdx \sum fdy}{N}}{\sum fdx^2 - (\sum fdx)^2 / N} \times \frac{i_y}{i_x}$$

$$byx = \frac{-26 - \frac{(-59)(57)}{100}}{33 - \left[\frac{57}{100}\right]^2} \times \frac{100}{5} = 0 \cdot 145$$

$$Y - 109 \cdot 1 = 0 \cdot 145 \left(X - 60 \cdot 35\right)$$

$$Y = 0 \cdot 145\ X + 100 \cdot 35$$

Regression equation of X on Y

$$X - X' = bxy\ \left(Y - Y'\right)$$

$$bxy = \frac{\sum fdxdy - \frac{\sum fdx \sum fdy}{N}}{\sum fdy^2 - \left(\sum fd\right)^2 / N} \times \frac{i_x}{i_y}$$

$$bxy = \frac{-26 - \frac{(-59)(57)}{100}}{129 - \left[\frac{-59}{100}\right]^2} \times \frac{5}{10} = 0 \cdot 044$$

$$X - 60 \cdot 35 = 0 \cdot 044\ \left(Y - 109 \cdot 1\right) = 0 \cdot 044\ Y - 4 \cdot 8$$

$$X = 0 \cdot 044 Y + 55 \cdot 55$$

8.5 Uses of Regression:

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. The use of regression analysis done for estimating or predicitng the unknown value of one variable from the known value of other variable. In economics as well as in the field of business this tool of statistical analysis is wedely used. Businessmen an interested in predicting future production, consumption, investment prices, profits, sales etc. In fact the success of a businessman depends much on the correctness of various estimates. In out day to day life also with the help of regression analyasis we can estimate or predict the effect of a variable on the other. In particular, regression analysis attempts an accomplish the following:

1. Regression analysis helps to estimate the dependent variable from the value of an independent variable.
2. With the help of regression coefficient we can calculate the correlative coefficient. The square of correlation coefficient (r) called coefficient of determination measures the degree of association or correlation that exists between the two variables.
3. Regression analysis is also used as a measure of error involved using the regression line as the basis for estimation.

The use of regression concept has varied applications in Physical as well as in social sciences.

8.6 Coefficient of Determination:

Coefficient of determination (r^2) is defined as the ratio of the explaine variance to the total variance.

$$\text{Coefficient of determination} = \frac{\text{Explained Variation}}{\text{Total Variance}}$$

Coefficient of determination is nothing but square of coefficient of correlation i.e. r^2. One very convenient and useful way of interpreting the value of coefficient of correlation between two variables is to use the square coefficient of correlation which is nothing but coefficient of determination. The maximum value of coefficient of determination is unity, because 'r' is never more than one and the explained variation of a dependent variable is an independent variable can never be more than total variation.

It should be noted that the fact that a correlation between two variables as a value of r = 0.66 and correlation between two other variables has a correlation of r = 0.30 does not demonstrate that first correlation is twice as strong as the second. The relationship between the two given values of r can be better understood by computing r^2. When r = 0.60, r^2 = 0.36 and when r = 0.3, r^2 =0.90. This implies that in first case 36% of the total variation is explained and in second case 90% of the total variation explained.

The coefficient of unexplained variance to total variance is called as coefficient of non determination. Thus the coefficient of non determination one minus coefficient of determination. The coefficient of non extermination is denoted by K^2 and its square root is called as coeeficient valienation.

Symbolically $K^2 = 1 - r^2$

$$\text{Coefficient of non determination} = 1 - \frac{\text{Explained Variance}}{\text{Total variance}}$$

$$\text{Coefficient of alienation} = \sqrt{1 - \frac{\text{Explained variance}}{\text{Total variance}}}$$

The coefficient of determination, however is often misinterpreted. It may be interpreted from the very term that X stands in a determining or relationship to Y or r^2 speaks of how much variation X is able to explain of the total variance on Y. However it is a neurtral term whether usually is there between X and Y is to be determined on the basis of evidences other than quantitative measurement on the top of it r^2 being a square it is always a positive number. It cannot tell whether the arrelation is positive or negative. Thus the square root of r^2, $\sqrt{r^2} = \pm r$ is frequently computed to indicate the direction of the relationship in addition to indicating the degree of relationship.

Limitations of Regression Analysis:

With the help of regression the estimates are made but it should be noted that for estimation unless the assumption that has been taken remains unchanged since the equation was computed estimates may so wrong. Another point to be remembered is that the relationship shown in the scatter diagram may not be the same if the equation is extended beyound the values in computing the equation. For example, if we find a close linear relationship between yield of a crop and amount of fertilizer applied, it would not be logical to extend this equation beyound the limits of the experiment for it is quire likely that it the amount of fertilizer were increased indefinitely, the yield would eventually decline as too much fertilizer is applied.

8.7 Summary:

In this lesson, we have shown the concept of regresson analysis and computation procedure. Further, one can predict the furture values using regression analysis.

8.8 Exercises:

1. What is regression analysis ? Indicate its uses in business.
2. Distinguish cleanly between correlation and regression analysis.
3. Given X = 68, Y = 150, $\sigma x = 25, \sigma y = 20$ and correlation coefficient between X and Y as $\pm 0 \cdot 6$ estimate the value of Y, where X is 60.
4. Calculate the regression equation of X on Y and Y on X from the following data and estimate X when Y is 20.

X	10	12	13	17	18
Y	5	6	7	9	13

5. In a regression analysis n = 25, $\sum x = 75, \sum y = 0, \sum Y = 50, \sum x^2 = 623, \sum xy = 30$ and $\sum y^2 = 228$. Find the estimated regression equation and the estimated variouce for the slop parameter.
6. Write down the two regression equations that may be associated with the following pairs of values:

X	152	114	138	154	144	153	141	117	136	154
Y	193	300	414	594	679	549	320	483	481	659

7. A study a price levels of a commodity at two places revealed the following:

	Place A Rs.	**Place B Rs.**
Average Price Per unit	124.60	135.90
Standard Deviation	13.50	17.10
Correlation between A and B	0.72	

(i) Obtain the two regression equations

(ii) Find the expected price level at place A when it is Rs 140 at place B

(iii) Find the expected price level at place B when it is Rs. 120 at place A

8.9 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 9

TIME SERIES ANALYSIS

Objectives:

After studying this lesson, you should be able to

- Identify the components of time-series and compute a linear trend and seasond variation, cyclical variation and irregular varidions.

Structure:

9.1 Introduction:

Time series refers to the chronologically ordered values of a variable over successive time periods. Thus time series refers to such a series in which time is one variable. The analysis of such figures chronologically arranged over the years successively are called anaylsis of time series.

The essential requirements of a time series are:

1. It must consist of a homogenous set of data
2. Data should be available for a sufficiently long period say 7 to 10 years or relevant time period.
3. The time gap between various values must, as far as possible be equal.
4. The gaps, if any, in the data must be made up by interpolation.

The following are some definitions which will clarify the concept of time series.

"A time series consists of data arranged chronologically" - Croxton and Cowden.

"A set of data depending on the time is called time series" - Kenny and Keeping.

"A time series may be defined as a sequence of values of some variable corresponding to successive points in time" - W. Herisch.

"Time series consists of statistical data which are collected, recorded and observed over successive incremetns" - Patterson.

"A time series may be defined as a sequence of repeated measurements of a variable made periodically through time".

9.2 Utility of Time Series Analysis:

The following are the uses of time series analysis:

1. It helps in the analysis of past behaviour of a variable - That is the effect of various factors like technological, economic etc. on a variable over a period of years can be studied by the help of time series analysis.
2. It helps in forecasting - The analysis of past condition becomes the basis for forecasting the behaviour of the variable in future. This helps in making future plans of action.
3. It helps in evaluating the achievements - The review and evaluation of progress in any field of economic and business activity is largely done on the basis of time series data.
4. It helps in making comparative studies - As in time series the data are chronologically arranged, it facilitaties comparison between one period of time with that of the other. It provides a scientific basis for making comparisons.
5. It helps in dissecting the data into various components. These components called by various names as seasonal variation, cyclical variation and irregular variation throw light on the economic behaviour pattern.

9.3 Variation in Time Series:

The term time series is used to refer to any group of ststistical information collected at regular intervals of time. There are four kinds of changes or variation involved in time series analysis. They are:

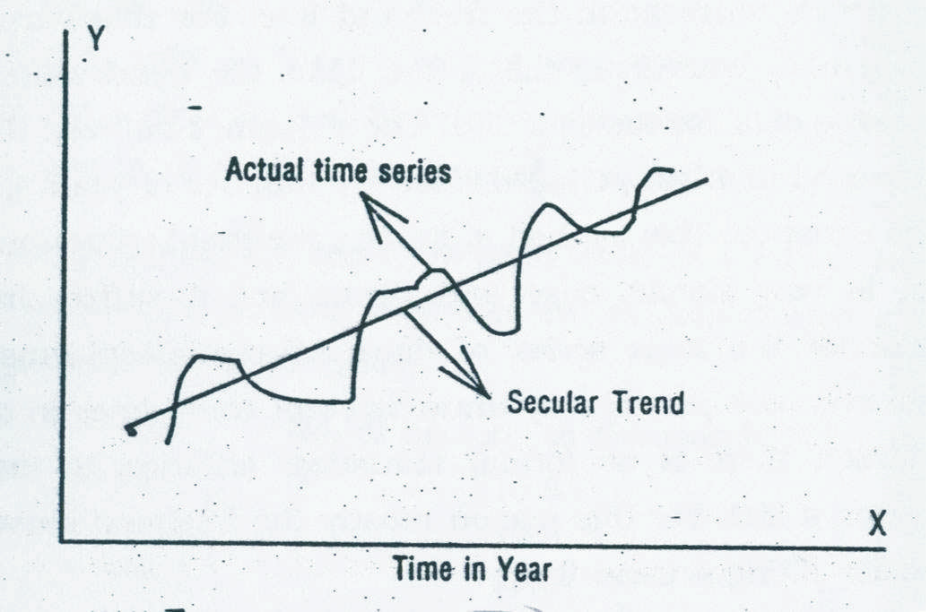
1. Secular Trend
2. Cyclical Fluctuation
3. Seasonal Variation
4. Irregular Variation

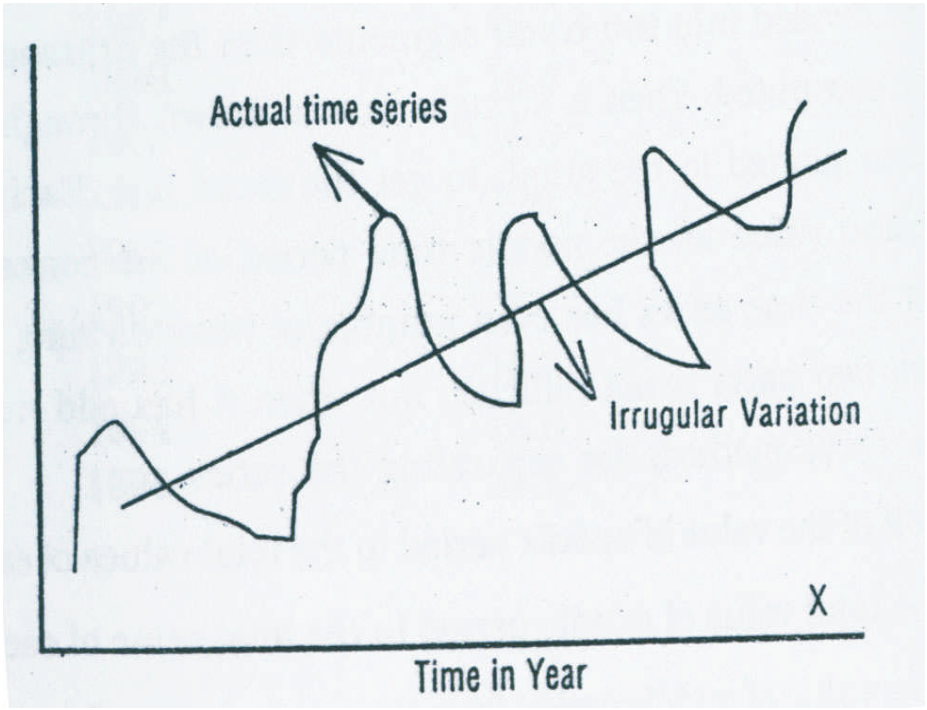
With the secular trend the value of the variable tends to increase or decrease over a long period of time. The steady increase in the cost of living recorded by the consumer price index is an example of secular trend. The cost of living may be varying a great deal from year to year. But when we consider it over a period of time we see that the trend is towards a steady increase. Thus when the variation in the time series is studied for a long period of time i.e., over several years the analysis is called trend analysis.

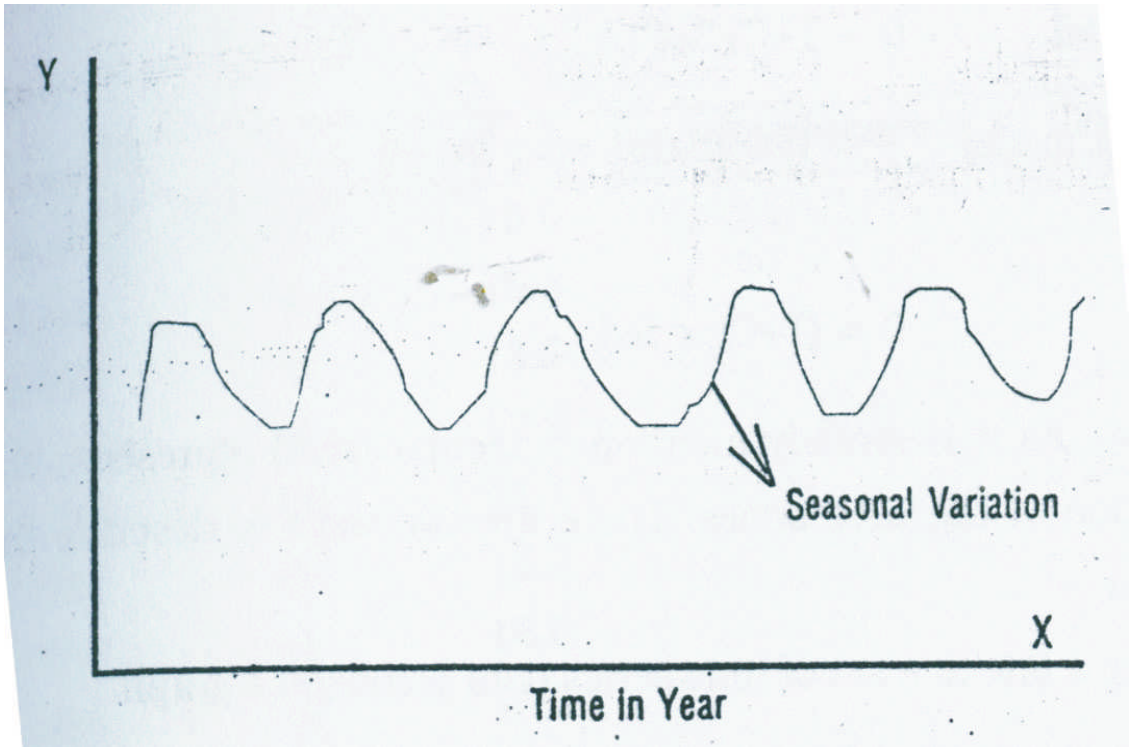
The second type of variation that can be observed in a time series is cyclical fluctuation over a period of time. There are years when numerical information will be in their peak or higher side and at other time will be in lower side. The most common example of cyclical fluction is the business cycle. Over 9 period of time there are years when the business cycle has a peak above the trend line and at other times the activity is likely to slump, touching the low point below the trend line. The time between touching the peak and falling to the low points is usually 3 to 5 years, but it also can be as may as 10 to 15 years also. It should be noted that the cyclical movements do not follow any definite trend but move in a somewhat unpredictable manner.

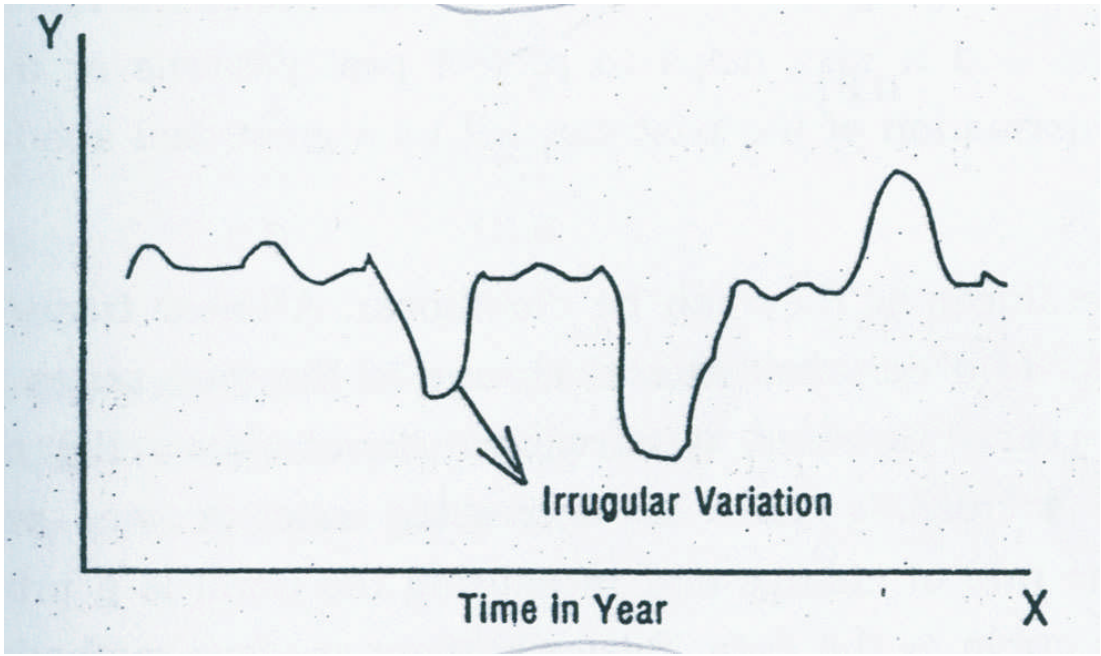
The third kind of fluctuation that can occur in a time series data is the seasonal variation. Seasonal variation involves pattern of changes within a year that tend to be repeated from year to year. For example sale of woollen garments is on the increase during the months of November to January every year because of the seasonal requirement. Since there are regular patterns they are useful in forecastingthe future production run.

Irregular variation is the fourth type of change that can be observed in a time series data. These variations may be due to (1) random fluctuations, no one of which is significantly important to warrant singling out for individual treatment (ii) non-recurring irregular influences that exerts a significant one time impact on the behaviour of time series and as such must be explicityly recognised. The events like flood, strike. war, earthquake etc., which incluence the time series data. The above four variations are shown diagrammatically below:




Actual time series
Irrugular Variation
X
Time in Year


Y
Seasonal Variation
X
Time in Year



The above four variations are generally considered as interacting in a multiplicative manner to produce observed values of the overall time series.

Multiplicative model : $0 = T \times C \times S \times I$

0 = observed values of time series

T = Trend component

C = Cyclical Component

S = Seasonal Component

I = Irregular Component

Other type of models that are possible are:

Additive model $0 = T + C + S + I$

Combination model $0 = T \times C\ (S + I)$

or

$0 = (T + C) \times S \times I$

9.4 Techniques of Measurement:

Trend Analysis:

As it is already mentioned secular trend represents the long term variation of the time series. There are two ways to describe the trend component.

(1) by fitting a line to a set of time series data points on a graph
(2) by fitting a trend line by the method of least square

The reasons for studying secular trend isto understand the historical pattern in the data and it also helps to project past patterns or trends into future. The information of the past can tell us a great deal about the future.

Trends can be linear or they can be curvilinear. A linear trend or a straight line speaks of a constant rate of change in the time series data over a period of years. However, a curvilinear trend shows the trend increasing in an increasing rate or decreasing rate or vice versa. Computation of the rate of change and measuring the trend is a process known as fitting a curve to the data. Basically there are four methods for fitting the trend in time series. These are:

(i) Free hand method
(ii) Method of semi average
(iii) Method of moving average
(iv) Method of least square

Fitting a trend curve involves assuming that a given time series exihibits a certain trend movement. Measuring a trend actually means computing the constants of the equation that we have chosen to be representative of the trend in the data.

(i) Free hand method:

This method is otherwise called as graphic method in the sense that the trend line is determined by inspecting the graph of the series. According to this method the trend values are determined by drawing free hand striaght line through the time series data that is judged by the analyst to represent adequately the longterm movement in the series. Once the free hand line is drawn, a trend equation for the line is approximated. This is done by first readign of the trend values of the first and he lst period from the chart with reference to the freehand line. For this purpose the first period is usually considered the origin. Thus the trend value of first period is the value of a for the equation. The difference between the trend value of first period and last period divided by number of years gives the 'b' value of the equation. this method of finding the trend values and trend line equation is very simple, easy and direct but it suffers from the limitations like for the same series of data different people may draw differet lines even one person may draw different trend lines in different times. In addition there is no formal statistical criterion to judge the adequacy of such a line. For this reason mostly the freehand curve is not recommended for fitting a trend line.

(ii) Semi average method:

To determine the trend values by semi average method, the series in question is first divided into two equal segments then the arithmetic mean of each part is computed. Then a straight line is drawn through the two arithmetic means plotted in the graph to get the trend line. Each average provides the trend value for the middle time period of the corresponding segment. When the time series has even number of years dividing the total time period into two parts is not difficult, but when it has odd number of years there are three methods for separating the series.

(a) Add half of the value of middle period to the total values of each part.

(b) Add the total value of middle period to the total value of each part.

(c) Drop the value of middle period from the computation of the averages.

With semi average method the middle time unit is considered as the origin and the values of Y intercept and slope of the straight line are derived by applying the following equations

$$a = \frac{S_1 + S_2}{t_1 + t_2}$$

$$b = \frac{S_2 - S_1}{t_1\left(n - t_2\right)}$$

where t_1 and t_2 refer to the number of time units for first and second segment in the series, S_1 and S_2 refer to the corresponding partial sum respectively and 'n' is the total number of periods in the series.

Example 1:

Compute the trend by semi average method for the data relating to number of persons registered in an employment exchange in AP during 1981 to 1995.

Year	No.of. persons in thoudands
1981	10.5
1982	15.3
1983	13.5
1984	12.9
1985	11.1
1986	15.9
1987	16.0
1988	16.5
1989	16.0
1990	16.4
1991	19.9
1992	21.7
1993	18.7
1994	18.6
1995	21.5

Solution:

Years	X	No.of. Persons	Semi -average	Trend Value
1981	- 7	10.5		11.6
1982	- 6	15.3		12.3
1983	- 5	13.5		12.9
1984	- 4	12.9	13.6	13.6
1985	- 3	11.1		14.3
1986	- 2	15.9		14.9
1987	- 1	16.0		15.6
1988	0	16.5		16.3
1989	1	16.0		17.0
1990	2	16.4		17.6
1991	3	19.9		18.3
1992	4	21.7	19.0	19.0
1993	5	18.7		19.6
1994	6	18.6		20.3
1995	7	21.5		21.0

The series contains 15 years and is divided into two parts with 7 years in each and the middle year being dropped. The arithmetic mean for the first half is 13.6 and that for second is 19.0. From this two average values we get

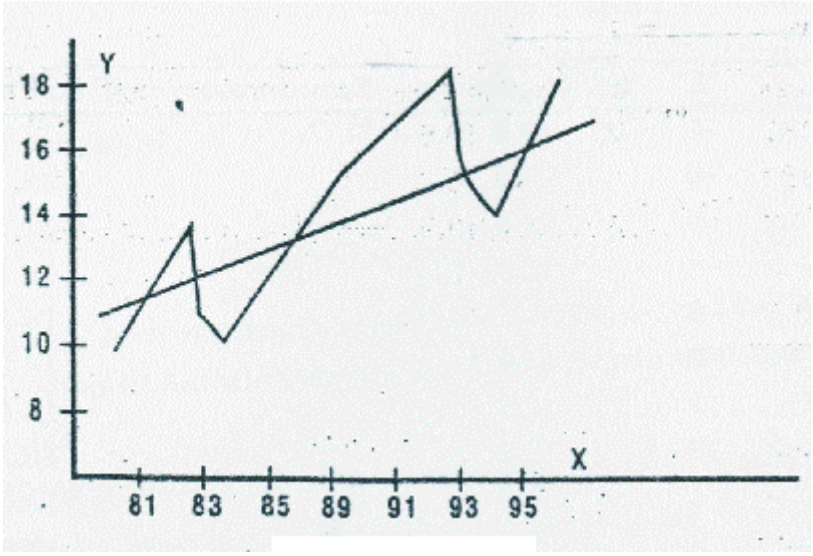
$$a = \frac{95 \cdot 2 + 132 \cdot 8}{7 + 7} = 16 \cdot 3$$

$$b = \frac{132 \cdot 8 - 95 \cdot 2}{7(15 - 7)} = 0 \cdot 67$$

Thus the trend equation becomes

$$Y = 16 \cdot 3 + 0 \cdot 67\ x$$

Trend value of each year can now be determined substituting the X value for that period. The straight line trend can be drawn through the two point 13.6 and 19.



This method of determining the trend is not a subjective one. The slope of the trend line depends on the difference between the averages that are computed from the original values with each average typical of the level of that segmetn of the data. The major drawback of this method is that the arithmetic means may be unduly affected by the extreme values in the series. Thus in the presence of extreme values in the time series data the trend line may not be the true representation of the secular movemetns of the series. Therefore the trend values obtained by this method are not accurate enough for the purpose of forecasting the future trend.

(iii) Method of moving average:

Another method that is used for determining the trend component in a time series is the method of moving average. This method may be considered as an artificially constructed time series in which each periodic figure is replaced by the mean of the value of that period and those of a number of preceeding and succeeding periods. The computation of moving average is simple and straightforward.

The steps in calculating the moving averages are:

1. Compute the moving totals for the number of years asked. If it is a three yearly moving average the value of first three years value is added and written in the center year i.e., against second year. This is first three year moving total. Then the first year value is deleted and fourth year value is included to form the second three year movign total which is centred at the third year i.e., the total of the value of 2nd, 3rd and 4th year is written against 3rd year. In a similar way the computation moves through the end of the series.

2. The moving average is obtained by dividing the moving totals calculated in the previous step by the number of years moving average asked. That is if it is a three yearly moving average the 3 yearly moving total is divided by 3.

However it is to be noted that in computting moving average for an even number of periods. the procedure is slightly complicated. For example calculation of 4 yearly moving average starts with adding up the first four years value in the series to form four yearly moving total. The second moving total is obtained by dropping the value of first period from and adding the value of fifth period to the first four year movign total and so on until all the moving totals have been obtained from the series. Then each moving total is divided by the number of year average (as in this case it is 4). Thus the movign totals are divided by four and 4 yearly moving average is obtained. However the moving totals and moving average so obtained fall between two periods i.e., in this case it falls between 2nd and 3rd year. Whereas data that are typical of a period should be written against the particular year. As the moving averages are written not against a particular year but rather in between two years, we need to centre it i.e., from the averages which is written between 2nd and 3rd year by adding the two averages and dividing it by two. Thus the first centred moving average will fall against 3rd year and secondcentred moving average will be against 4th year and so on if we are calculating 4 yearly moving average.

A moving average of equal length period eliminates the periodic fluctuations and the moving average of equal length willbe linear if the series changes on the average by a constant per time unit and its fluctuations are periodic.

Example 2:

Find the trend value of the following data by the moving average method (take 3 years cycle).

Year	Population in million
1989	512
1990	519
1991	538
1992	548
1993	560
1994	575
1995	590
1996	599

Solution:

Year	Population in million	3 Yearly moving total	3 yearly moving average
1989	512		
1990	519	1569	523
1991	538	1605	535
1992	548	1646	549
1993	560	1683	561
1994	575	1725	575
1995	590	1764	588
1996	599		

Example 3: (Even period of moving average)

From the following series of observations find out 4 yearly moving average.

Year	1989	90	91	92	93	94	95	96
Annual Sales (Rs. '0000)	3	7	1	6	4	9	8	3

Solution:

Year	**Sales in '0000**	**4 yearly moving total not centred**	**4 yearly moving average**	**4 yearly moving average centred**
1989	3			
1990	7			
	→	17	4.25	
1991	1			
	→	18	4.5	4.375
1992	6			
	→	19	3.8	6.4
1993	4			
	→	27	6.75	7.17
1994	9			
	→	24	6	6.39
1995	8			
1996	3			

Moving average method constitutes a stisfactory trend for a series that is basically linear and that is regular in duration and amplitude is a useful technique for analysing the time series data. However the limitations of moving average method are that in computing moving averages we lose some years at the beginning and end of the series. Another drawback of the method is that the moving average is not represented by any mathematical formula and therefore is not capable of objective future projection. Since the major objective of trend analysis is that of forecasting moving average is not used as a trend measure.

(iv) Method of least square:

The earlier discussed methods for trend analysis have certain defects particularly providing a satisfactory projection of the future. To overcome this defect a convenient method is to follow a mathematical approach. The device of getting an objective fit of a straight line to a series of data is the least square method. The line fitted by the method means that the estimates of the constants a and b are the best linear unbiased estimates of those constants.

To determine the value of a and b in a linear equation by least square method we are required to solve the following two normal equations simultaneously.

$$\sum y = Na + b\sum x$$

$$\sum xy = a\sum x\ b\sum x^2$$

In the case of time series analysis the solution of a and b from these equation is simplified by using the middle of the series as the origin. Since the time units in a series are usually of uniform duration and are consecutive numbers then the middle point is taken as the origin the sum of time units i.e. $\sum x$ will be zero. As a result the above two normal equations reduce to

$$\sum y = Na$$

$$\sum xy = b\sum x^2$$

Therefore we can get

$$a = \frac{\sum y}{N} \quad \text{and} \quad b = \frac{\sum xy}{\sum x^2}$$

It should be noted that in computing the trend it is convenient to use the middle of the series as the origin and get $\sum x$ as zero. If the serie contains an odd number of periods the origin is the middle of the given period. If an even number of periods is involved the origin is ser between te two middle periods.

Example 4:

Given below are the figures of production (in thousand tons) of a sugar factory.

Year	1990	91	92	93	94	95	96
Production (in '000 tons)	40	47	49	42	45	50	52

Fit a straight line trend by least square method and tabulate the trend value.

Solution:

Year	**X (taking 93 as origin)**	**Y**	X^2	**XY**	**Trend value**
1990	- 3	40	9	- 120	42.35
1991	- 2	47	4	- 94	43.71
1992	- 1	49	1	- 49	45.07
.1993	0	42	0	0	46.43
1994	1	45	1	45	47.79
1995	2	50	4	100	49.15
1996	3	52	9	156	50.51
	$\sum x = 0$	$\sum y = 325$	$\sum x^2 = 28$	$\sum xy = 38$	

$N = 7 \; ; \sum y = 325 \; ; \sum xy = 38 \; ; \sum x^2 = 28$

$a = \sum y / N = 325/7 = 46 \cdot 43$

$b = \sum xy \Big/ \sum x^2 = 38/28 = 1 \cdot 36$

Example 5:

Fit a straight line trend equation by the least square method and estimate the trend values.

Year	1988	89	90	91	92	93	94	95
Value	80	90	92	83	95	99	94	106

Solution:

In this case since N is 8 (even number) we have the origin to the time which is the arithmetic mean of two middle terms 1991 and 1992.

Year	**X (Deviation from origin 91.5)**	**Y (Deviation multiplied by 2)**		**XY**	X^2	**Yc**
1988	- 3.5	- 7	80	- 560	49	87.475
1989	- 2.5	- 5	90	- 450	25	88.875
1990	- 1.5	- 3	92	- 276	9	90.275
1991	- 0.5	- 1	83	- 83	1	91.675
1992	0.5	1	95	95	1	93.075
1993	1.5	3	99	297	9	94.475
1994	2.5	5	94	470	25	95.875
1995	3.5	7	106	742	49	97.275
		$\sum x = 0$	$\sum y = 739$	$\sum xy = 235$	$\sum x^2 = 168$	

$Y = a + bx \;\; ; \; a = \sum y / N \;\;\; ; \;\;\; b = \sum xy \Big/ \sum x^2$

$a = 739/8 = 92 \cdot 375 \qquad b = 235/168 = 1 \cdot 40$

The merits of this method is that it is a mathematical method and it does not have any subjectivity. Trend values can be obtained for all the given time periods in the series, which is not possible in other methods like moving average and it is useful for future prediction. However this method is tedious, time consuming as well as very rigid. In this method only long term variation can be studied and the impact of cyclical, seasonal and irregular variations are ignored.

Measurement of Seasonal Index:

By seasonal variation in a time series we mean the variations of reular and periodic nature with period less than one year. If the variation is for longer than one year period they will be considered under cyclical variation. The variations can be measured in terms of relative seasonal factor in rates or percentages. The climination of seasonal variation from time series is known as deseasonalisation. The deseasonalisation helps in the process of decomposition of atime series into various components viz., trend, seasonal, variation, cyclical variation and treegular or random variation.

The following are important methods of studying seasonal variations:

1. Method of simple average
2. Ratio to trend method
3. Ratio to moving average method
4. Link relative method

1. Method of simple average:

The procedure of computing seasonal index by this method is

(1) Arrange the data by year, moths or quarters as the case may be.

(2) Compute the arithmetic average for each period i.e., month, year or quarter.

(3) Obtain the overall average X for a month, or quarter from all monthly or quarterly averages obtained in step (2).

$$\text{i.e., } X = \frac{X_1 + X_2 + X_3 + \cdots\cdots + X_{12}}{12} \text{ or } X = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

(4) Seasonal indices for different months or quarters are obtained by expressing each monthly (quarterly) averages as a percentage of overall average X

Seasonal index for kth month (or quarter)

$$\frac{\text{Monthly Average}}{\text{Total Average}} \times 100 \quad \text{i.e.,} \quad \frac{X_k}{x} \times 100$$

where k = 1,2, ... to 12 & k (1 ..... to 4)

Example 6:

Use the method of monthly average to determine the quarterly indices from the following data ofproduction of a certain commodity for the year 1993, 1994, 1995 and 1996.

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1993	33	44	40	37
1994	42	47	42	39
1995	44	48	39	36
1996	50	54	42	34

Solution:

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1993	33	44	40	37
1994	42	47	42	39
1995	44	48	39	36
1996	50	54	42	34
Total	169	193	163	146
Average	42.25	48.25	40.75	36.50
Seasonal Index	100.74	115.05	97.17	87.03

$$\text{The average of averages} = \frac{42\cdot25 + 48\cdot25 + 40\cdot75 + 36\cdot5}{4}$$

$$= 41.938$$

$$\text{Seasonal index} = \frac{\text{quarterly average}}{\text{grand average}} \times 100$$

$$\text{Thus seasonal index of 1st quarter} = \frac{42\cdot25}{41\cdot938} \times 100 = 100\cdot74$$

$$\text{Thus seasonal index of 2st quarter} = \frac{48\cdot25}{41\cdot938} \times 100 = 115\cdot05$$

$$\text{Thus seasonal index of 3st quarter} = \frac{40\cdot75}{41\cdot938} \times 100 = 97\cdot17$$

$$\text{Thus seasonal index of 4st quarter} = \frac{36\cdot50}{41\cdot938} \times 100 = 87\cdot03$$

Ratio to Trend Method:

This method is an improvement over the average method. This method also is otherwise known as percentage to trend method. This method assumes that seasonal variation for various seasons (month for monthly data quarter for quarterly data) is a constant factor of trend. This method takes into cognizance of the effect of trend on time series. The steps involved in computaion of seasonal indices by this method are as follows:

1. Obtain the trend values by least square method by reson wise.
2. Divide the original data by coresponding trend values and multiply these ratios by 100. Thus we express the original data as a percentage to trend value.
3. The effect of cyclical and irregular movements are eliminated by the process of averaging the percentage for each unit of time. Either arithmetic mean or median can be used for averaging. Thus season wise figures for various years are averaged and this average gives the preliminary indices of seasonal variation.
4. The preliminary indices obtained in step 930 are adjusted to a total of 1200 for monthly data or 400 for quarterly data by multiplying each of them by a constant factor 'k' given by

$$k = \frac{1200}{\text{Sum of monthly indices}} \text{ and } \frac{400}{\text{Sum of quarterly indices}}$$

Example 7:

Find the seasonal variation by ratio to trend method from the data given below:

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1991	40	50	46	44
1992	44	62	60	54
1993	50	68	64	58
1994	64	86	78	72
1995	90	102	96	92

Calculation of trend by least square method:

Year	X (Deviation from middle year 93)	X^2	Y (Yearly total	Yearly average	XY	Trend value
1991	- 2	4	180	45	- 90	42
1992	- 1	1	220	55	- 55	54
1993	0	0	240	60	0	66
1994	1	1	300	75	75	78
1995	2	4	380	95	190	90
N = 5		$\sum x^2 = 10$		$\sum y = 330$	$\sum xy = 120$	

The equation of straight line trend is $Y = a + bx$

$a = \sum y/N$ $\qquad b = \sum xy \Big/ \sum x^2$

$a = 330/5 = 66$ $\qquad b = 120/10 = 12$

Quarterly incremetn $= 12/4 = 3$

Now we calculate quarterly trend value consider 1991 the trend value of the middle quarter is 42 i.e. half of 2nd and half of 3rd quarter. Quarterly increment is 3 so the trend value of 2nd quarter is $42 - 3/2 = 40 \cdot 5$ and 3rd quarter is $42 + 3/2 = 43 \cdot 5$ trend value of first quarter is $40 \cdot 5 - 3 = 37 \cdot 5$ and 4th quarter is $43 \cdot 5 + 3 = 46 \cdot 5$

We thus get quarterly values as shown below:

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1991	37.5	40.5	43.5	46.5
1992	49.5	52.5	55.5	58.5
1993	61.5	64.5	67.5	70.5
1994	73.5	76.5	79.5	82.5
1995	85.5	88.5	91.5	94.5

The given values are expressed as percentage of the corresponding trend value. Thus the value for first quarter will be $(40/37 \cdot 5 \times 100) = 106 \cdot 67$ for 2nd quarter $(50/40 \cdot 5 \times 100) = 123 \cdot 46$ etc.

Quaterly values as percentage to trend value would be

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1991	106.67	123.46	105.75	94.65
1992	88.89	118.09	108.11	92.31
1993	81.30	105.43	94.81	82.69
1994	87.07	112.42	91.11	87.27
1995	105.26	115.25	109.92	97.35
Total	469.19	574.65	504.7	454.24
Average	93.84	114.93	100.94	90.85
SI adjusted	93.71	114.77	100.80	90.72

Total of average = 93.84 + 114.93 + 100.94 + 90.85 = 400.56

Since the total of average is more than 400 an adjustment is made by multiplying each average by 400/400.56.

Ratio to moving average method is a more logical procedure for measuring seasonal index. This method eliminates both trend and cyclical fluctuations from the time series. This method eliminates both trend and cyclical fluctuations fromt he time series. But if cycles are not regular and are of different intensity seasonal index calculated by above method would contain some effect of cyclical variation.

Ratio to moving average method:

This method is also known as percentage of moving average method. The computation of seasonal indices by this method is similar to that of ratio to trend method except that in place of least square trend, moving average trend is used. The various steps involved in the computation of seasonal indices by the ratio to moving average method are as follows:

1. Obtain 12 monthly (4 quarterly) centered moving average for thegiven series
2. Express each original value on the time series as a percentage of the trend value
3. Arrange these percentages season - wise for all the years and average it. These would be preliminary seasonal indices.
4. If the sum of the indices is not 1200 (for monthly) or 400 (for quarterly) figures multiply them by the correction factor (c.f.) 1200/sum if monthly indices or 400/sum of quarterly indices as the case may be.

   These are the ratio to movign average seasonal indices.

Example 8:

Calculate the seasonal indices by the ratio to moving average method from the following data:

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1993	70	64	63	65
1994	67	60	68	63
1995	70	65	65	69

Solution:

Year	Quarter	Given figure	4 figure moving total	2 figure moving total	4 figure moving average	Given figure % to moving average
1993	I	70				
	II	64				
			→ 262			
	III	63	→	521	65.125	96.73
			→ 259			
	IV	65	→	514	64.250	101.67
			→ 255			
1994	I	67	→	515	64.675	104.08
			→ 260			
	II	60	→	518	64.750	92.66
			→ 258			
	III	68	→	519	64.875	107.82
			→ 261			
	IV	63	→	527	65.875	95.64
			→ 266			
1995	I	70	→	529	66.125	105.83
			→ 263			
	II	65	→	532	66.500	97.74
			→ 269			
	III	65				
	IV	69				

Calculation of seasonal index (Trend eliminated values)

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1993	-	-	96.73	101.67
1994	104.08	92.66	104.82	95.64
1995	105.86	97.74	-	-
Total	209.94	190.40	201.55	197.31
Average	104.97	95.20	100.76	98.66
SI adjusted	105.08	95.30	100.86	98.76

$$\text{Arithmetic average of averages} = \frac{104 \cdot 97 + 95 \cdot 20 + 100 \cdot 76 + 98 \cdot 66}{4}$$

$$= \frac{399 \cdot 59}{4} = 99 \cdot 90$$

By expressing each quartetly average as percentage the average of average is 99.90.

This method of measuring seasonal variation is considered to be most satisfactory as in practice most widely used. This index does not fluctuate so much as the index based on straight line method. However seasonal index cannot be obtained for each month if 12 month moving average is taken six month in the beginning and six month at the end are left out forwhich we cannot calculate seasonal indices.

Link relative method:

The constructuon if indices of seasonal variation by link relatives method is also known as Pearson's method, involves the following steps:

1. Convert the original data into link relatives by the formula

$$\text{Link relative} = \frac{\text{Current season's figure}}{\text{Previous season's figure}} \times 100$$

2. Average these link relatives for each month (or quater or other time period) while calculating the average mean or median may be used. Howwever median would be a better average as it is not influenced by extreme item.

3. Convert the average link relatives (LR) into chain relatives (CR) on the base of the first season by the formula

$$\text{CR of any month} = \frac{\text{LR of that month} \times \text{CR of preceding month}}{100}$$

4. Earlier it was assumed the CR of first month or quarter as 100. The new CR of January based on CR of December would not necessarily be 100. The difference between two CRs of January. Therefore needs correction.

5. The correction is done by subtracting a correction factor from each chain relative. The correction factor is

$$\text{Cf} = \frac{\text{New chain relative} - \text{old chain relative or } 100}{12}$$

If figures are given quarterly the correction factor would be

$$\text{Cf} = \frac{\text{CR (New)} - \text{CR (Old)}}{4}$$

The correction factors for February would be 2 x Cf for march 3 x Cf. Likewise Cf for second quarter would be 2 x Cf for third quarter 3 x Cf and fourth quarter 4 x Cf. These are corrected chain relatives.

6. Express the corrected chain relatives as percentages of their averages. These provide the required seasonal indices by the method of link reltives.

Example 9:

Calculate sesonal indices by the link relative method for the following data:

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1991	70	64	63	65
1992	67	60	58	63
1993	70	65	65	69
1994	72	61	58	64
1995	62	57	53	58

Solution:

Link Relative

Year	1st Quarter	2nd Quarter	3rd Quarter	4th Quarter
1991		91.43	98.44	103.17
1992	103.08	85.71	96.67	108.62
1993	111.11	92.86	100.00	106.15
1994	104.35	94.72	95.08	110.34
1995	96.87	91.93	92.98	109.43
Total	415.41	446.65	481.17	537.71
Mean LR	103.85	89.33	96.23	107.54
CR	100.00	$\frac{100\times89}{100}$ = 89	$\frac{89\times96\cdot23}{100}$ = 85.64	$\frac{85\cdot64\times107\cdot57}{100}$ = 92.09
Adjusted CR	100	89 (1.09) = 90.09	85.64 - (- 2.18) = 87.82	92.09(- 3.27) = 95.36
Seasonal Indices	100 x 1.071 = 107.2	90.09 x 1.071 = 96.49	87.82 x 1.071 = 94.06	95.36 x 1.071 = 102.13

The second CR of first quarter $= \frac{92 \cdot 09 \times 103 \cdot 85}{100} = 95 \cdot 64$

The difference between the old and new chain relatives of first quarter is 95.64 - 100.00 = - 4.36. Thus difference in one quarter $-4 \cdot 36/4 = -1 \cdot 09$. Subtract $(1 \times 1 \cdot 09)$, $(2 \times 1 \cdot 09)$ and $(3 \times 1 \cdot 09)$ from 2nd, 3rd and 4th quarter respectively. The total of adjusted CRs is $100 + 90 \cdot 09 + 87 \cdot 82 + 95 \cdot 36 = 373 \cdot 27$.

Thus seasonal indices will be obtained by multiplying adjusted chain relative by a correction factor $400/373 \cdot 3 = 1 \cdot 071$

Measuremetn of Cyclical Variation:

Measurement of cyclical fluations is a difficult proposition. This is because successive cycles vary so widely in timing amplitude and pattern and because the cyclical rhythm is very closely intertwined with irregular factors. Because of this reason it is impossible to construct meaningful typical index.

1. Residual method
2. Referenic cycle analysis method
3. Direct method
4. Harmonic analysis method

Residual Method:

This is the most common method for estimating the cyclical movement of time series. This method consists of eliminating seasonal variation and trend thus obtaining the cyclical irregular movents. Symbolically

$$\frac{T \times S \times C \times I}{S} = T \times C \times I \text{ and } \frac{T \times C \times I}{T} = C \times I$$

Next the data are usually smoothed in order to obtain the cyclical relatives, since they are expressed in percentage. Thus under residual method the steps involved are:

1. Obtain trend value (T) and seasonal indices
2. Divide the original data by T to get SCI then divide SCI by S to get CI
3. Take the moving total of CI value for monthly series generally monthly moving total is taken. The moving total is generally taken by giving weights of 1.2.1 to the three months. After this the weighted moving total is divided by 4 to get the moving average. If 5 monthly moving average is taken the weights are 1, 2, 4, 2, 1. These moving averages are the cyclical variation of the series.

Referenic Cycle Analysis:

Under this method the index of variation in each series is calculated with reference to a given referenic year which may be a peak or trough year of the cycle. Obviously from peak year of reference the economic series will show downward falling index till the cycle takes a turn or vice - versa will betrue if the reference year is the peak year.

The steps involved in this method are as follows:

1. Select reference data which are the data of peaks and troughs of a business cycle refers to general business movements.
2. Obtain a cyclical pattern for each series for the period between each two successive troughs. Each period is same for all series so that comparison between the series may be possible.
3. To obtain a cyclical pattern for aseries, the data are adjusted for seasonal variation. These are then divided into reference cycle segments for each segment the monthly value are expressed as percentage of average of all values in the segment.
4. Each reference cycle segmetn is broken into nine stages to correspond to the same 9 stages in the business cycle and the reference cycle relative calculated ins tage 3 are averaged for each of the 9 stages.

The nine stages are:

1. The 3 months centred on initial trough
2. The first third of expansion period
3. The second third of expansion period
4. The last third of expansion period
5. The 3 months centred on the peak
6. The first third of contraction period
7. The second third of contraction period
8. The last third of centraction period
9. The three month centred on the terminal trough

The nine state averages for each reference cycle segment helps to reduce the ematic movement in a series and thus give a reference cycle pattern for a particular series under consideration.

Direct Method:

It is a method based on calculating variation each month or quatter with respect to previous year same month and quarter to see upward changes in case of rising cycle and downwards incase of decliming cycle. This roughly results in eliminating seasonal variation and trend.

Harmonic Analysis:

If the cyclical variations are of same duration and the amptitude of its various phases in constant a suitable curve may be fitted. Cuve is fitted after the irregular movements have been smoothed the advantage of it is that it will have a prediction value.

Measurement of irregualr variation:

Irregylar movements are those which are left after the cyclical irregular movents have been smoothened. But the very nature of these movements which are erratic no special formula can be suggested to isolate or identify irregular fluctuation. However from a time series if trend, cyclical and seasonal variations are taken at the residual is nothing but the irregular variation.

In multiplicate model $\frac{Y}{TSC}$ or $\frac{TSCI}{TSC} = I$

(where S and C are intractional form not percentase)

In additive model the irresular ariations are

$$Y - (T + S + C) \quad \text{or} \quad T + S + C + I - (I + S + C)$$

However in actual practive only trend and seasonal variations are isolated and the cyclical and irregualr fluctuationsare kept to gether because cycles differ in period and amptitude and irregular fluctuations are so mixed up with cyclical fluctions that it is impracticable to separate them in a meanningful manner.

9.5 Summary:

(a) Time series is a series of observations recorded over time. Its components are

(i) Secular trend - long term smooth change over time;

(ii) Seasonal variation - regular cyclical variation with a period of one year;

(iii) Cyclical variation - Oscillatory movement with periods of varying length (eg: one day, week, year or even 50 years)

(iv) Irregular (random) variation - unpredictable erratic variation

(b) A few methods of trend determination are

(i) free - hand method

(ii) semi - averages method

(iii) moving averages method

(iv) method of least squares

Trend is eliminated by substracting the trend value under an additive model and by division by the trend value under a multiplicative model.

(c) Seasonal variation is eliminated by 12 monthly moving averages method for monthly data. It may be computed by the trend ratio method.

(d) Trend is used for prediction and control purposes by planners and institutions like Reserve Bank of India. Refirement of prediction can be mode by using seasonal and cyclical indices.

If the time series is short, then one cannot expect the prediction to be accurate. If there is no stability in the system, or if the irregular variation is large, trend value might not be reliable.

9.6 Exercises:

1. What is a time - series? How is it represented graphically? Explain with a diagram.
2. Describe different components of a time series.
3. The following data are annual sales of bicycles (in millions) in India during 1970 to 1989 (Read the data row wise)

6.7	5.3	4.3	6.1	5.6	7.9	5.8	6.1	4.3	5.6
6.7	5.5	6.9	7.6	7.8	9.3	8.6	7.4	6.2	5.8

(a) Draw a historigram

(b) Does the series appear to have only irregular component a part from a constant trend?

4. Explain secular trend. Describe the method of fitting a linear trend by the free - hand method and the method of semi - averages.
5. For the data in exercise 3 fit a linear trend by (a) free - hand method (b) the method of semi-averages.
6. Describe the method of moving averages for measuring secular trend. Point out its advantages and disadvantages.
7. Obtain the 3 - yearly moving average for the following data. Plot the original and trend values on a graph paper.

Year	1964	65	66	67	68	69	70	71	72
Imports (000 Kgs)	87	62	67	74	85	87	96	97	84

8. Describe the method least squares for determining trend. Find the linear trend - values by the least squares method from the following data.

Year	1984	85	86	87	88
Production (in 000 Kgs)	35	55	79	80	60

Plot the original and the trend - values on the some graph.

9. How will you compare goodress of fit of two trends?

10. Discuss the additive and multiplicative models in time - series.

11. What are the uses of (i) trend - analysis (ii) trend elimination?

12. The following is the price of rice in Rs. for 50 Kg.

Year	1960	64	68	72	76	80	84	88	92
Price	55	80	135	180	245	260	300	390	550

(i) Draw a historigram

(ii) Fit a linear trend by the method of semi - averages and by the free hand method.

(iii) Fit (a) linear trend and (b) quadratic trend by the method of least squares and draw both of them on the same graph as historigram.

(iv) Compute the residual sum of squares and trend - free data in the cases (a) and (b) of (iii)

(v) Estimate the price of rice in the year (a) 1989 (b) 1994 using the trend curves (a) and (b) of (iii)

13. The following data are based on radio/T.V. marketing research survey in an area.

Year	Radio audience (in millions)	T.V. audience (in millions)
1985	31	3
1986	32	3
1987	33	6
1988	30	7
1989	29	10
1990	30	11
1991	28	14
1992	26	14
1993	24	17

(a) Fit suitable trend curves for radio and T.V. audience series. Comment on their relation.

(b) Forecast the total audience for both radio and T.V. in 1994 what is the likely percentage of radio - audience in 1994?

14. Production figures (in lakhs of cases0 of apples in Himachal Pradesh during 1960 - 85 are given below. (i) Fit a trend by taking 5 year moving averages (ii) Eliminate trend under model II.

Year	Production	Year	Production	Year	Production
1960	16.8	1968	29.9	1976	53.7
1961	12.2	1969	25.6	1977	58.6
1962	14.5	1970	28.6	1978	66.2
1963	15.9	1971	27.2	1979	76.9
1964	15.6	1972	37.2	1980	70.5
1965	15.9	1973	46.2	1981	89.1
1966	19.1	1974	42.8	1982	70.5
1967	23.9	1975	49.8	1983	89.1
				1984	85.9
				1985	88.2

15. When do you fit a parabolic (a quadratic) trend to a given time series? Does it always provide a better fit than a linear trend? Explain L.S. method of fitting a parabolic trend.

16. Food grain and steel production in India during the years 1985 - 86 to 1991 - 92 are given below (in million tonnes). Fit a parabolic trend to the series on food grains and linear trend to that on steel production. Estimate the values for 1995.

Year	Food Grain	Steel
1985 - 86	150.4	9.49
1986 - 87	143.4	9.55
1987 - 88	140.4	11.68
1988 - 89	169.9	12.84
1990 - 91	176.4	13.53
1991 - 92	167.1	14.33

17. Opinion polls in U.S.A. gave the following data on the question "would you vote for a woman as president if she is qualified and seemed best for the job?"

	1937	1945	1955	1967	1971	1985	1991
Yes %	34	33	52	57	66	80	81
No%	66	55	44	39	29	18	19

Using a computer package,

(i) Fit a straight line to the yes (%) over time by L.S. method. What is the likely percentage in 1995 of those who say 'yes'?

(ii) Fit a quadratic curve to the above time - series data on No (%) and forecast the percentage in 1995.

9.7 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 10

INTERPOLATION AND EXTRAPOLATION

Objective:

After going through this lesson, you will learn:

- The concept and significance of Inter polation and Extrapolation
- Computation of Binomial Expansion method, newton's method Lag Tange's Method and Parabolic Curve Method.

Structure:

10.1 Introduction:

Many a time in practical work we come across situations where we have to estimate a value which is not available in a given series or predict future value. For example the census of population of India takes place every 10 years i.e. we have figures for 1931, 1941, 1951, 1961, 1971, 1981, 1991. Now if we require the population of 1989 and 1996 what should we do? This can be done either by guessing or by analysing the data and inserting a value in between a series of data or projecting the forward. The process of inserting a value in between the series of data is called interpolation and projecting forward or backward onthe basis of series of values given is known as extrapolation. Thus interpolation supplies us with the missing link and extrapolation helps in forecasting.

10.2 Definition:

The following definitions give a formal expressionto the basic idea of interpolation and extrapolation.

"Interpolation consists in reading a value which lies between two extreme points. Extrapolation means reading a value that lies outside the extreme points" - W.M. Harper.

"Interpolation is the estimation of a most likely estimate in a given condition. The technique of estimation fo a past figure is termed as interpolation while estimating a probable figure for the future is called extrapolation" - Hirach.

There is no difference between interpolation and extrapolation so far as the methods are concerned but for distinguishing past from the future we give them two different names. Interpolation relates to past whereas extrapolation gives us the forecast for the future.

10.3 Significance of Interpolation and Extrapolation:

The tools of interpolation and extrapolation are of great practical significance. Their utility can be well imagined in the following situations.

(1) Estimation of Intermediate Values:

It often happens in business or economic situations, a particular type of information is collected at regular intervals. It is likely that at some future date it may be felt that data for the intermediate period is necessary. The only alternative is to use the technique of interpolation. The most common examples are intercensus population figures, mid term figures of industrial production etc.,

(2) Non - availability of data or loss of data:

In case of non - availability of data extrapolation helps to estimate the value of some past period and to fulfill the gap in the data caused on account of any loss or destruction of data and interpolation becomes useful.

(3) Derivation of median and mode:

In case of continuous frequency distribution the interpolation technique is used to derive the formula for computation of median and mode.

(4) Bringing uniformity in the data:

Sometimes data pertaining to a particular phenomenon are grouped by different agencies by different agencies in different types of groups which make them unfit for comparison. To bring uniformity in groups interpolation technique is used.

(5) Making of forecast:

For a number of phenomena estimates for future have to be made. Extrapolation is a scientific technique that can be used for estimating or projecting certain phenomena. However the accuracy of interpolation depends on (1) knowledge of the possible fluctuations of the figures to be obtained by a general inspection of the fluctuations at dates for which they are given and (2) knowledge of course of events with which the figures are connected.

10.4 Assumptions:

The following assumptions are made while making use of the techniques of interpolation and extrapolation.

1. There is no sudden jump in the figures from one period to another within the period under consideration. In other words the given data do not refer to abnormal periods such as famine, war, drought, epidemic, etc. which may result in sudden change in the series.
2. The rate of change of figures from one period to another is uniform.

The limitaions however of these techniques may be

1. A number of consecutive missing values in a series cannot be estimated at any rate they cannot be reliable.
2. Unless there is fairly good number of observations, the technique of interpolation and extrapolation fail to deliver desired result.
3. Unless the gaps between the values are equal these methods cannot be used.

10.5 Methods of Interpolation:

Broadly speaking the various methods of interpolation can be divided under two heads.

1. Graphic method and
2. Algebric method

Under algebric head there are several methods. The following are some of the important and more popular methods:

1. Binomial expansion method
2. Newton's method
3. Lagrange's method
4. Parabolic curve method

10.5.1 Graphic Method:

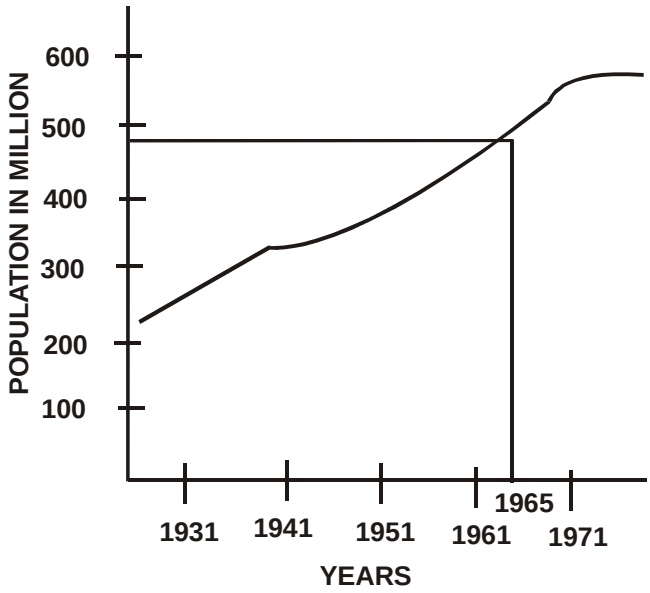
This is the simplest of all types of interpolation methods. When this method is used the given data are plotted in a graph paper and the plotted points are joined. If there are only two values a straight line is obtained otherwise a curve will be obtained. On the X - axis the years are taken and in the Y - axis the values of the variables. For the year the value is to be interpolated, a perpendicular line is drawn on the line or curve, whatever may be depending on the number and values of the variables. From the point where the perpendicular drawn meets the curve, another perpendicular is drawn on the Y -axis. The corresponding value in Y -axis is the required value of the variable for the desired year. The following example will illustrate it.

Illustration:

From the following data determine the population for the year 1965.

Year	Population in millions
1931	251
1941	279
1951	319
1961	439
1971	548

Solution:



Thus population for the year 1965 will be 472. This method is the simplest method of interpolation but suffers with the limitations that different smooth curves can be drawn through the point. Thus the value arrived at is not free from subjectively. In addition, the larger the volume of figures, the narrower the scale has to be in the graph paper and consequenctly greater will be the error of approximation. The method also will not be helpful for future projection.

10.5.2 Binomial expansion method:

This method is simple to understand and requires very little calculation. But it is applicable only in those situations where the following two conditions are satisfied:

1. It can be used only when the independent variable X advances by equal interval say 5, 10, 15, 20, 25 etc. If the increase is not uniform this method is not applicable for example if X is 5, 8, 13, 17, 24 this method cannot be applied.

2. The value of X for which Y is to be interpolated is one of the class limits of X series. For example if

X	5	10	15	20	25
Y	30	32	?	38	40

We can determine the values of Y corresponding to X = 15 but not when X is 12 and 19. The same is true for extrapolation i.e. we can extrapolate the values of X = 30 but X = 27.

When this method is applied the formula for the binomial theorem is expanded and equated with zero.

$$(Y-1)^n = Y^n - nY^{n-1} + \frac{n(n-1)}{2!}Y^{(n-2)} - \frac{n(n-1)(n-2)}{3!}Y^{(n-3)}$$

$$+ \frac{n(n-1)(n-2)(n-3)}{4!}Y^{(n-4)} = 0$$

or

$${}^nC_n Y^n - {}^nC_{n-1}Y^{n-1} + {}^nC_{n-2}Y^{n-2} + {}^nC_{n-3}Y^{n-3} \ldots\ldots {}^nC_0 Y^{n-0} = 0$$

where n is the number of known values of Y the expansion of the binomial for some values of n is given as

Number of known values of Y	Formula	Expansion
2	Δ_0^2 or $(Y-1)^2 = 0$	$Y_2 - 2Y_1 + Y_0 = 0$
3	Δ_0^3 or $(Y-1)^3 = 0$	$Y_3 - 3Y_2 + Y_1 - Y_0 = 0$
4	Δ_0^4 or $(Y-1)^4 = 0$	$Y_4 - 4Y_3 + 6Y_2 - 4Y_1 + Y_0 = 0$
5	Δ_0^5 or $(Y-1)^5 = 0$	$Y_5 - 5Y_4 + 10Y_3 - 10Y_2 + 5Y_1 - Y_0 = 0$
6	Δ_0^6 or $(Y-1)^6 = 0$	$Y_6 - 6Y_5 + 15Y_4 - 20Y_3 + 15Y_2 - 6Y_1 + Y_0 = 0$

The expansion of the binomial formula as shown appears to be a little difficult and complex. However this can be done by a simple procedure as follows:

1. The first subscript of Y will be the number equivalent of which we have to find the binomial expansion. Thus if $(Y-1)^4 = 0$ is to be expanded the first Y will be Y^4.

After that each Y's subscript will be reduced by 1 till it reaches Y_0 i.e., Y_4, Y_3, Y_2, Y_1 and Y_0.

2. The plus and minus signs are to be placed alternatively starting from first which is a plus

$$+Y_4, -Y_3, +Y_2, -Y_1, +Y_0$$

3. The numerical coefficient would be determined as follows:

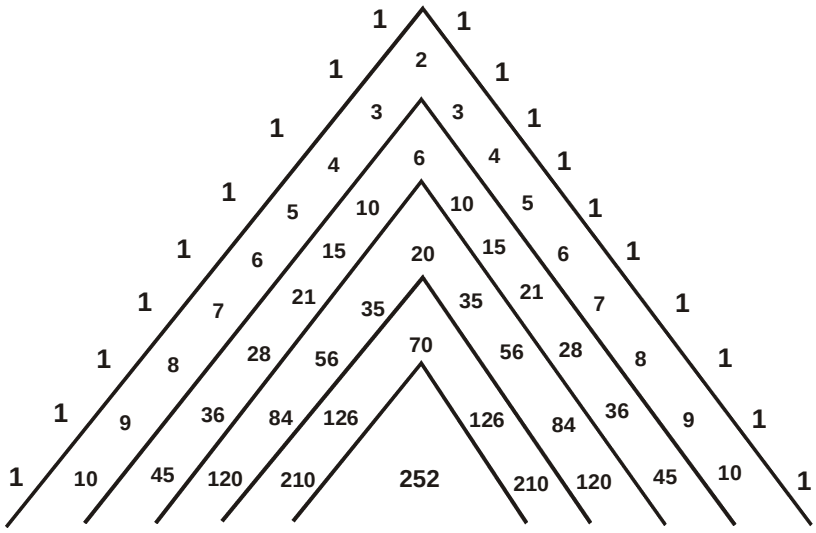
   (a) The first numerical coefficient will be always 1. Thus in this case the term becomes 1 Y_4 but the value 1 is not written, so it is Y_4.

   (b) The second coefficient would be the product of numerical coefficient of first value and its subscript divided by one. Thus in the present illustration it would be $1 \times 4/1 = 4$ so the term would be $4\,Y_4$.

   (c) The third coefficient would be the product of the numerical coefficient of second value and its subscript divided by 2. Thus in our illustration it would be $4 \times 3/2 = 6$ and the third term would be $6\,Y_3$.

   (d) Likewise 4th coefficient would be $6 \times 2/3 = 4$ the 4th term would be $4Y_4$ and so on. The coefficeints can be found out by referring Pascal's Triangle given below:

Pascal's Triangle



Example 1: Find the missing figure in the following data:

Year	1970	1975	1980	1985	1990	1995
Sales of Umbrella	320	300	?	280	278	250

Solution:

	X_0	X_1	X_2	X_3	X_4	X_5
Year	1970	1975	1980	1985	1990	1995
Sales of Umbrella	320	300	?	280	278	250
	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5

Since the known value of Y are 5 therefore $\Delta_0^5 = 0$ or $(Y-1)^5 = 0$

or

$$Y_5 - 5Y_4 + 10Y_3 - 10Y_2 + 5Y_1 - Y_0 = 0$$

Substituting the values of Y from the given data we get

$$250 - (5 \times 278) + 10(280) - 10(Y_2) + 5(300) - 1(320) = 0$$

or $$250 - 1390 + 2800 - 10Y_2 + 1500 - 320 = 0$$

$$10\ Y_2 = 250 - 1390 + 2800 + 1500 - 320 = 2840$$

or $$Y_2 = 284$$

when there are two or more missing values

We get two unknown quantities in the equation obtained by the binomial expansion. In such case if we are given n values we assume that (n - 1)th difference are constant i.e., we assume

$\Delta^{n-1}\ \Delta^{n-2}\ \Delta^{n-3}$ ..... are constant.

If (n - 1)th difference is constant n^{th} difference is zero.

i.e., $\Delta_{Y_1}^n = 0, \Delta_{Y_2}^n = 0$ and so on.

The following example will illustrate the procedure.

Example 2:

From the following data of profits of a firm (in lakh rupees) interpolate the missing figure.

Year	1965	1970	1975	1980	1985	1990	1995
Profit (in lakhs)	20	22	26	?	36	?	43
	Y_0	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6

Solution:

As five figures are known we shall assume that the fifth order difference will be zero. In the problem there are two unknown figures hence two equations will be required to determine them

$$\Delta_0^5 = Y_5 - 5Y_4 + 10Y_3 - 5Y_2 + Y_0 = 0 \text{ and}$$

$$\Delta_1^5 = Y_6 - 5Y_5 + 10Y_4 - 10Y_3 + 5Y_2 + Y_1 = 0$$

Substituting the values

$$Y_5 - 5(35) + 0Y_2 - 10(26) + 5(22) - 20 = 0 \cdots\cdots\cdots(1)$$

$$43 - 5Y_5 + 10(36) - 10(Y_3) + 2(26) - 220 = 0 \cdots\cdots\cdots(2)$$

or

$Y_5 + 10\,Y_2 = 345$ substituting equation (1) from (2) we get

$$\frac{5Y_5 - 10Y_2 = 501}{4Y_5 - 156} \text{ or } Y_5 = 156/4 = 39$$

Substituting the values of Y_2 in the above equation

$$39 + 10Y_2 = 345$$

$$10Y_2 = 345 - 39 \text{ or } Y_2 = 306/10 = 30 \cdot 6$$

Thus missing values corresponding to 1980 and 1990 are 30.6 and 39 respectively.

10.5.3. Newton's Method:

The Newton's Method can be classified into four different heads:

1. Newton's Advancing Difference Method
2. Newton's Gauss (forward) Method
3. Newton's Gauss (backward) Method
4. Newton's Divided Difference Method

1. Newton's Advancing Difference Method:

The advancing difference method is applicable where the independent variable X advances by equal interval like 10, 20, 30, 40 etc., However unlike binomial method it is not necessary that the value of X to be interpolated should be one of the class intervals. for example value of X = 25 or 33 can be interpolated. Similarly, we can extrapolate X = 59.

The formulafor interpolation is

$$Y_x = Y_0 \times \Delta_0^1 + \frac{x(x-1)}{2!}\Delta_0^2 + \frac{x(x-1)(x-2)}{3!}\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{4!}\Delta_0^4 + \cdots\cdots$$

where Y_0 represents the value of Y at origin.

Y_x represents the figure to be interpolated and Δ_0^1, Δ_0^2, Δ_0^3 are the first, second and third leading differences of each column to various degree of difference. The value of x is obtained as follows:

$$x = \frac{\text{The value be interpolated} - \text{The value at origin}}{\text{Difference between two adjoinging values}}$$

When applying this method the difference between the various value of Y are to be calculated. The difference are indicated by the sign Δ. Thus the first difference would be Δ^1 second Δ^2 and Δ^3 likewise. The following is the table of differences.

Table of Difference

X	Y	$1^{st}\ \Delta$ Δ^1		$2^{nd}\ \Delta$ Δ^2		$3^{rd}\ \Delta$ Δ^3		$4^{th}\ \Delta$ Δ^4	
X_0	Y_0	$Y_1 - Y_0$	Δ_0^1	$\Delta_1^1 - \Delta_0^1$	Δ_0^2	$\Delta_1^2 - \Delta_0^2$	Δ_0^3	$\Delta_1^3 - \Delta_0^3$	Δ_0^4
X_1	Y_1	$Y_2 - Y_1$	Δ_1^1	$\Delta_2^1 - \Delta_1^1$	Δ_1^2	$\Delta_2^2 - \Delta_1^2$	Δ_1^3		
X_2	Y_2	$Y_3 - Y_2$	Δ_2^1	$\Delta_3^1 - \Delta_2^1$	Δ_2^2				
X_3	Y_3	$Y_4 - Y_3$	Δ_3^1						
X_4	Y_4								

Example 3:

Given the following annual premium charged by LIC of India for a policy of Rs. 1000. Calculate the premium payable at the age 26.

Age in Years	20	25	30	35	40
Premium (Rs.)	23	27	32	39	48

Solution:

Applying Newton's Method

Age X	**Premium Y**		**1st Diff.**		**2nd Diff.**		**3rd Diff.**		**4th Diff**	
20	23	Y_0	+ 4	Δ_0^1	+ 1	Δ_0^2				
25	27	Y_1	+ 5	Δ_1^1	+ 2	Δ_1^2	1	Δ_0^3		
									- 1	Δ_0^4
30	32	Y_2	+ 7	Δ_2^1	+ 2	Δ_2^2	0	Δ_1^3		
35	39	Y_3	+ 9	Δ_3^1						
40	48	Y_4								

$$Y_x = Y_0 \text{ x } \Delta_0^1 + \frac{x(x-1)}{2!}\Delta_0^2 + \frac{x(x-1)(x-2)}{3!}\Delta_0^3 + \frac{x(x-1)(x-2)(x-3)}{4!}\Delta_0^4$$

where $x = \frac{26-20}{5} = 1\cdot 2$

$$Y_{26} = 23 + 1\cdot 2 \times 4 + \frac{(1\cdot 2)(1\cdot 2-1)}{1\times 2} \times 1 + \frac{(1\cdot 2)(1\cdot 2-1)(1\cdot 2-2)}{3\times 2}\times 1$$

$$+ \frac{(1\cdot 2)(1\cdot 2-1)(1\cdot 2-2)(1\cdot 2-3)}{4\times 3\times 2} \times -1$$

= 23 + 4.8 + 0.12 + (- 0.032) + 0.0144

= 27.90

Thus the premium payable at the age 26 is Rs. 27.90

2. Newton's Gauss (forward) Method:

This method of interpolation is applicable when

(a) The independent variable x advances by equal intervals and

(b) The figure to be interpolated is in the middle of the series.

In this method the value preceding the value to be interpolated is denoted by X_0 and previous values are denoted by X_{-1}, X_{-2}, X_{-3} and value after X_0 are denoted by $X_1, X_2, X_3 \cdots\cdots$ similarly difference are denoted according to Y's rotation

$$\Delta^1 Y0, \Delta^2 Y-1, \Delta^3 Y-1, \Delta^4 Y-3 \cdots\cdots$$

The formula for interpolation is

$$Y_x = Y_0 + x\Delta^1 Y0 + \frac{x(x-1)}{1\times 2}\Delta^2 Y-1 + \frac{x(x-1)(x-2)}{1\times 2\times 3}\Delta^3 Y-1$$

$$+ \frac{x(x-1)(x-2)(x-3)}{1\times 2\times 3\times 4}\Delta^4 Y-2 + \cdots\cdots$$

where $x = \dfrac{\text{Item to be interpolated} - \text{Preceding item}}{\text{Difference between adjusting items}}$

Example 4:

The following are the annual sales of a concern for the last few years. Estimate sales for the year 1993.

Year	1988	1990	1992	1994	1996
Sales in '000 Rs.	23	26	30	35	42

Solution:

Table of Difference

X **Year**	**Y** **Sales**		$1^{st}\ \Delta$ Δ^1		$2^{nd}\ \Delta$ Δ^2		$3^{rd}\ \Delta$ Δ^3		$4^{th}\ \Delta$ Δ^4	
1988	23	Y_2	+ 3	Δ^1_{-2}	+ 1	Δ^2_{-2}	0	Δ^3_{-2}	+ 1	Δ^4_{-2}
1990	26	Y_1	+ 4	Δ^1_{-1}	+ 1	Δ^2_{-1}	1	Δ^3_{-1}		
1992	30	Y_0	+ 5	Δ^1_0	+ 2	Δ^2_0				
1994	35	Y_1	+ 7	Δ^1_1						
1996	42	Y_2								

The value of $x = \frac{1993-1992}{1994-1992} = \frac{1}{2} = 0\cdot5$

Substituting the values in Newton's Gauss forward formula we get

$$Y_x = 30 + (0\cdot5 \times 5) + \frac{(0\cdot5)(0\cdot5-1)}{1\times2} \times 1 + \frac{(0\cdot5)(0\cdot5-1)(0\cdot5-2)}{1\times2\times3} \times 1$$

$$+ \frac{(0\cdot5)(0\cdot5-1)(0\cdot5-2)(0\cdot5-3)}{1\times2\times3\times4} \times 1$$

= 30 + 2.5 + (- 0.125) + 0.0625 + (- 0.0391)

= 32.3984

3. Newton's Gauss (backward) Method:

This method is applicable when

(a) X series advances by equal intervals

(b) When the figures to be interpolated is near the end of the series

The formula is

$$Y_x = Y_0 - X\Delta^1 Y - 1\frac{(x+1)X}{1\times2}\Delta^2 Y - 1 + \frac{(x+1)X(x-1)}{1\times2\times3}\Delta^3 Y - 2$$

$$+ \frac{(x+1)X(x-1)(x-2)}{1\times2\times3\times4}\Delta^4 Y - 2$$

where Y_0 is the figure succeeding the missing figure and

$$x = \frac{\text{Item succeeding the item to be interpolated} - \text{Item to be interpolated}}{\text{Difference between adjoining items}}$$

Example 5:

The following table gives some values of x and y variable. Inerpolate the values of y when x is 45.

X	10	20	30	40	50	60
Y	22	26	32	41	47	51

Solution:

The figure to be interpolated is towards the end of the table hence Newton's Gauss (backward) method is used.

Table of Difference

X		Y		$1^{st}\ \Delta$ Δ^1		$2^{nd}\ \Delta$ Δ^2		$3^{rd}\ \Delta$ Δ^3		$4^{th}\ \Delta$ Δ^4	
10	X - 4	22	Y - 4	+ 4	Δ^1_{-4}	+ 2	Δ^2_{-4}				
20	X - 3	26	Y - 3	+ 6	Δ^1_{-3}	+ 3	Δ^2_{-3}				
								+ 1	Δ^3_{-4}		
30	X - 2	32	Y - 2	+ 9	Δ^1_{-2}	- 3	Δ^2_{-2}			+ 7	Δ^4_{-4}
								- 6	Δ^3_{-3}		
40	X - 1	41	Y - 1	+ 6	Δ^1_{-1}	- 1	Δ^2_{-1}			+ 8	Δ^4_{-3}
								+ 2	Δ^3_{-2}		
50	X - 0	51	Y - 1	+ 5	Δ^1_0						
60	X - 1										

$$Y_x = Y_0 - X\Delta^1 Y_{-1} + \frac{(x+1)X}{1\times 2}\Delta^2 Y_{-1} + \frac{(x+1)X(x-1)}{1\times 2\times 3}\Delta^3 Y_{-2}$$

$$+ \frac{(x+1)X(x-1)(x-2)}{1\times 2\times 3\times 4}\Delta^4 Y_{-2}$$

$$= 45 - 3 - 0\cdot 375 + 0\cdot 125$$

$$= 41.75$$

4. Newton's Divided Difference Method:

This method is to be used when the value of the independent variable x advances by unequal intervals. The formula is

$$Y_x = Y_0 + (X - X_0)\Delta^1 0 + (X - X_0)\Delta^2 0 + (X - X_0)(X - X_1)(X - X_2)\Delta^3 0 + \cdots\cdots$$

where $\Delta^1 0, \Delta^2 0, \Delta^3 0$ are first, second and third leading divided differences respectively. The divided difference are obtained in the manner as follows:

Divided Difference

X	Y	1st Δ Δ^1		2nd Δ Δ^2		3rd Δ Δ^3	
X_0	Y_0	$\frac{Y_1 - Y_0}{X_1 - X_2}$	Δ_0^1	$\frac{-\Delta_3^1 - \Delta_0^1}{X_2 - X_0}$	Δ_0^2	$\frac{-\Delta_1^2 - \Delta_0^2}{X_3 - X_0}$	Δ_0^3
X_1	Y_1	$\frac{Y_2 - Y_1}{X_2 - X_1}$	Δ_1^1	$\frac{-\Delta_2^1 - \Delta_1^1}{X_3 - X_1}$	Δ_1^2		
X_2	Y_2	$\frac{Y_3 - Y_2}{X_3 - X_2}$	Δ_2^1				
X_3	Y_3						

Example 6:

The observed values of a function are respectively 168, 120, 72, 63 at four positions 4, 9, 12, 16 of independent variable, what best estimate can you give for the value of the function at the position 6 of the independent variable?

Solution:

Since the independent variable is advancing by unequal interval Newton's divided difference method is used.

X		Y		1st Δ Δ^1		2nd Δ Δ^2		3rd Δ Δ^3	
4	X_0	168	Y_0	$\frac{120-168}{5}$	$-9\cdot6\ \Delta_0^1$	$\frac{-16-(-9\cdot6)}{12-44}$	$0\cdot8\ \Delta_0^2$	$\frac{-1\cdot96-(-8)}{16-4}$	= 0.23
9	X_1	120	Y_1	$\frac{72-120}{3}$	$-16\ \Delta_1^1$	$\frac{2\cdot256-(-16)}{16-9}$			
12	X_2	72	Y_2	$\frac{63-72}{4}$	$2\cdot25\ \Delta_2^1$				
16	X_3	63	Y_3						

$$Y_x = Y_0 + (X - X_0)\Delta^1 0 + (X - X_0)(X - X_1)\Delta^2 0 + (X - X_0)(X - X_1)(X - X_2)\Delta^3 0$$

$$= 168 + (6-4)(-9\cdot 6) + (6-4)(6-9)(6-12)(\cdot 23)$$

= 168 - 19.2 + 4.8 + 8.28

= 161.88

10.5.4 Langrang's Method:

Like Newton's divided difference method this formula is also used when x series does not advance by equal interval. This also can be used when x series advances by equal intervals. The Langrang's formula is

$$Y_x = Y_0 \frac{(X - X_1)(X - X_2)(X - X_3)(X - X_4)\cdots\cdots(X - X_n)}{(X_0 - X_1)(X_0 - X_2)(X_0 - X_3)(X_0 - X_4)\cdots(X_0 - X_n)}$$

$$+ Y_1 \frac{(X - X_0)(X - X_2)(X - X_3)(X - X_4)\cdots\cdots(X - X_n)}{(X_1 - X_0)(X_1 - X_2)(X_1 - X_3)(X_1 - X_4)\cdots\cdots(X_0 - X_n)}$$

$$+ Y_2 \frac{(X - X_0)(X - X_1)(X - X_2)\cdots\cdots(X - X_n)}{(X_2 - X_0)(X_2 - X_1)(X_2 - X_3)\cdots\cdots(X_2 - X_n)}$$

$$+ Y_n \frac{(X - X_0)(X - X_1)(X - X_2)\cdots\cdots(X - X_{n-1})}{(X_n - X_0)(X_n - X_1)(X_n - X_2)\cdots\cdots(X_n - X_{n-1})}$$

where Y_x is the figure to be interpolated, X is the value of x series for which Y_x is to be obtained. $X_0, X_1, X_2, X_3, \cdots\cdots\cdots, X_n$ are given values of X variable. The following example would illustrate the formula.

Example 7:

The following table gives the normal weight of a baby during the first six months of life.

Age in months	0 2	3	5	6
Weight in lbs	5 8	9	12	14

Estimate the weight of a baby at the age of 4 months.

Solution:

X		Y	
0	X_0	5	Y_0
2	X_1	8	Y_1
3	X_2	9	Y_2
5	X_3	12	Y_3
6	X_4	14	Y_4

By substituting the values in Lagrang's formula we get

$$\frac{(4-2)(4-3)(4-5)(4-6)}{(0-2)(0-3)(0-5)(0-6)} \times 5 + \frac{(4-0)(4-3)(4-5)(4-6)}{(2-0)(2-3)(2-5)(2-6)} \times 8$$

$$+ \frac{(4-0)(4-2)(4-5)(4-6)}{(3-0)(3-2)(3-5)(3-6)} \times 9 + \frac{(4-0)(4-2)(4-3)(4-6)}{(5-0)(5-2)(5-3)(5-6)} \times -12$$

$$+ \frac{(4-0)(4-2)(4-3)(4-5)}{(6-0)(6-2)(6-3)(6-5)} \times 14$$

$$\frac{2x1x-1x-2}{-2x-3x-5x-6} \times 5 + \frac{4x\,1x-1x-2}{2x-1x-3x-4} \times 8 + \frac{4x2x-1x-2}{3x1x-2x-3} \times 9$$

$$= \frac{4x2x\,1x-2}{5x\,3x\,2x\,1} \times 12 + \frac{4x\,2x\,1x-1}{6x\,4x\,3x\,1} \times 14$$

$$\frac{4}{180} \times 5 + \frac{8}{-24} \times 8 + \frac{16}{18} \times 9 + \frac{-16}{30} \times 12 + \frac{-8}{72} \times 14$$

= 0.11 + (- 2.67) + 8 + 6.4 - 1.55

= 10.29

Thus the weight os the baby at the age of 4 months will be 10.29 pounds.

10.5.5 Parabolic Curve Method:

This method of interpolation is also known as method of simulataneous equation. This method is based on the assumption that the values of X and Y are interdependent and the values of X are known. The variable Y is taken as dependent variable and X as independent. Consequenctly for a given X we can find out the value of Y. The equation of this curve is

$$Y = a + bx + cx^2 + dx^3 + ex^4 \cdots\cdots\cdots\cdots kx^n$$

This equation represents a parabolic curve of nth degree and a, b, c, d etc are the constants. The power to which this equation is to be raised depends upon the number of known quantities. The curve is raised to the power 'one' less than the number of known quantities. For example if known quantities are 4 we would take a curve of 3rd order.

$$Y = a + bx + cx^2 + dx^3$$

Example 8:

Estimate the profits for the year 1993 from the following data:

Year	1991	1992	1994
Profit	8.5	12	10

Solution:

Value of X by taking deviation from 1993 and the value of Y are

Year	1991	1992	1993	1994
X	- 2	- 1	0	1
Profit	8.5	12	Y_0	10
Y				

Since the known values are 3 we raise a parabola of second order.

$$Y = a + bx + cx^2$$

Substituting X and Y we get

$$8 \cdot 5 = a - 2b + 4c \cdots\cdots\cdots\cdots (1)$$

$$12 = a - b + c \cdots\cdots\cdots\cdots (2)$$

$$Y = a \cdots\cdots\cdots\cdots (3)$$

$$10 = a + b + c \cdots\cdots\cdots\cdots (4)$$

Adding (2) and (4) we get

$$22 = 2a + 2c \cdots\cdots\cdots\cdots (5)$$

Multiplying equation (4) by 2 and adding it with equation (1)

$$20 = 2a + 2b + 2c$$
$$8 \cdot 5 = a - 2b + 4c$$

$$28 \cdot 5 = 3a + 6c \quad \ldots\ldots\ldots\ldots (6)$$

Multiplying equation (5) by (3) and subtracting equation (6) from it we get

$$66 = 6a + 6c$$
$$28.5 = 3a + 6c$$
$$37.5 = 39 \quad \text{or} \quad a = 12.5$$

Thus profit for the year 1993 would be 12.5 lakhas.

10.5.6 Extrapolation:

As pointed out earlier extrapolation refers to estimating a value beyond the given values of Y. It is a futuristic estimation there is no specific formula for extrapolation can be adopted for extrapolation. The choice of method could depend on the nature of data given. The unditions applicable with different formula for interpolation hold good for extra polation also.

Thus if we are given population figure if india for 1941, 49, 51, 61, 71, 81 and we have to extrapolate 1991 we will use binomial expansion method. If the population figure has to be extrapolated for 86 we can use newton's methods of advancing difference.

10.6 Summary:

In this lesson, we have shown the concept of interpolation and Extropolation. We have explained the methods of intropolation and extropolation such as Binomial expansion method, Newton's method, lagrangl's method parabolic cur ve method. These methods are very important in practical series problems.

10.7 Exercises:

1. Explain briefly the usefulness of interpolation and extrapolation in statistical studies.
2. Give the meaning of Interpolation and Extrapolation. Mention the assumption underlying Interpolation and Extrapolation.
3. State Newton's formula for interpolation and discuss some its uses.
4. Explain Lagrange's method of Interpolation. Point out its usefulness.
5. Using the binomial expansion method estimate the missing figure for 2006 - 07 from the data given below:

Year	2002 - 03	2003 - 04	2004 - 05	2005 - 06	2006 - 07	2007 - 08
Average Value	89.56	52.54	37.36	27.85	?	23.36

(Rs. Crores)

6.

Extrapolate the business done in 2009 from the following data:					
Year:	2004	2005	2006	2007	2008
Business done	150	235	365	525	780

(Rs. Lakhs)

10.8 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 11

TEST OF HYPOTHESIS

Objective:

After going through this lesson you will learn:

- The concept of test of hypothesis and errors in sampling tests
- The computation of one tailed and two tailed tests and test of significance for large samples.

Structure:

11.1 Introduction:

There are many problems, in which, rather than estimating the value of a parameter we need to decide whether to accept or reject a statement about the parameter. This statement is called a hypothesis. This is one of the most useful aspects of statistical inference. Since many types of decision - making problems, tests or experiments in the practical problems can be formulated as hypothesis - testing problems.

11.2 Definition:

A statistical hypothesis is a statement about the parameter of one or more populations.

Since we use probability distributions to represent population, a statistical hypothesis may also be thought of as a statement about the probability distribution of a random variable. The hypothesis will usually involve one or more parameters of this distribution.

Example:

1. The majority of men in the city are smokers
2. The teaching methods in both the schools are effective

There are two types of hypothesis.

(1) Null hypothesis

(2) Alternative hypothesis

1. Null Hypothesis:

For applying the tests of significance, we first set up a hypothesis - a definite statement about the population parameter. Such a hypothesis is usually a hypothesis of no - difference, is called Null Hypothesis, and it is denoted by H_0.

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics (H_0) will be that the sample statistics do not differ significantly.

2. Alternative Hypothesis:

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1. For example, if we want to test the null hypothesis that the population has a specified mean μ_0 (say) i.e., $H_0 = \mu = \mu_0$ then the alternative hypothesis would be

(i) $H_1 = \mu \neq \mu_0 \quad (\text{i.e, either } \mu > \mu_0 \text{ or } \mu < \mu_0)$

(ii) $H_1 : \mu > \mu_0$

(iii) $H_1 : \mu < \mu_0$

The alternative hypothesis (i) is known as a two tailed alternative and the alternative in (ii) is known as right tailed and in (iii) is known as left tailed.

The setting of alternative hypothesis is very important to decide whether we have to use a single tailed (right or left) or two tailed test.

11.3 Errors In Sampling:

The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or to reject the lot after examining a sample form it. As such we have two types of errors.

(i) Type I error : Reject H_0 when it is true.

(ii) Type II error : Accept H_0 when it is wrong i.e., accept H_0 when H_1 is true.

If we write

P (Reject H_0 when it is true) = P (Type I error) = α

and P (Accept H_0 when it is wrong) = P (Type II error) = β

then α and β are called sizes of Type I and Type II errors respectively

i.e., $\alpha = P$ (Rejecting a god lot)

$\beta = P$ (Accepting a bad lot)

The sizes of Type I and Type II errors are also known as producer's risk and consumer's risk respectively.

1. Critical Region:

A region corresponding to a statistic 't', in the sample space S which leads to the rejection of H_0 is called Critical Region or Rejection Region. Those region which lead to the acceptance of H_0 gives us a region called Acceptance Region.

2. Critical values or Significant Values:

The value of the test statistic which separates the critical region (or rejection region) and the acceptance region is called the critical value or significant value. This value is dependent on

(i) The level of significance used, and

(ii) The alternative hypothesis whether it is one tailed or two tailed

For larger samples, corresponding to the statistic t, the variable $Z = \frac{t - E(t)}{S \cdot E(t)}$

$E(t) = ??$ is normally distributed with mean 0 and variance 1.

The value of Z gives above under the null hypothesis is known as test statistic. The critical value Z_α of the test statistic at level of significance α for a two - tailed test is given by

$$P\left(|Z| > Z_\alpha\right) = \alpha \quad \cdots\cdots\cdots\cdots(1)$$

That is, Z_α is the value of Z so that the total area of the critical region on both rails is α. Since the normal curve is a symmetrical one, equation (1) implies,

$$P\left(Z > Z_\alpha\right) + P\left(Z < -Z_\alpha\right) = \alpha$$

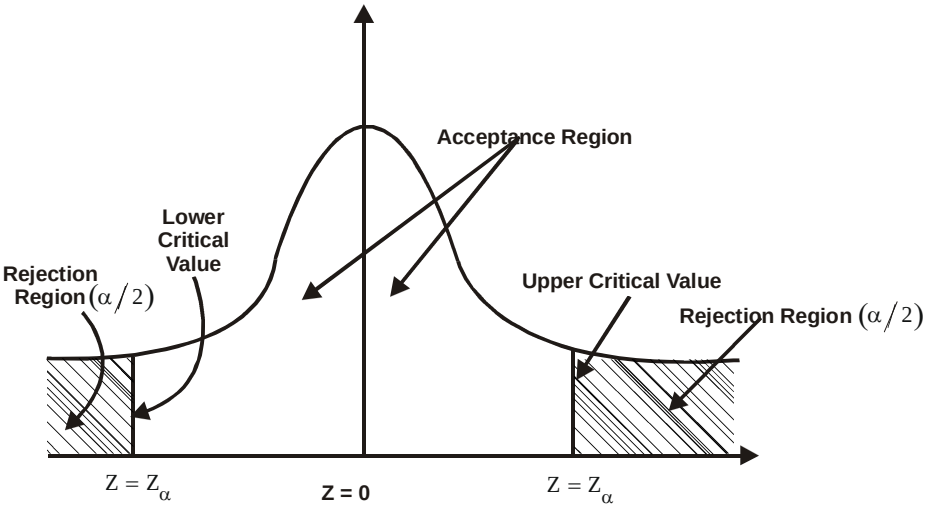
i.e., $2P\left(Z > Z_\alpha\right) = \alpha$

or $P\left(Z > Z_\alpha\right) = \alpha/2$

That is the area of each tail is $\alpha/2$.

The critical value Z_α is that value such that the area to the right of Z_α is $\alpha/2$ and the area to the left of $-Z_\alpha$ is $\alpha/2$. Refer figure.

3. Two - tailed test at level of significance 'α':



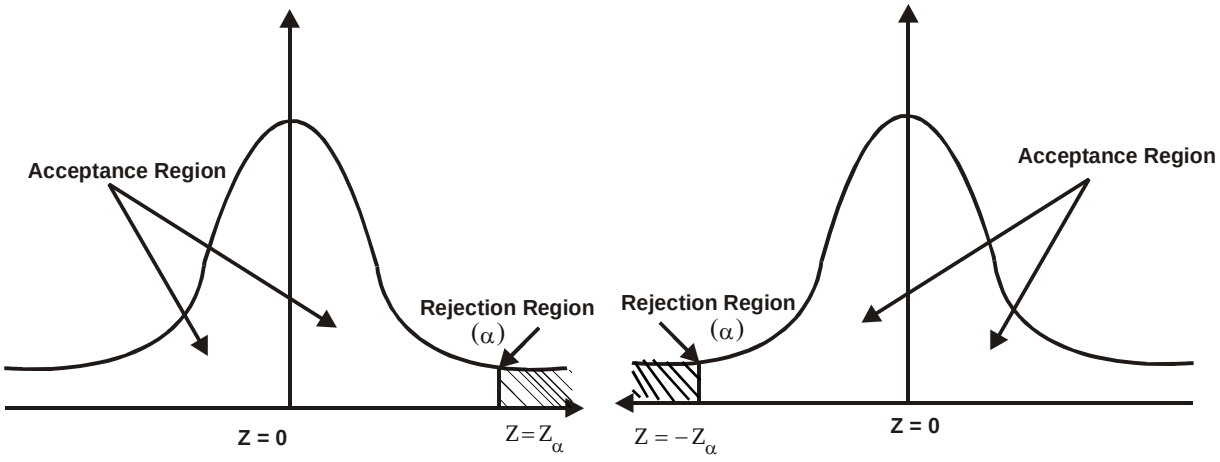
In the case of one - tailed alternative,

$P(Z > Z_\alpha) = \alpha$ if it is one - tailed (right)

$P(Z < Z_\alpha) = \alpha$ if it is one - tailed (left)

For Level of Significance 'α'

Right Tailed Test **Left Tailed Test**



From the above figures, it is clear that the critical value of Z for a single - tailed test (right or left) at level of significance 'α' is same as the critical value of Z for two-tailed test at level of significance '2α'.

The critical values of Z at different level of significance (α) for both single tailed and two tailed tests are calculated from equations.

$$P\left(|Z| > Z_\alpha\right) = \alpha$$

$$P\left(Z > Z_\alpha\right) = \alpha$$

$$P\left(Z < -Z_\alpha\right) = \alpha$$

Using the normal tables. They are listed below:

Critical values $\left(Z_\alpha\right)$ of Z :

	Level of Significance		
	1% (.01)	**5% (.05)**	**10% (.1)**
Two - tailed test	$\left\|Z_\alpha\right\| = 2\cdot58$	$\left\|Z_\alpha\right\| = 1\cdot96$	$\left\|Z_\alpha\right\| = 1\cdot645$
Right tailed test	$Z_\alpha = 2\cdot33$	$Z_\alpha = 1\cdot645$	$Z_\alpha = 1\cdot28$
Left tailed test	$Z_\alpha = -2\cdot33$	$Z_\alpha = -1\cdot645$	$Z_\alpha = -1\cdot28$

11.4 One Tailed and Two - Tailed Tests:

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called a one tailed test.

For example, in a test for testing the mean of a population in a single tailed test we assume that the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis

$H_0 : \mu > \mu_0$ (Right Tailed)

or $H_0 : \mu < \mu_0$ (Left Tailed)

is called one tailed test.

In a test of statistical hypothesis where the alternative hypothesis is two tailed, we assume that the null hypothesis

$H_0 : \mu = \mu_0$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_0 \left[\mu > \mu_0 \text{ or } \mu < \mu_0\right]$$

Applying one tailed or two tailed test for a particular problem depends entirely on the nature of the alternative hypothesis. If the alternative test is two - tailed we apply two - tailed test and if alternative hypothesis is one - tailed we apply one - tailed test.

Example:

Consider two population brands of bulbs one manufactured by routine process (mean μ_1) and the other manufactured by new technique (mean μ_2). If we want to test if the bulbs differ significantly then the hypothesis is $H_0 : \mu_1 = \mu_2$ and the alternative hypothesis will be $H_0 : \mu_1 \neq \mu_2$. This gives us a two - tailed test. Suppose if we want to test if the bulbs produced by new process (μ_2) have higher average life than those produced by standard process (μ_1), then we have

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 < \mu_2$$

in this case we have to adopt a left tail test.

If we want to test whether the product of new process (μ_2) is inferior to that of standard process (μ_1) then we have

$$H_0 : \mu_1 = \mu_2 \text{ and } H_1 : \mu_1 > \mu_2$$

which gives a right - tail test.

Hence the decision about applying a two - tail test or a single - tail (left or right) test will mainly depend on the problem under study.

Working Rule for Testing of Hypothesis:

The following working rule or procedure may be adopted in testing of a statistical hypothesis.

Step 1: Null Hypothesis: Define or set up a null hypothesis H_0 is clear terms.

Step 2: Alternative hypothesis: Setup the Alternative Hypothesis H_1 so that we could decide whether we should use one - tailed or two tailed test.

Step 3: Level of Significance: Select the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. That is a suitable α is selected in advance.

Step 4: Test Statistic: Compute the test statistic $Z = \dfrac{t - E(t)}{S \cdot E(t)}$ under the null hypothesis.

Step 5: Conclusion: We compare the computed value of the test statistic Z with the critical value Z_α at given level of significance (α).

If $|Z| < Z_\alpha$ (that is, if the absolute value of the calculated value of Z is less than the critical value Z_α) we conclude that it is not significant. We accept the null hypothesis.

If $|Z| > Z_\alpha$ then the difference is significant and hence the null hypothesis is rejected at the level of significance α.

Clearly,

For two - tailed test:

If $|Z| < 1{\cdot}96$ accept H_0 at 5% level of significance.

If $|Z| > 1{\cdot}96$ reject H_0 at 5% level of significance.

If $|Z| < 2{\cdot}58$, accept H_0 at 1% level of significance.

If $|z| > 2{\cdot}58$ reject H_0 at 1% level of significance.

For single - tailed (right or left) test:

If $|Z| < 1{\cdot}645$ accept H_0 at 5% level of significance.

If $|Z| > 1{\cdot}645$ reject H_0 at 5% level of significance.

If $|Z| < 2{\cdot}33$, accept H_0 at 1% level of significance.

If $|z| > 2{\cdot}33$ reject H_0 at 1% level of significance.

11.5 Test of Significance For Large Samples:

If the sample size n is greater than 30, we usually take sample as large sample. If n is large, the distributions such, as Binomial, Posson, Chi-square etc. Are closely approximated by normal distributions. Therefore, for large samples, we apply normal test assuming the population as normal.

Under large sample tests, we will see four important tests to test the significance.

1. Testing of significance for single proportion
2. Testing of significance for difference of proportions
3. Testing of significance for single mean
4. Testing of significance for difference of means

1. Test for significance for single proportion:

Suppose a large sample of size n is taken from a normal population. To test the significant difference between the sample proportion p and the population proportion P, we use the statistic

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$ where n is the sample size

Note:

Limits for population proportion P are given by $p \pm 3\sqrt{\frac{pq}{n}}$ where q = 1 - p.

Example 1:

A manufacturer claimed that atleast 95% of the equipment which he supplied to a factory conformed to specifications. An examination of a sample of 200 pieces of equipment revealed that 18 were faulty. Test his claim at 5% level of significance.

Solution:

Given sample size n = 200

Number of pieces confirming to specification = 200 - 18 = 182

$\therefore$ p = Proportion of pieces confirming to specifications

$$= \frac{182}{200} = 0 \cdot 91$$

P = Population proportion $= \frac{95}{100} = 0 \cdot 95$

Null Hypothesis H_0 : The proportion of pieces confirming to specifications

i.e. P = 95%

Alternative Hypothesis H_1 : P < 0.95 (left - tail test)

The test statistic $$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0 \cdot 91 - 0 \cdot 95}{\sqrt{\frac{0 \cdot 95 \times 0 \cdot 05}{200}}} = \frac{-0 \cdot 04}{0 \cdot 0154} = -2 \cdot 59$$

Since alternative hypothesis is left tailed, the tabulated value of Z at 5% level of significance is 1.645.

Since calculated value of $|Z| = 2 \cdot 6$ is greater than 1.645, we reject the null hypothesis H_0 at 5% level of significance. Hence the manufacture's claim is rejected.

Example 2:

In a sample of 1000 people in Karnataka 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular in this state at 1% level of significance?

Solution:

Given $n = 1000$

p = sample proportion of rice eaters $= \frac{540}{1000} = 0 \cdot 54$

P = population proportion of rice eaters $= \frac{1}{2} = 0 \cdot 5$

$\therefore Q = 0 \cdot 5$

Null Hypothesis H_0**:** Both rice and wheat are equally popular in the state.

Alternative Hypothesis H_1 **:** $P \neq 0 \cdot 5$ (two - tailed alternative)

Test statistic is $Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}} = \frac{0 \cdot 54 - 0 \cdot 5}{\sqrt{\frac{0 \cdot 5 \times 0 \cdot 5}{1000}}} = 2 \cdot 532$

The calculated value of Z = 2.532

The tabulated value of Z at 1% level of significance for two - tailed test is 2.58

Since calculated Z < tabulated Z we accept H_0. i.e. both rice and wheat are equally popular in the state at 1% level of significance.

Example 3:

In a big city 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers?

Solution:

Given n = 600

Number of smokers = 325

p = sample proportion of smokers $= \frac{325}{600} = 0 \cdot 5417$

P = Population proportion of smokers in the city $= \frac{1}{2} = 0 \cdot 5$

$Q = 1 - P = 1 - 0 \cdot 5 = 0 \cdot 5$

Null Hypothesis H_0 : The number of smokers and non - smokers are equal in the city.

Alternative Hypothesis: P > 0.5 (Right tailed)

Test statistic $Z = \dfrac{p - P}{\sqrt{\dfrac{PQ}{n}}} = \dfrac{0 \cdot 5417 - 0 \cdot 5}{\sqrt{\dfrac{0 \cdot 5 \times 0 \cdot 5}{600}}} = 2 \cdot 04$

Confidence interval at 99% level of significance is

$$\left(P - 3\sqrt{\frac{PQ}{n}} \ , \ P + 3\sqrt{\frac{PQ}{n}} \right)$$

i.e, $\left(0 \cdot 15 - 3 \times 0 \cdot 028,\ 0 \cdot 15 + 3 \times 0 \cdot 028\right)$

i.e., $\left(0 \cdot 065,\ 0 \cdot 234\right)$

2. Testing of significance for difference of proportions:

Suppose two large samples of sizes n_1 and n_2 are taken respectively from two different populations. To test the significant difference between the sample proportions.

$$p_1 \text{ and } p_2 \quad \text{find} \quad Z = \frac{p_1 - p_2}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

where $p = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ and q = 1 - p

Example 4:

A manufacture of electronic equipment subjects samples of two completing brands of transistors to an accelerated performance test. If 45 of 180 transistors of the first kind and 34 of 120 transistors of the second kind fail the test, what can he conclude at the level of significance $\alpha = 0 \cdot 05$ about the difference between the corresponding sample proportions?

Solution:

We have

$$n_1 = 180,\ x_1 = 45,\ x_2 = 34$$

and $p_1 = \dfrac{x_1}{n_1} = \dfrac{45}{180} = 0 \cdot 25, \quad p_2 = \dfrac{x_2}{n_2} = \dfrac{34}{120} = 0 \cdot 283$

$$p = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{45 + 34}{180 + 120} = \frac{79}{300} = 0 \cdot 263$$

$$q = 1 - p = 1 - 0 \cdot 263 = 0 \cdot 737$$

1. **Nill Hypothesis** H_0**:** $p_1 = p_2$ i.e. there is no difference

2. **Alternative Hypothesis** H_1**:** $p_1 \neq p_2$ i.e., there is a difference

3. **Level of significance** = 0.05

4. The test statistic is $$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0 \cdot 25 - 0 \cdot 283}{\sqrt{(0 \cdot 263)(0 \cdot 737)\left(\frac{1}{180} + \frac{1}{120}\right)}}$$

$$= \frac{-0 \cdot 033}{\sqrt{(0 \cdot 194)(0 \cdot 0134)}} = \frac{-0 \cdot 033}{0 \cdot 051}$$

$$= - 0.647$$

$$|Z| = 0 \cdot 647$$

Since $|Z| < 1 \cdot 96$ we accept the null hypothesis H_0 at 5% level of significance i.e., the difference between the proportion is not significant.

Example 5:

A cigarette manufacturing firm claims that its brand A line of cigarettes outsells its brand B by 8%. If it is found that 42 out of a sample of 200 smokers prefer brand A and 18 out of another sample of 100 smokers prefer brand B, test whether the 8% difference is a valid claim.

Solution:

Given $n_1 = 200,\ n_2 = 100$

$$p_1 = \frac{42}{200} = 0 \cdot 21,\ p_2 = \frac{18}{100} = 0 \cdot 18$$

and $$P_1 - p_2 = 8\% = \frac{8}{100} = 0 \cdot 08$$

1. Null Hypothesis H_0 :

Assume that 8% difference in the sale of two brands of cigarettes is a valid claim i.e.,

$$H_0 : p_1 - p_2 = 0 \cdot 08$$

2. Alternative Hypothesis H_1 : $p_1 - p_2 \neq 0 \cdot 08$ (two tail test)

3. The test statistic is, $$Z = \frac{(p_1 - p_2)-(P_1 - P_2)}{\sqrt{pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Now $$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{200(0 \cdot 21) + 100(0 \cdot 18)}{200 + 100} = \frac{42 + 18}{300} = 0 \cdot 2$$

and $q = 1 - 0 \cdot 2 = 0 \cdot 8$

$$\therefore \quad Z = \frac{0 \cdot 03 - 0 \cdot 08}{\sqrt{0 \cdot 2 \times 0 \cdot 8\left(\frac{1}{200} + \frac{1}{100}\right)}} = \frac{-0 \cdot 05}{0 \cdot 0489} = -1 \cdot 02$$

$$\therefore |Z| = 1 \cdot 02$$

Since $|Z| < 1 \cdot 96$ we accept the null hypothesis H_0 at 5% level of significance.

i.e., 8% difference in the sale of two brands of cigarettes is a valid claim.

Example 6:

In two large populations, there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two population.

Solution:

Given $n_1 = 1200$, $n_2 = 900$

P_1 = Proportion of fair haired people in the first population $= \frac{30}{100} = 0 \cdot 3$

P_2 = Proportion of fair haired people in the second population $= \frac{25}{100} = 0 \cdot 25$

1. Null Hypothesis H_0 :

Assume that the sample proportions are equal i.e., the difference in population proportions likely to be hidden in sampling i.e.,

$$H_0 : p_1 - p_2$$

2. Alternative Hypothesis H_1 : $p_1 \neq p_2$

3. The test statistic is $Z = \dfrac{P_1 - P_2}{\sqrt{\dfrac{P_1Q_1}{n_1} + \dfrac{P_2Q_2}{n_2}}}$

where $Q_1 = 1 - P_1 = 1 - 0 \cdot 3 = 0 \cdot 7$

$$Q_2 = 1 - P_2 = 1 - 0 \cdot 25 = 0 \cdot 75$$

$$Z = \frac{0 \cdot 3 - 0 \cdot 25}{\sqrt{\dfrac{0 \cdot 3 \times 0 \cdot 7}{1200} + \dfrac{0 \cdot 25 \times 0 \cdot 75}{900}}} = \frac{0 \cdot 05}{0 \cdot 0195} = 2 \cdot 55$$

i.e., Z = 2.5

i.e, Z > 1.96

$\therefore$ we reject the null hypothesis H_0 at 5% level of significance (Two - tailed test)

i.e., the sample proportions are not equal.

i.e., The difference in population proportions is unlikely to be hidden in sampling.

3. Test of Significance for single mean:

Suppose we want to test whether the given sample of size n has been drawn from a population with mean μ. We setup null hypotheses is that there is no difference between $\bar{x}$ and μ where $\bar{x}$ is the sample mean.

The test statistic is, $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

where σ is the S.D. of the population

If the population S.D is not known, then use the statistic

$Z = \dfrac{\bar{x} - \mu}{S / \sqrt{n}}$ where S is the sample S.D

Note:

The values $\bar{x} \pm 1{\cdot}96 \frac{\sigma}{\sqrt{n}}$ are called 95% fiducial limits or confidence limits for the mean of the population corresponding to the given sample.

Similarly $\bar{x} \pm 2{\cdot}58 \frac{\sigma}{\sqrt{n}}$ are called 99% confidence limits.

Example 7:

An oceanographer wants to check whether the depth of the ocean in a certain region is 57.4 fathoms, as had previously been recorded. What can he concluded at the level of significance $\alpha = 0{\cdot}05$, if readings taken at 40 random locations in the given region yielded a mean of 59.1 fathrooms with a standard deviation of 5.2 fathoms.

Solution:

Given $n = 40, \bar{x} = 59{\cdot}1$ and $\sigma = 5{\cdot}2$

1. **Null Hypothesis** H_0**:** $\mu = 57{\cdot}4$
2. **Alternative Hypothesis** H_1**:** $\mu \neq 57{\cdot}4$
3. Level of significance = 0.05

The test statistic is $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{59{\cdot}1 - 57{\cdot}4}{5{\cdot}2/\sqrt{40}} = 2{\cdot}067$

Tabulated value of Z at 5% level of significance is 1.96

Hence calculated Z > tabulated Z.

$\therefore$ The null hypothesis H_0 is rejected.

Example 8:

An ambulance service claims that it takes on the average less than 10 minutes to reach its destination in emergency calls. A sample of 36 calls has a mean of 11 minutes and the variance of 16 minutes. Test the significance at 0.05 level.

Solution:

Given $n = 36, \bar{x} = 11, \mu = 10$ and $\sigma = \sqrt{16} = 4$

1. **Null Hypothesis** H_0**:** $\mu = 10$
2. **Alternative Hypothesis** H_1**:** $\mu \neq 10$

3. **Level of significance** $\alpha = 0 \cdot 05$

4. **The test statistic is,** $Z = \dfrac{\bar{x} - \mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{11-10}{4/\sqrt{36}} = \dfrac{6}{4} = 1 \cdot 5$

Tabulated value of Z at 5% level of significance is 1.64

Hence calculated Z < tabulated Z

$\therefore$ we accept the null hypothesis H_0

Example 9:

An insurance agent has claimed that the average age of policy holders who issue through him is less than the average for all agents which is 30.5 years. A random sample of 100 policy holders who had issued through him gave the following age distribution.

Age	16 - 20	21 - 25	26 - 30	31 - 35	36 - 40
No.of persons	12	22	20	30	16

calculate the Arithmetic mean and Standard deviation of this distribution and use these values to test his claim at 5% level of significance.

Solution:

Take $A = 28, d_1 = x_i - A$

$$\bar{x} = A + \frac{h \sum f_i d_i}{N} = 28 + \frac{5 \times 16}{100} = 28 \cdot 8$$

S.D. $S = h \sqrt{\dfrac{\sum fd^2}{N} - \left(\dfrac{\sum fd}{N}\right)^2} = 5 \cdot \sqrt{\dfrac{164}{100} - \left(\dfrac{16}{100}\right)^2}$ $\quad [\because h = 5]$

1. **Null Hypothesis** H_0**:** The sample is drawn from a population with mean μ i.e. $\bar{x}$ and μ do not differ significantly where $\mu = 30 \cdot 5$ years.

2. **Alternative Hypothesis** H_1**:** $\mu < 30 \cdot 5$ years (left tail test)

Now, $\bar{x} = 28 \cdot 8, S = 6 \cdot 35, \mu = 30 \cdot 5$ years and n = 100

3. **The test statistic is,** $Z = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} = \dfrac{28 \cdot 8 - 30 \cdot 5}{6 \cdot 35/\sqrt{100}} = -2 \cdot 677$

$\therefore |Z| = 2 \cdot 68$

Tabulated value of Z at 5% level of significance is 1.645 (left tail test)

Here calculated Z > tabulated Z.

$\therefore$ The null hypothesis H_0 is rejected.

i.e., $\bar{x}$ and μ difference significantly

i.e, the sample is not drawn from a population with mean $\mu = 30 \cdot 5$ years.

4. Test of significance for difference of means:

Let $\bar{x}_1$ be the mean of a sample of size n_1 from a population with mean μ_1 and variance σ_1^2.

Let $\frac{1}{x_2}$ be the mean of a sample of size n_2 from a population with mean μ_2 and variance σ_2^2.

To test whether there is any significant difference between between $\bar{x}_1$ and $\bar{x}_2$ we have to use the statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Note:

If the samples have been drawn from the same population then $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\therefore \quad Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}}$$

If σ is not known we can use an estimate of σ^2 given by $\sigma^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2}$

Example 10:

The mean yield of wheat from a district A was 210 pounds with S.D. 10 pounds or acre from a sample of 100 plots. In another district the mean yield was 220 pounds with S.D. 12 pounds from a sample of 150 plots. Assuming that the S.D. of yield in the entire state was 11 pounds, test whether there is any significant difference between the mean yield of crops in the two districts.

Solution:

Given $\bar{x}_1 = 210$ $\bar{x}_2 = 200$

and $n_1 = 100$ $n_2 = 150$

Population S.D. $\sigma = 11$

1. **Null Hypothesis H_0:** There is no difference between $\bar{x}_1$ and $\bar{x}_2$ i.e., $H_0 : \bar{x}_1 = \bar{x}_2$

2. **Alternative Hypothesis H_1:** $\bar{x}_1 \neq \bar{x}_2$

4. **The test statistic is** $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{210 - 200}{\sqrt{11^2\left(\frac{1}{100} + \frac{1}{150}\right)}} = 7 \cdot 04178$$

i.e. $Z = 7 \cdot 041 > 1 \cdot 96$

The null hypothesis H_0 is rejected at 5% level of significance i.e., there is a difference between $\bar{x}_1$ and $\bar{x}_2$.

i.e., There is a significant difference between the mean yield of crops in the two distincts.

Note:

If the two samples are drawn from two populations with unknown Standard deviations, then we take the test statistic Z though

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Example 11:

In a survey of buying habits, 400 women shoppers are chosen at random in super market 'A' located in a certain section of the city. Their average weekly food expenditure is Rs. 250 with a S.D. of Rs. 40. For 400 women shoppers chosen at random in super market 'B' in another section of the city, the average weekly food expenditure is Rs. 220 with a S.D. of Rs 55. Test at 10% level of significance whether the average weekly food expenditure of the two populations of shoppers are equal.

Solution:

Given $n_1 = 400$ $\bar{x}_1 = \text{Rs. } 250$ $S_1 = \text{Rs. } 40$

$n_2 = 400$ $\bar{x}_2 = \text{Rs. } 220$ $S_2 = \text{Rs. } 55$

1. **Null Hypothesis H_0 :** Assume that the average weekly food expenditure of the two populations of shoppers are equal i.e., $H_0 : \bar{x}_1 = \bar{x}_2$

2. **Alternative Hypothesis H_1 :** $\bar{x}_1 \neq \bar{x}_2$

4. **The test statistic is** $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{250 - 220}{\sqrt{\frac{(40)^2}{400} + \frac{(55)^2}{400}}} = \frac{30}{3 \cdot 4} = 8 \cdot 82$$

i.e., $Z = 8 \cdot 82 > 2 \cdot 58$

$\therefore$ The null hypothesis H_0 is rejected.

i.e., The average weekly food expenditure of the two populations of shoppers are not equal.

Example 12:

Samples of students were drawn from two universities and from their weights in kilograms, mean and standard deviations are calculated and shown below. Moke a large sample test to test the significance of the difference between the means.

	Mean	S.D	Size of the sample
University A	55	10	400
University B	57	15	100

Solution:

Given $\bar{x}_1 = 55,\ \bar{x}_2 = 57,\ n_1 = 400,\ n_2 = 100$

$S_1 = 10$ and $S_2 = \text{Rs} \cdot 15$

1. **Null Hypothesis H_0 :** $\bar{x}_1 = \bar{x}_2$ i.e., there is no difference

2. **Alternative Hypothesis H_1 :** $\bar{x}_1 \neq \bar{x}_2$

3. **Level of significance $\alpha = 0.05$ (assumed)**

4. **Critical Region:** Accept H_0 if - 1.96 < Z < 1.96

5. The test statistic is $$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{55 - 57}{\sqrt{\frac{100}{400} + \frac{225}{100}}} = \frac{-2}{\sqrt{\frac{1}{4} + \frac{9}{4}}} = -1 \cdot 26$$

$$\therefore \left| Z \right| = 1 \cdot 26 < 1 \cdot 96$$

So we accept the null hypothesis H_0 at 5% level of significance i.e., there is no significant difference between the means.

11.6 Summary:

In this lesson we have shown concept of test of hypothesis and computation procedure. Further, we have given important of test of hypothesis in practical problems.

11.7 Exercise:

1. In a random sample of 400 persons from a large population, 120 are females. Can it be said that males and females are in the ratio 5 : 3 in the population? Use 1% level of significance.
2. A coin is tossed 900 times and heads appear 490 times. Does this result support the hypothesis that the coin is unbiased.
3. A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random sample of 600 apples contained 36 defective apples. Test the claim of the wholesaler.
4. In a sample of 500 people in Tamil Nadu 280 are tea drinkers and the rest are coffee drinkers. Can we assume that both coffee and tea are equally popular in this state at 1% level of significance.
5. A manufacturer claimed that atleast 98% of the steel pipes which he supplied to a factory conformed to specifications. An examination of a sample of 500 pieces of pipes revealed that 30 were defective. Test his claim at a significance level 5%.
6. The machine puts out 16 imperfect articles in a sample of 500. After machine is overhauled, it puts out 3 imperfect articles in a batch of 100. Has the machine improved?
7. In a sample of 600 students of a certain college 400 are found to use dot pens. In another college, from a sample of 900 students 450 were found to use dot pens. Test whether the two colleges are significantly different with respect to the habit of using dot pens.
8. 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 often 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?
9. A machine produced 20 defective articles in a batch of 400. After overhauling it produced 10 defectives in a batch of 300. Has the machine improved.
10. A random sample of 100 articles selected from a batch of 2,000 articles shows that the average diameter of the articles = 0.354 with a S.D. = 0.048. Find 95% confidence interval for the average of this batch of 2,000 articles.
11. A sample of 100 iron bars is said to be drawn from a large number of bars whose lengths are normally distributed with mean 4 feet and S.D. 0.6 feet. If the sample mean is 4.2 feet, can the sample be regarded as a truly random sample.

12. The mean life of a sample of 100 electric bulbs produced by a company is found to be 1570 hrs with a S.D. of 120 hrs. If μ is the mean life time of all the bulbs produced by the company, test the hypothesis $\mu = 1600$ hrs against the alternative hypothesis $\mu \neq 1600$ hrs at 5% level of significance.

13. A sample of 400 male students is found to has a mean height of 171.38 cm. Can it be reasonably regarded as a sample from a large population with mean height of 171.17 cm and S.D. 3.30cm.

14. A sample of 100 workers in a large plant gave a mean assembly time of 294 seconds with a S.D.of 12 seconds in a time and motion study. Find a 95% confidence interval for the mean assembly time for all the workers in the plant.

15. The mean breaking strength of the cables supplied by a manufacturer is 1800 with a S.D. 100. By a new technique in the manufacturing process, it is claimed that the breaking strength of the cables have increased. in order to test this claim, a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance.

16. An investigation of the relative merits of two kinds of flash light batteries showed that a random sample of 100 batteries of brand A tested on the average 36.5 hrs with a S.D. of 1.8 hrs while a random sample of 80 batteries of brand B tested on the average 36.8 hrs with a S.D. of 1.5 hrs. Use a level of significance of 0.05 to test whether the observed difference between the average life times is significant.

17. Given the following information relating to two places A and B. Test whether there is any significant difference between their mean wages:

	A	**B**
Mean wages (Rs.)	47	49
S.D. (Rs.)	28	40
Number of workers	1000	1500

18. The mean consumption of food grains among 400 sampled middle class consumers is 180 gms per day per person with a S.D. of 120 gms. A similar sample survey of 600 working class consumers gave a mean of 410 gms with a S.D. of 80 gms. Are we justified in saying that the two classes consume the same quantity of food grains. Use 5% level of significance.

19. In a city 250 men out of 750 were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers.

11.8 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer
Dr. K. CHANDAN

Lesson - 12

STUDENT 't' TEST

Objective:

After going through this lesson you will learn:

- The concept of students 't' test
- Computation of students 't' test

Structure:

12.1 Introduction:

The small size sample of 'n' the sampling distribution is assumed to follow a new distribution called 't - distribution'. It also hold good only when the samples are drawn from a normal population. Students t distribution was given and used by W.S. Gosset in 1908. The statistician W.S. Gosset is better known by his pseudonym, student.

12.2 Test of Significance:

A very important aspect of the sampling theory is the study of tests of significance which enable us to decide on the basis of the sample results if

(i) The deviation between the observed sample statistic and the hypothetical parameter value is significant.

(ii) The deviation between two sample statistics is significant.

Test of significance for small samples:

Definition:

When the size of the sample (n) is less than 30, then that sample is called a small sample. i.e., when $n < 30$, the sample is called a small sample. Students t is the deviation of estimated mean from its population mean expressed interms of standard error. The test criteria is; reject H_0 if $t \geq t_{\alpha/2,\ n-1}$ or $t \leq -t_{\alpha/2,\ n-1,}$ otherwise accept H_0.

12.3 Students 't' Test:

The student's t is defined by the statistic

$$t = \frac{\bar{x} - \mu}{S.D./\sqrt{n-1}}$$

where $\bar{x} = \frac{1}{n}\sum x_i$ = sample mean, μ = population mean,

n = sample size, $(n-1) \rightarrow$ degrees of freedom

S.D., 'S' $= \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$

Note:

1. If the standard deviation of a sample is given directly, then the statistic is given by

$$\bar{t} = \frac{\bar{x} - \mu}{S \cdot D \cdot /\sqrt{n-1}}$$

2. If the calculated value of t exceeds the table value of t at 5% level of significance, the null hypothesis H_0 is rejected. If the calculated value of t is less than the tabulated value of t at 5% level the null hypothesis is accepted.

Confidence or Fiducial Limits For μ:

If $t_{0\cdot05}$ is the table value of t for $(n-1)$ degrees of freedom at 5% level of significance then 95% confidence limit for μ is given by $\bar{x} \pm t_{0\cdot05} \cdot \frac{S}{\sqrt{n}}$

For $P(|t| > t_{0\cdot05}) = 0\cdot05$

i.e. $P(|t| < t_{0\cdot05}) = 0\cdot95$

95% confidence limits for μ are given by

$$|t| \leq t_{0\cdot05}$$

i.e. $$\left|\frac{\bar{x} - \mu}{S/\sqrt{n}}\right| \leq t_{0\cdot05}$$

$$\Rightarrow \quad -t_{0\cdot05} \le \frac{\bar{x}-\mu}{S/\sqrt{n}} \le t_{0\cdot05}$$

$$\Rightarrow \quad \bar{x} - t_{0\cdot05} \cdot S/\sqrt{n} \le \mu \le \bar{x} + t_{0\cdot05} \cdot S/\sqrt{n}$$

Similarly 99% confidence limits for μ is $\bar{x} \pm t_{0\cdot01} \cdot S/\sqrt{n}$

where $t_{0\cdot01}$ is the tabulated value of t for (n - 1) degrees of freedom at 1% level of significance.

Students 't' test for single mean:

Suppose we want to test

(a) If a random sample x_i of size n has been drawn from a normal population with a specified mean μ_0.

(b) If the sample mean differs significantly from the hypothetical value μ_0 of the population mean.

In this case the statistic is given by

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n-1}}$$

or $$t = \frac{\bar{x}-\mu}{S \cdot D \cdot /\sqrt{n-1}}$$

where $\bar{x}$, μ, S, n have usual meanings.

Example 1:

A mechanist is making engine parts with axle diameters if 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a S.D. of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specification.

Solution:

Here the sample size $n = 10 < 30$

Hence the sample is small sample.

Also sample mean $\bar{x} = 0\cdot742$ inches and population mean $\mu = 0\cdot700$ inches are given

S.D. = 0.040 inches

We use student's t test.

Null Hypothesis H_0 : The product is confirming to specification.

Alternative Hypothesis H_1 : $\mu \neq 0 \cdot 700$

The test statistic is, $t = \dfrac{\bar{x} - \mu}{S \cdot D \cdot / \sqrt{n-1}}$

Here $\bar{x} = 0 \cdot 742$ inches, μ = 0.700 inches, S.D. = 0.040 inches and n = 10.

Degrees if freedom (d.f.) = n - 1 = 10 - 9 = 9

$$\therefore\ t = \frac{0 \cdot 742 - 0 \cdot 700}{\frac{0 \cdot 040}{\sqrt{10-1}}} = 3 \cdot 15$$

$\therefore$ The calculated value of t = 3.15

The table value of t are 5% level with 9 degrees of freedom

$t_{0 \cdot 05} = 2 \cdot 26$

Since calculated value of t > tabulated value of t, therefore H_0 is rejected.

$\therefore$ The product is not meeting the specification.

Example 2:

A sample of 26 bulbs gives a mean life of 990 hours with a S.D. of 20 hours. The manufacturer claims that the mean life of bulbs is 1000 hours. Is the sample not upto the standard.

Solution:

Here sample size 26 < 30

$\therefore$ The sample is small sample.

Also given sample mean $\bar{x} = 990$

Population mean $\mu = 1000$

S.D. = 20

Degrees of freedom = n - 1 = 26 - 1 = 25

Here we know $\bar{x}$, μ, $S \cdot D \cdot$ and n.

$\therefore$ We use students 't' test.

Null Hypothesis H_0**:** The sample is upto the standard.

Alternative Hypothesis H_1**:** $\mu < 1000$ (The sample is below standard) (left tail test)

The test statistic is $t = \frac{\bar{x} - \mu}{S \cdot D \cdot / \sqrt{25}} = \frac{990 - 1000}{20 / \sqrt{25}} = -2 \cdot 5$

$|t| = 2 \cdot 5$

$\therefore$ calculated value of t = 2.5

Table value of 't' at 5% level with 25 degrees of freedom for left - tailed test is 1.708

Since calculated t > tabulated t, we reject H_0

i.e., We reject the null hypothesis

$\therefore$ The sample is not upto the standard.

Example 3:

A machine is designed to produce insulating washers for electrical devices of average thickness of 0.025 cm. A random sample of 10 washers was found to have a thickness of 0.024 cm with a S.D. of 0.002 cm. Test the significance of the deviation. Value of t for 9 degrees of freedom at 5% level is 2.262.

Solution:

Here the sample size is 10 < 30

$\therefore$ The sample is small

Also given sample mean $\bar{x}$ = 0.024 cm

Population mean μ = 0.025 cm

S.D. = 0.002 cm

Degrees of freedom = (n-1) = 10 - 1 = 9

Null Hypothesis H_0**:** The difference between $\bar{x}$ and μ is not significant.

Alternative Hypothesis H_1**:** $\mu_1 \neq 0 \cdot 025$

Students 't' $= \frac{\bar{x} - \mu}{S \cdot D \cdot / \sqrt{n-1}} = \frac{0 \cdot 024 - 0 \cdot 025}{\frac{0 \cdot 002}{\sqrt{10-1}}} = -1 \cdot 5$

$\Rightarrow |t| = 1 \cdot 5$

$\therefore$ calculated value of t = 1.5 for two tailed test.

Tabulated value of t for 9 degrees of freedom at 5% level = 2.262

Since calculated t < tabulated t we accept the null hypothesis.

i.e., difference between $\bar{x}$ and μ is not significant.

Example 4:

The mean life time of a sample of 25 fluorescent light bulbs produced by a company is computed to be 157 hours with a S.D. of 120 hours. the company claims that the average life of the bulbs produced by the company is 1600 hours using the level of significance of 0.05. Is the claim acceptable?

Solution:

Given sample size n = 25

Sample mean, $\bar{x} = 1570$

Population mean, $\mu = 1600$

S.D. (S) = 120

Degrees of freedom = n - 1 = 24

Null Hypothesis H_0**:** The claim is acceptable. $\mu = 1600$ hrs

Alternative Hypothesis H_1**:** $\mu \neq 1600$ hrs

The test statistic is $t = \frac{\bar{x} - \mu}{S \cdot D / \sqrt{n\,1}} = \frac{1570 - 1600}{120/\sqrt{24}} = \frac{-30}{24 \cdot 49} = -1 \cdot 22$

$\therefore |t| = 1 \cdot 22$

$\therefore$ Calculated t = 1.22

The tabulated value of t at 5% level with 24 degrees of freedom for two tailed test is 2.06

Since the calculated value of t < tabulated value of t, we accept the null hypothesis H_0, i.e. the claim that the average life of the bulbs produced by the company is 1600 hrs is acceptable.

Example 5:

The average breaking strength of the steel rods is specified to be 18.5 thousand pounds. To test this sample of 14 rods was tested. The mean and standard deviations obtained were 17.85 and 1.955 respectively. Is the result of experiment significant?

Solution:

Given sample size, n = 14

Sample mean, $\bar{x}$ = 17.85

S.D. (S) = 1.955

Population mean, μ = 18.5

Degrees of freedom = n - 1 = 13

null Hypothesis H_0 : The result of the experiment is not significant.

Alternative Hypothesis H_1 : $\mu \neq 18 \cdot 5$

The test statistic is, $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n-1}} = \dfrac{17 \cdot 85 - 18 \cdot 5}{1 \cdot 955/\sqrt{13}} = \dfrac{0 \cdot 65}{0 \cdot 542} = -1 \cdot 199$

$|t| = 1 \cdot 199$

$\therefore$ Calculated t = 1.199

Tabulated t at 5% level for 13 d.f. for two tailed test = 2.16

Since calculated t < tabulated t. We accept H_0 at 5% level

i.e., the result of the experiment is not significant.

Example 6:

A random sample of size 16 values from a normal population showed a mean of 53 and a sum of squares of deviation from the mean equals to 150. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% confidence limits of the mean of the population.

Solution:

Given, sample size n = 16

Sample mean, $\bar{x} = 53$

$\sum (x - \bar{x})^2 = 150$ $\qquad \therefore S^2 = \dfrac{\sum (x - \bar{x})^2}{n-1} = \dfrac{150}{15} = 10$

$\therefore S = \sqrt{10}$

Degrees of freedom, $\gamma = n - 1 = 15$

Null Hypothesis H_0 : The sample is taken from the population having 56 as mean i.e., $\mu = 56$

Alternative Hypothesis H_1 : $\mu \neq 56$

The test statistic, $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$ (Note that S.D. is not given directly)

$$= \frac{53 - 56}{\sqrt{10}/\sqrt{16}} = -3 \cdot 79$$

$$\therefore \ |t| = 3 \cdot 79$$

The tabulated value of t at 5% level of significance for 15 d.f. for two tailed test is 2.13.

Since calculated t > tabulated t the null hypothesis H_0 is rejected i.e., the sample cannot be regarded as taken from the population.

The 95% confidence limits of the mean of the population are given by

$$\bar{x} \pm t_{0 \cdot 05} \frac{S}{\sqrt{n}}$$

$$= 53 \pm 2 \cdot 13 \times 0 \cdot 79$$

$$= 53 \pm 1 \cdot 6827 = 54 \cdot 68 \text{ and } 51 \cdot 31$$

Hence 95% confidence limits are $[51 \cdot 31,\ 54 \cdot 68]$

Example 7:

A random sample of six steel beams has a mean compressive strength of 58,392 p.s.i (pounds per square inch) with a standard deviation of 648 p.s.i. Use this information and the level of significance $\alpha = 0 \cdot 05$ to test whether the true average compressive strength of the steel from which this sample came is 58,000 p.s.i. Assume normality.

Solution:

We have

n = Sample size (number of steel beams 0 = 6 < 30. $\therefore$ The sample is small.

$\bar{x}$ = sample mean (average compressive strength) = 58392 psi.

S = Standard deviation of six beams = 648 psi

Degree of freedom (d.f.) = n - 1 = 6 - 1 = 5

In this problem σ is unknown and n < 30. Hence we use t - distribution.

1. **Null Hypothesis is** H_0**:** $\mu = 58000$

2. **Alternative Hypothesis** H_1**:** $\mu \neq 58000$

3. Level of significance $\alpha = 0 \cdot 05$

4. **Critical Region:** Since alternative hypothesis is of the type $'\neq$, the test is two tailed and the critical region is - 3.365 < t < 3.365

5. The test statistic is $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}} = \dfrac{58392 - 58000}{648/\sqrt{6}} = 1 \cdot 482$

Since $t = 1 \cdot 482 < 3 \cdot 365 = t_{\alpha/2}$ we accept the null hypothesis H_0.

Hence the average compressive strength of the steel beam is not equal to 58000 psi.

Problems related to student's t - test (when S.D. of the sample is not given diveetly)

Example 8:

Samples of two types of electric light bulbs were tested for length of life and following data were obtained

	Type I	**Type II**
Sample number	$n_1 = 8$	$n_2 = 7$
Sample means	$\bar{x}_1 = 1234$ hours	$\bar{x}_2 = 1036$ hrs
Sample S.D.	$S_1 = 36$ hrs	$S_2 = 40$ hrs

If the difference in the means sufficient to warrant that type I is superior to type II regarding length of life.

Solution:

Null Hypothesis H_0**:** The two types I and II of electric bulbs are identical.

Alternative Hypothesis $H_1: \mu_1 > \mu_2$

Since two sample means $\bar{x}_1$ and $\bar{x}_2$ are given and also sample standard deviations S_1 and S_2 are given, we use the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S^2 \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

where $S^2 = \frac{1}{n_1 + n_2 - 2}\left(n_1 S_1^2 + n_2 S_2^2\right)$

$$= \frac{1}{8 + 7 - 2}\left[8(36)^2 + 7(40)^2\right] = 1659 \cdot 08$$

$$\therefore\ t = \frac{1234 - 1036}{\sqrt{1659 \cdot 08\left(\frac{1}{8} + \frac{1}{7}\right)}} = 9 \cdot 39$$

Degree of freedom (d.f) = $n_1 + n_2 - 2$

$= 8 + 7 - 2 = 13$ at 5% level.

Tabulated value of t for 13 d.f. at 5% level is 1.77 (one - taile test)

Since calculated t > tabulated t, we reject the null hypothesis H_0.

i.e., The two types I and II of electric bulbs are not identical.

Example 9:

The means of two random samples of sizes 9 and 7 are 196.42 and 198.82 respectively. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the sample be considered to have been drawn from the same normal population.

Solution:

Given $n_1 = 9$ $n_2 = 7$

$\bar{x}_1 = 196 \cdot 42$ $\bar{x}_2 = 198 \cdot 82$

$\Sigma\left(x_1 - \bar{x}_1\right)^2 = 26 \cdot 94$ $\Sigma\left(x_2 - \bar{x}_2\right)^2 = 18 \cdot 73$

$$S^2 = \frac{\Sigma\left(x_1 - \bar{x}_1\right)^2 + \Sigma\left(x_2 - \bar{x}_2\right)^2}{n_1 + n_2 - 2} = \frac{26 \cdot 94 + 18 \cdot 73}{9 + 7 - 2} = 3 \cdot 26$$

$S = 1 \cdot 81$

null Hypothesis H_0: The two samples are drawn from the same population.

i.e., $\mu_1 = \mu_2$

Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$

The test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\dfrac{1}{n_2} + \dfrac{1}{n_2}}} = \dfrac{196 \cdot 42 - 198 \cdot 82}{1 \cdot 81\sqrt{\dfrac{1}{9} + \dfrac{1}{7}}}$

$$= \frac{-2 \cdot 4}{0 \cdot 912} = -2 \cdot 63$$

i.e., calculated value of $|t| = 2 \cdot 63$

Tabulated value of t for 9 + 7 - 2 = 14 d.f. at 5% level of significance is 2.15

Since calculated value of t > tabulated value of t, we reject the null hypothesis H_0.

i.e., The two samples are not drawn from the same population.

Example 10:

Below are given the gain in weights (in bls) of pigs fed on two diets A and B.

Diet A	25	32	30	34	24	14	32	24	30	31	35	25	-	-
Diet B	44	34	22	10	47	31	40	32	35	18	21	35	29	22

Test if the two diets significantly as regards their effect on increase in weight.

Null Hypothesis H_0 : There is no significant difference between the mean increase in weight due to diets A and B i.e., $\mu_1 = \mu_2$

Alternative Hypothesis H_1 : $\mu_1 \neq \mu_2$

Calculation for sample means and S.D.'s

$$\bar{x} = \frac{336}{12} = 28, \quad \bar{y} = \frac{450}{15} = 30$$

Here $n_1 = 12, \; n_2 = 15$

x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
25	- 3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	- 8	64
34	6	36	10	- 20	400
24	- 4	16	47	17	289
14	- 14	196	31	1	1
32	4	16	40	10	100
24	- 4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	- 12	144
25	- 3	9	21	- 9	81
			35	5	25
			29	- 1	1
			22	- 8	64
336		380	450		1410

$$\sum(x-\bar{x})^2 = 380, \quad (y-\bar{y})^2 = 1410$$

$$\therefore\ S^2 = \frac{1}{n_1+n_2-2}\left[\sum(x-\bar{x})^2 + \sum(y-\bar{y})^2\right]$$

$$= \frac{1}{12+15-2}\left[380+1410\right] = 71\cdot 6$$

$$\therefore\ t = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{30-28}{\sqrt{71\cdot 6\left(\frac{1}{12}+\frac{1}{15}\right)}} = 0\cdot 609$$

Degrees of freedom $= n_1 + n_2 - 2$

$= 12 + 15 - 2 = 25$

Tabulated t for 25 d.f. at 5% level = 2.06

Since calculated t < tabulated t,

$\therefore$ We accept the null hypothesis H_0

i.e., There is no significant difference between the mean increase in weigth due to diets A and B.

Example 11:

A groups of 5 patients treated with medicine A weigh 42, 39, 48, 60 and 41 kgs. Second group of 7 patients from the same hospital treated with medicine B weight 38, 42, 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine B increases the weight significantly.

Solution:

Calculation for sample means and S.D.'s

x	$x - \bar{x}$	$(x - \bar{x})^2$	y	$(y - \bar{y})$	$(y - \bar{y})^2$
42	- 4	16	38	- 19	361
39	- 7	49	42	- 15	225
48	2	4	56	- 1	1
60	14	196	64	7	49
41	- 5	25	68	11	21
			69	12	44
			65	5	25
230	0	290	399	0	926

Now $\bar{x} = \frac{230}{5} = 46, \quad \bar{y} = \frac{399}{7} = 57$

$\sum (x - \bar{x})^2 = 290, \ \sum (y - \bar{y})^2 = 926$

$$\therefore S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2\right] = \frac{1}{5 + 7 - 2}\left[290 + 926\right] = 121 \cdot 6$$

Null Hypothesis H_0: There is no significant difference between the medicines A and B as regards their effect on increase in weight i.e., $H_0 = \mu_1 = \mu_2$

Alternative Hypothesis H_1 **:** $\mu_1 > \mu_2$

The test statistics is $t = \dfrac{\bar{x} - \bar{y}}{S\sqrt{\dfrac{1}{n_2} + \dfrac{1}{n_2}}} = \dfrac{46 - 57}{11 \cdot 07\sqrt{\dfrac{1}{5} + \dfrac{1}{7}}} = -1 \cdot 7$

$\therefore$ Calculated value of $|t| = 1 \cdot 7$

Tabulated value of t for 5 + 7 - 2 = 10 d.f. at 5% level of significant is 1.81 (right - tail - test).

Since calculated t < tabulated t, we accept H_0 i.e., The medicines A and B do not differ significantly as regards their effect on increase in weight.

Example 12:

Two horses A and B were tested according to the time (in seconds) to run a particular track with the following results.

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity.

Solution:

Given $n_1 = 7,\ n_2 = 6$

We first compute the sample means and standard deviations.

$\bar{x}$ = Mean of first sample $= \dfrac{1}{7}(28 + 30 + 32 + 33 + 33 + 29 + 34)$

$= \dfrac{1}{7}(219) = 31.286$

$\bar{y}$ = Mean of second sample $= \dfrac{1}{6}(29 + 30 + 30 + 24 + 27 + 29)$

$= \dfrac{1}{6}(169) = 28.16$

x	$x-\bar{x}$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$
28	- 3.286	10.8	29	0.84	0.7056
30	- 1.286	1.6538	30	1.84	3.3856
32	0.714	0.51	30	1.84	3.3856
33	1.714	2.94	24	- 4.16	17.3056
33	1.714	2.94	27	- 1.16	1.3456
29	- 2.286	5.226	29	0.84	0.7056
34	2.714	7.366			
219		31.4358	169		26.8336

Now $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum(\bar{x} - \bar{x})^2 + \sum(\bar{y} - \bar{y})^2\right]$

$= \frac{1}{11}[31\cdot4358 + 26\cdot8336] = \frac{1}{11}(58\cdot2694) = 5\cdot23$

$\therefore\ S = \sqrt{5\cdot23} = 2\cdot3$

1. **Null Hypothesis** H_0**:** $\mu_1 = \mu_2$

2. **Alternative Hypothesis** H_1**:** $\mu_1 \neq \mu_2$

3. Level of significance, $\alpha = 0\cdot05$

4. **The test statistic is t** $= \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{31\cdot286 - 28\cdot16}{(2\cdot3)\sqrt{\frac{1}{7} + \frac{1}{6}}} = 2\cdot443$

Tabulated value of t for 7 + 6 - 2 = 11 d.f. at 5% level of significance is 2.2.

Since calculated t > tabulated t, we reject the null hypothesis H_0. That is both horses A and B do not have the same running capacity.

12.4 Summary:

In this lesson we have shown importance of student's t-test and computational procedures.

12.5 Exercises:

1. A random sample of 10 boys had the following I.Q.'s

   70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100? Find a reasonable range in which most of the mean. I.Q. values of samples of 10 boys lie.

2. The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inches. Is it reasonable to believe that the average height is greater than 64 inches? Test at 5% significance level assuming that for 9 degrees of freedom P (t > 1.83) = 0.05.

3. A random sample from a company's very extensive files shows that the orders for a certain kind of machinery were filled, respectively in 10, 12, 19, 14, 15, 18, 11 and 13 days. Use the level of significance $\alpha = 0 \cdot 01$ to test the claim that on the average such orders are filled in 10.5 days. Assume normality.

4. To examine the hypothesis that the husbands are more intelligent than the wives, an investigator took a sample of 10 couples and administered them a test which measures the I.Q. The results are as follows:

Husbands	117	105	97	105	123	109	86	78	107	107
Wives	106	98	87	104	116	95	90	69	108	85

   Test the hypothesis with a reasonable test at the level of significance of 0.05

5. Find the maximum difference that we can expect worth probability 0.95 between the means of samples of sizes 10 and 12 from a normal population if their standard deviations are found to be 2 and 3 respectively.

6. The independent samples of 7 items respectively had the following values.

Sample I	11	11	13	11	15	9	12	14
Sample II	9	11	10	13	9	8	10	-

   Is the difference between the means of samples significant?

7. Measuring specimens of nylon yarn, taken from two machines, it was found that 8 specimens from first machine had a mean denier of 9.67 with a standard deviation of 1.81 while 10 specimens from second machine had a mean denier of 7.43 with a standard deviation of 1.48. Assuming that the proportions are normal, test the hypothesis $H_0 : \mu_1 - \mu_2 = 1 \cdot 5$ against $H_1 : \mu_1 - \mu_2 > 1 \cdot 5$ at 0.05 level of significance.

8. Ten soldiers participated in a shooting competition in the first week. After intensive training they participated in the competition in the second week. Their scores before and after training are given as follows:

Scores Before:	67	24	57	55	63	54	56	68	33	43
Scores After:	70	38	58	58	56	67	68	75	42	38

Do the data indicate that the soliders have been benefited by the training.

9. To compare two kinds of bumper guards, 6 of each kind were mounted on a car and then the car was run into a concrete wall. The following are the costs of repairs.

Guard 1	107	148	123	165	102	119
Guard 2	134	115	112	151	133	129

Use the 0.01 level of significance to test whether the difference between two sample means is significant.

11.8 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer

Dr. K. CHANDAN

Lesson - 13

χ^2 - TEST

Objectives:

After going through this lesson you will learn:

- The concept of χ^2 - test and computation of χ^2 test.

Structure:

13.1 Introduction:

The idea behind the chi-square goodness-of-fit test is to see if the sample comes from the population with the claimed distribution. Another way if looking at that is to ask if the frequerey ditribution fits a specific pattern. The idea is that if the observed frequency is really close to the expected frequency, then the square of the deviation will be small. The square of the deviation is divided by the expected frequency to weight frequencies. A difference of 10 many bevery significant if 12 was the expected frequency, but a difference if 10 is n't very significant at all if the expected frequency was 1200. Only when the sum is large is the a reason to question the distribution. Therefore the chi-square goodness of fit test is always a right tail test.

13.2 Chi - Square Distribution:

If the population various is σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi - square distribution with n-1 degrees if freedom. To set a confidence interval σ^2 we can locate two points on the χ^2 distribution, χ^2 UCL cuts off 0.025 of the area in the upper tail of the distribution and χ^2 LUL cut of 0.025 in lower tail for 5 percent level of significance.

13.3 Applications of χ^2 analysis:

(i) Chi - square test allows the comparison of more than two population proportions, as such comparisons cannot be handled by t and z distributions.

(ii) Chi - square test helps in determining the independence or otherwise of the attributes when data is classified into several proportions.

(iii) Chi - square test also helps in deciding whether a particular probability distribution, such as binomial, poisson or normal is appropriate distribution of data under consideration.

13.4 Hypothesis Testing of Population Variance:

Many times there sponsible decision makers may wish to know the average maximum and minimum variability in attempting to deliver a particular service in an organization. In such a situation they have to make inferences about population variability in order to schedule machines and men to meet peak working schedules. To illustrate the issues involved in such a situation, let us consider the following illustration.

Example 1: Customer Service in a Bank

Having received complaints about slow customer service in a certain public counters in a Bank branch, the chief - executive of a most busy bank office ordered for a preliminary investigation. The investigation officer enquired about the time spent by 10 customers in receiving similar bank service in a peak day of banking activities. He has provided the following schedule of information.

Customer Service Time In a Particular Bank

Time Spent by a Customer X (minutes)	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$
60	54	6	36
45	54	- 9	81
25	54	- 29	841
75	54	21	441
60	54	6	36
30	54	- 24	576
90	54	36	1296
15	54	39	1521
80	54	26	676
60	54	6	36
540			$\sum(X - \bar{X})^2 = 5540$

$$\bar{X} = \frac{540}{10} = 54 \text{ minutes} \qquad S^2 = \frac{(X-\bar{X})^2}{n-1} = 615\cdot55$$

$$S = \sqrt{S^2} = 248 \text{ minutes}$$

From the above information, the investigator can report to the top executive about population standard deviation of approximately 25 minutes in getting the bank service rendered. In order to estimate a range for such variance at 95 percent confidence interval, the value of X^2 distribution at two different points.

X^2 UCL cut off for 0.025 of area in upper tail : 19.023

X^2 LCL cut off for 0.025 of area in lower tail : 2.700

Lower Limit $\rightarrow \sigma LCL^2 = \frac{(n-1)S^2}{X_U^2} = \frac{9(615\cdot55)}{19\cdot023} = 291\cdot22$

Upper Limit $\rightarrow \sigma UCL^2 = \frac{(n-1)S^2}{X_L^2} = \frac{9(615\cdot55)}{2700} = 2051\cdot83$

In terms of standard deviation, the average deviation of time in customer service ranges between 17 minutes to 45.3 minutes.

Example 2: Adjustment to the diameter of the mouth of a tooth paste tube:

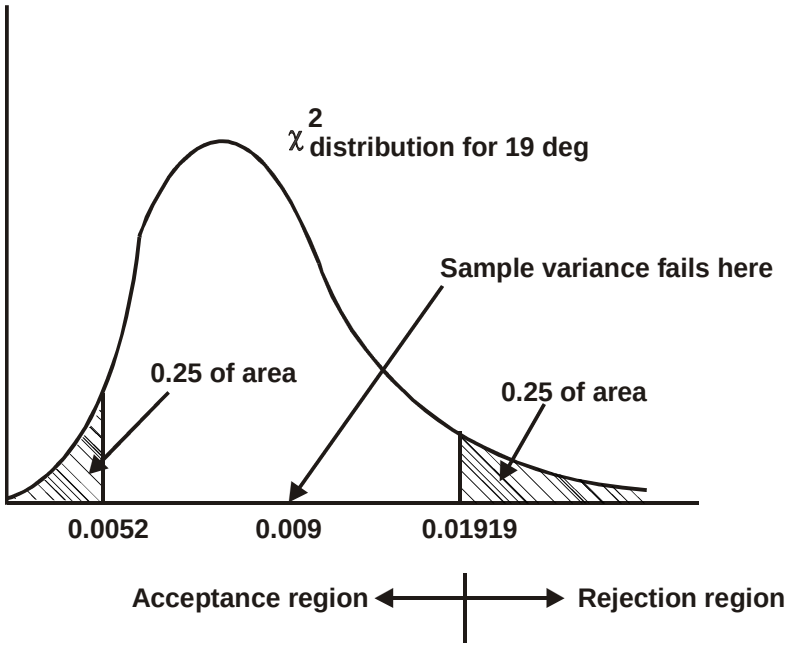
A manufacturer of tooth paste tubes as a packaging material to a large sized tooth paste giant claims that they have perfected the to diameter of the mouth of a tooth paste tube that they supply. The original variance was 0.0110 inches. The production manager has selected 20 tooth paste tubes after the adjustment made and found that the variance is reduced to 0.009. Does it support the claim of the manufacturer?

To validate it, let us consider limits for a hypothesised value of 0.010 inches. At 5% significance level, the area of X^2 for $(20-1)$ degrees of freedom would be 32.852 for Upper Limit and 8.907 for Lower Limit. Then the range of variance would be:

$$\sigma^2 \text{ UCL} = \frac{(n-1)S^2}{X_{LCL}} = \frac{19(0\cdot009)}{8\cdot907} = 0\cdot01919$$

$$\sigma^2 \text{ LCL} = \frac{(n-1)S^2}{X_{ULC}^2} = \frac{19(0\cdot009)}{32825} = 0\cdot0052$$

Fig Chi - Square Distribution For 19 d.f.



Against a hypothesised value of variance of 0.010 the actual adjusted variance falls within the acceptance region. Hence the claim of perfection is not tenable as such adjustment has not brought down the variance to a sizeable small size.

13.5 Hypotheses About Two Populations:

Difference in two sample variances:

In order to analyses variance, we assume that each of the samples is drawn from a Normal distribution and each with same variance. If the samples are selected from two normally distributed populations, with σ_1^2 equal to σ_2^2, a ratio of the sample variances, S_1^2 / S_2^2 (in case of sample variance denoted as S^2), has a probability distribution known as F distribution. The F statistic is

$$F = \frac{S_1^2}{S_2^2}$$

Example 3: Comparison of two mutual fund investments:

Mr. Chaterjee has some money to invest. Instead of investing directly in company stocks, he has preferred investment in mutual fund companies. He is considering two mutual funds viz., CANBANK, SBI MAGNAM. SBI CANBANK's. If SBI MAGNAM's variability in rate of return is significantly lower than CANBANK's, he will invest his money there. If there is no significant, difference in variability, he will go with CANBANK. To make a decision Mr. Chaterjee has considered

a sample 25 monthly returns of both the firms by taking the market prices quoted for the funds. For SBI MAGNAM the standard deviation was 3 and for CANBANK the standard deviation was 5. Advise Mr. Chatterjee.

Here, Let us make Null and Alternative Hypothesis

$$\left.\begin{array}{c} H_0 \sigma_1^2 = \sigma_2^2 \\ (\text{or}) \\ \sigma_1^2 / \sigma_2^2 = 1 \end{array}\right\}$$ Null Hypothesis: The variances in the returns of two Mutual Fund companies are the same

$$\left.\begin{array}{c} H_1 \sigma_1^2 \neq \sigma_2^2 \\ (\text{or}) \\ \sigma_1^2 / \sigma_2^2 \neq 1 \end{array}\right\}$$ Alternative Hypothesis : that the variance are different

Significance level : 0.05 (say)

Problem details:

$\sigma_1^2 = 9$ variance of SBI MAGNAM

σ_2^2 = variance of CANBANK

n = 25 sample size

F statistic:

$$F = \frac{S_1^2}{S_2^2} = \frac{9}{25}$$

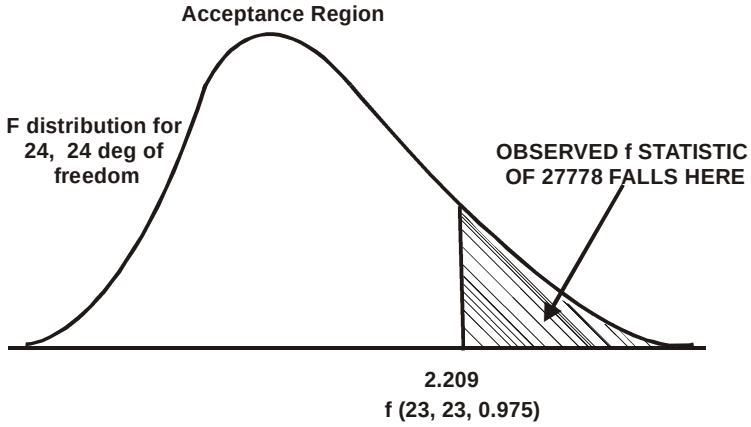
in order to find the ratio, which is larger is to be considered as numerator

$$= \frac{25}{9} = 2 \cdot 7778$$

This statistic comes from F distribution with $(n_1 - 1)$ degrees of freedom in the numerator and $(n_2 - 1)$ degrees of freedom for the denominator.

As per the F distribution tables, the critical value of Acceptance region extence upto

$$\begin{bmatrix} \text{F value} & 24 \text{ df (numerator)} \\ & 24 \text{ df (denominator)} \\ & @ \ 5\% \text{ level of signature} \end{bmatrix} \quad 2 \cdot 209$$



Since observed F = 2.778 exceeds the critical range value of 2.209, the null hypothesis is to be rejected. In other words, the data supports the view that the variance in the rates of returen of two mutual fund investments differ much.

13.6 Test of Independence:

Test of Proportions:

Comparison of two sample proportions for their significancy can be made with the help of Z or t statistics. But many times, managers may need to know whether the differences among several sample proportions are significant or not.

The first step in making comparison if more and more sample proportions and then testing for the equality of such proportions, is the arrangement of data in a contingency table. A contingency table is nothing but a classified table where in data are presented across different attributes under study. For example, let us consider the customer purchasing a certain brand of talcum powder. Suppose we have eleicited the information about the reasons in selecting that brand, the number of respondents can be classified against reasons as well as any other characteristic variable of the customer like the age. Then the data presentation would be as given in table below.

Classification of Customers by age

Reasons for selecting the brand	Age of the customer (year)		
	0 - 25	25 - 40	40 - above
Previous use	10	7	5
Advertisement in Magazines and TV	27	11	4
Display in store / retailer Persuations	12	10	4
Total	49	28	13

When the company's marketing executive wants to analyze the above data to see whether reasons vary according to age of the customer, then we have to hypothesis that the proportion of customers against each age class tentatively be the same. If advertisement as a major reason is considered, then let us say

P_y = proportion of respondents in the age group of 0 - 25 years.

P_m = proportion of respondents in the age group of 25 - 40 years.

P_0 = proportion of respondents in the age group of over 40 years.

The null hypothesis would be

$$H_0 : P_y = P_m = P_0$$

$$H_0 : P_y \neq P_m \neq P_0$$

Similarly we can draw the hypothesis for other reasons.

The X^2 statistic, in order to validate the equality of proportions against different classes, considers overall proportions as the estimate of population proportion. For the earlier illustration of reasons for selecting a brand, the overall proportions are as follows:

Proportion of customers

Reasons	Total Respondence	Proportion
(1) Previous use	22	0.2444
(2) Advertisement	42	0.4667
(3) Display	26	0.2889
Total	90	100.0

Calculation of expected frequencies:

Based on above reason - wise proportions, each cell wise proportions have to be estimated. These proportions when multiplied with number of respondents the expected number of frequencies in each cell can be arrived at. The observed frequencies in each cell and the expected frequencies would be taken to construct X^2 statistic. Below table shows the expected frequency against each cell of the continegency table.

Calculation of expected number of customers under each category

Reasons	Age		
	0 - 25	**25 - 40**	**40 - above**
(1) Previous - used observed	10	7	5
Observed frequency expected frequency	11.98	6.84	3.18
(Column x proportion)			
Total	(49 x 0.2444)	(28 x 0.2444)	(13 x 0.2444)
(2) Advertisement	27	11	4
Observed frequency	19.60	13.07	6.07
Expected frequency	(49 x 0.4667)	(28 x 0.4667)	(18 x 0.4667)
(3) Display observed Frequency	12	10	4
Expected Frequency	14.16	8.09	3.76
	(49 x 0.2889)	(28 x 0.2889)	(18 x 0.2889)
Total	49	28	13

The commonly used straightforward way to estimate the expected frequencies is

$$\text{Expected Frequency of a cell} = \frac{\text{Respective}(R)\ \text{RowTotal} \times \text{Respective}(C)\ \text{Column total}}{\text{Grand total}\ (G)}$$

For example,

$$\text{Cell 1}: \frac{R \times C}{G} : \frac{22 \times 49}{90} = 11 \cdot 98$$

↑

Expected frequency of cell 1

To test whether the Null Hypothesis of all proportions are equal among age categories, we have to compare the observed frequencies with the estimated frequencies, and work out with an assumption that null hypothesis is true. The test statistic of comparison is:

Test Statistic:

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

Where X^2 = Chi - square

O = Observed Frequency in each cell

E = Expected frequency in each cell

The chi - square is the sum of squared differences duly divided by the expected frequency. It is expected to follow chi - square distribution with (column 1 - Row - 1) degrees of freedom symbolically, it is shown as

Degrees of freedom = (r - 1) (c - 1)

Calculation of Chi - Square

Row	Column	O	E	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
1	1	10	11.98	- 1.98	3.92	0.3272
1	2	7	6.84	0.16	0.03	0.0037
1	3	5	3.19	1.82	3.31	1.0416
2	1	27	19.60	7.4	54.76	2.7939
2	2	11	13.07	- 2.07	4.28	0.3278
2	3	4	6.07	- 2.07	4.28	0.7059
3	1	12	14.16	- 2.16	4.67	0.3295
3	2	10	8.09	1.91	3.65	0.4509
3	3	4	3.76	0.24	0.06	0.0153
			Total			$X^2 = 5 \cdot 9958$

The table chi - square value with 4 degrees of freedom (i.e., (C - 1) (R - 1)) at 5% level of significance is found to be 9.488. The calculated value falls much below the critical table value. It indicates that the proportions are not different from each other. In other words, the reasons for using the specific brand and the age are independent and age has no role to specifically identify a reason.

Example 4: Testing the efficiency of a drug for flue:

A pharmaceutical company has developed a drug for common flue and was administered to 500 individuals out of 900 in a certain locality identified to decrease prone. Although the effectiveness of the drug depends upon many factors including an individual's physical stamina, an enquiry is made to test the effectiveness of the drug. The details are as follows:

	Attacked by Flue	Attacked by Flue	Total
After Administering the Drug	200	300	500
without administering the drug	250	150	400
	450	450	900

Let us test the effectiveness of the drug. It means that whether the incidence of Flue and the administration of the drug are independent or dependent. If the incidence of Flue is dependent on administration of drug, the drug is said to be effective.

Null Hypothesis:

Two attributes of 'incidence of Flue' and administration of drug are independent i.e., proportion of persons affected by Flue are similar among two groups of individuals of those who used the drug and those who do not

$$H_0 = P_A = P_{NA}$$

$$H_1 = P_A \neq P_{NA}$$

Expected Frequencies:

Column	Row	$\frac{C \times R}{G}$	Expected Frequency
1	1	$\frac{450 \times 500}{900}$	= 250
1	2	$\frac{450 \times 400}{900}$	= 200
2	1	$\frac{450 \times 500}{900}$	= 250
2	2	$\frac{400 \times 450}{900}$	= 200

Calculation of Chi - Square:

Column	Row	O	E	$(O-E)^2$	$(O-E)^2/E$
1	1	200	250	2500	10.0
1	2	250	200	2500	12.5
2	1	300	250	2500	10.0
2	2	150	200	2500	12.0
					$X^2 = 45 \cdot 5$

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

$$= 45.0$$

For qa contingency table of r rows and c columns, the chi - square test follows (r - 1) (c - 1) degrees of freedom. Assuming a level of significance of 5% the critical X^2 value for $(2-1)\ (2-1)$ degrees of freedom is 3.841. The calculated X^2 value is higher than the critical table value. Therefore, it rejects the null hypothesis of equal proportion of individuals getting effected by Flue with or without using the drug. In other words, there exists dependency between the drug and incidence of Flue and thus, the drug is found to be effective.

13.7 Test of Goodness of Fit:

Statistical theoretical distribution have a great role in decision making, especially during the periods of uncertainty. We can test whether Binomial, Poisson or Normal Distributions are appropriate or not in such situation.

Suitability of a distribution:

In order to calculate expected frequencies and compare them with observed ones, an appropriate fit of the hypothesised distribution be made for the given attributes of samples. To illustrate the testing of goodness of fit, let us consider the following examples.

Example 5: Testing the fitness to an assumed ratio:

A garment manufacturing firm manufactures men's pyjamas in three sizes, small, medium and large. The firm's production has been tuned to produce these three sizes of pyjamas on the ratio of 1 : 4 : 3. The marketing consultant wishes to test the above pattern and considers the measures from 50 male customers and finds that numbers requiring small, medium and large pyjamas are 4, 32, 14 respectively. The consultant wishes to determine whether the results are consistent at 5% level of significant, with the firm's practices.

(a) Hypothesised Model : 1 : 4 : 3

(b) Observed Frequencies :

		Expected Frequencies
Small	4	6.25
Medium	32	25.00
Large	14	18.75
	50	50.00

(c) Calculation of X^2 : $X^2 = \sum \frac{(O - E)^2}{E}$

Modes	O	E	$(O-E)^2$	$(O-E)^2/E$
Small	4	6.25	50.0625	0.81
Medium	32	25.00	49.000	1.96
Large	14	18.75	22.563	1.20
				$X^2 = 3 \cdot 97$

(d) Decision:

The critical value of X^2 at 5% level of significance with (K - 1) degrees of freedom is 5.991. The calculated X^2 value is less than the critical one. Therefore, we accept the null hypothesis. In other words the norms being followed by unit are correct.

Example 6: Testing Binomial fit to defectives in manufacture:

A tractor manufacturing firm expects that the existing process is resulting in 10% of all tractors coming off the assembly line requires adjustments. With a view to revalidate this a quality control engineer has been considering daily samples and tractors coming of the assembly line for 300 consecutive working days and obtains the following data.

Number requiring Adjustments	0	1	2	3
No. of days	151	119	28	2

Let us fit a binomial distribution to the data and test how far it is a better fit. An analysis of such fit would enable management to make approx. Inferences even without completing the production in assembly line:

(a) Hypothesis:

H_0 : The data fts well in Binomial distribution

H_1 : The data does not fit into Binomial distribution

(b) Calculation of Binomial Frequencies:

$= {}^nC_r \; P^r \; q^{n-r} \times N$

Where n = sample size

r = no. of defectives

p = probability of defectives

q = 1 - p

N = Total number of days observed.

Following the Binomial probability given in tables, the frequencies are as follows:

Number requiring Adjustment out of 4	Binomial probability Density Function Where P = 0, 10, n = 4 r = 0, 1, 2, 3, ...	Probability	Expected Frequency
0	${}^nC_0P^0Q^4$	0.6561	196.8
1	${}^nC_1P^1Q^3$	0.2916	87.5
2	${}^nC_2P^2Q^2$	0.0486	14.6
3	${}^nC_3P^3Q^1$	0.0036	1.1
			300.0

(c) Calculation of X^2:

Number requiring Adjustment	Observed No.of days O	Expected No.of Days E	$(O-E)^2$	$(O-E)^2/E$
0	151	196.8	2097.64	10.66
1	119	87.5	992.25	11.34
2	28	14.6	179.56	12.30
3	2	1.1	0.81	0.74
				$X^2 = 35 \times 04$

The critical value of X^2 at 5% level of significance and for (K - 1) degrees of freedom is 7.815. The calculated value exceeds this value and it rejects the null hypothesis of suitability of Binomial distribution to the given data. In other words, the data is fit for some other distribution rather than the Binomial Distribution.

Example 7: The following data give the number of aircraft acciolent that occurred during the various days of a week

Day	Mon	Tues	Wed	Thu	Fri	Sat
No.of. Accidents	15	19	13	12	16	15

Test whether the accidents are uniformly distributed over the week.

Solution:

H_0 : Accidents occur uniformly over the week.

Total number of accidents = 90

Based on H_0 the expected number of accidents on any day $= \frac{90}{6} = 15$

O_i :	15	19	13	12	16	15
E_i :	15	15	15	15	15	15

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{1}{15}(0 + 16 + 4 + 9 + 1 + 0) = 2$$

Since $\sum E_i = \sum O_i$, $\nu = 6 - 1 = 5$

From the χ^2 table, $\chi^2_{0.05}(\nu = 5) = 11 \cdot 07$

Since $\chi^2 < \chi^2_{0.05}$, H_0 is accepted. That is accidents may be regarded to occur uniformly over the week .

Example 8: The following data show defective articles produced by 4 machines:

Machine	:	A	B	C	D
Production Time	:	1	1	2	3
No.of. Defectives	:	12	30	63	98

Do the figures indicate a significant difference in the performance of the machines?

Solution:

H_0 : Production rates of the machines are the same.

Total number of defectives = 203

Based on H_0, the expected numbers of defectives produced by the machines are

O	E (rounded E)	$(O - E)^2/E$
1154	$\frac{1872 \times 2257}{3759} = 1124$	$30^2/1124 = 0 \cdot 80$
475	$\frac{1872 \times 917}{3759} = 457$	$18^2/457 = 0 \cdot 71$
243	$\frac{1872 \times 585}{3759} = 291$	$48^2/291 = 7 \cdot 92$
1103	$\frac{1887 \times 2257}{3759} = 1133$	$30^2/1133 = 0 \cdot 79$
442	$\frac{1887 \times 917}{3759} = 460$	$18^2/460 = 0 \cdot 70$
342	$\frac{1887 \times 585}{3759} = 294$	$48^2/294 = 7 \cdot 84$
	$\nu = (3-2)(2-1) = 2$	$\chi^2 = 18 \cdot 76$

From the χ^2 - table, $\chi^2_{0 \cdot 05}(\nu = 2) = 5 \cdot 99$

Since $\chi^2 > \chi^2_{0 \cdot 05}$, H_0 is rejected. Therefore sex and attitude are not independent i.e., there is some association between sex and attitude.

13.8 Summary:

In this lesson, we have explained the concept of χ^2 - test and also given caculation procedure. Further, we have provided applications of χ^2 analysis.

13.9 Exercises:

1. What are the characteristics of Chi - Square distribution? What are its uses?
2. Explain how chi - square distribution can be used for judging the agreement between a hypothetical and an observed distribution.
3. What is the purpose of a test of goodness of fit? Find the applicability of Chi - Square as test statistic for goodness of fit in different situations.
4. The administrator of Grindley's Bank wants to determine whether the number of transactions in Teller's counter is greater in one day of the week than the other. The records of past six months yields the following information.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
33	28	24	22	30	19

Test whether the number of transactions are uniformly distributed or not.

5. Tata company services is interested in analysing the movement of different brands of a consumer good in which their client is engaged in. They have surveyed four leading departmental stores in metropolitan area and enquired about the sale of four popular brands of the product. The data are

Number items sold in a month

Brand A	Brand B	Brand C	Brand D
45	62	47	67
61	27	72	63
72	55	49	68
35	57	49	59

Test whether sale of units differs from brand to brand.

6. A T.V. advertisement copy producer is bringing out a new exposure to his customer. In order to map out his advertisement at appropriate T.V. relay timing, he wants to determine which age group viewers would be attracted to new exposure much. Responses from respondents attended the Pre - reviewing show are collected and arranged according to their age groups.

	Age Groups			
Persons	**Under 25**	**25 - 35**	**35 - 50**	**50 above**
Liked the Exposure	60	50	50	30
Do not like the Exposure	15	10	20	10
Indifferent	10	5	15	20

Test whether the liking and disliking the new advertisement exposure by individuals is independent of the age group.

7. Kinetic Engineering Ltd. has introduced two different models of 50 c.c. mopeds viz., Luna Super and Luna Magnam. The performance executive claims that both the models would equally perform in terms of mileage expecting the fact that the later type is a sturdier one. However the experience of 30 customers of each vehicle provides the following summary statistics.

	Mileage in km	
	Luna Super	**Luna Magnam**
Mean	55	60
Standard Deviation	20	15

Test whether the difference in variances are significant at 5% level.

8. In 250 digits from the lottery numbers, the frequencies of the digits were as follows:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency:	23	25	20	23	23	22	29	25	33	27

Test the hypothesis that the digits were randomly drawn.

9. In 120 throws of a single die, the following distribution of faces are obtained:

Face :	1	2	3	4	5	6
Frequency:	20	25	18	10	22	15

Do there results support the equal probability hypothesis.

10. A certain drug is claimed to be effective curing cold. In an experiment on 500 persons with cold, heat of they were given the drug and half of they were given the sugar pills. The patients reaction to the treatment are recorded in the following table:

	Helped	Harmed	No effect
Drug	150	30	70
Sugar Pills	130	40	80

On the basis of this data can it be concluded that the drug and sugar pills differ significantly in curing cold?

11. A survey of radio listener's preference for two types of music under various age group d gave the following information.

Type of music	Age group		
	19 - 25	26 - 35	above 35
Carnatic music	80	60	90
Film music	210	325	44
Indiff even	16	45	132

Is preference for type of music influenced by age?

13.10 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer

Dr. K. CHANDAN

Lesson - 14

ANALYSIS OF VARIANCE (ANOVA)

Objective:

After going through this lesson, you will learn:

- The concept of Analysis of variance computational of one way and two way classification.

Structure:

14.1 Introduction:

The Analysis of variance is a widdy used technique developed by R.A. Fisher. It enables us to divide the total variation (represented by variance) in a group into parts which are ascribable to different factors and a residual random variation which could not be accounted for by any of there factors. The variation due to any specific factor is compared with the residual variation for significance by applying the F-test, with which the reader is addceured to be familiar.

Analysis of variance:

Analysis of variance was introduced by Prof. R.A. Fisher in 1920's to deal with the problem in the analysis of africultural data. The basic purpose of analysis of variance is to test the homogeneity of several means.

According to Prof. R.A. Fisher analysis of variance is the separation of variance due to one group of causes from the separation of variance due to another group of causes.

The basic assumptions for the validity of F-test in ANOVA:

The following are the basic assumptions for the validity of F-test in ANOVA.

1. All the observations must be independent.
2. The parent population from which the samples have been drawn must be normal.
3. Variuos treatment and environment effects are additive in nature.

The difference between variability with in classes and variability between the classes:

Variations among the observations of each specific class are called its internal variation and the totality of the internal variation is called variability within the classes.

The totality of variation with in each class reflects chance variation under the assumption that the variation due to classes is equal to the total variation and is often called the experimental error. This variation is due to the control and non specific factors. The totality of variation from one class to another i.e. variation due to classes is called variability between classes.

For example let us consider a simple of 4 provinces and 10 sugar shops from each priovince. We note down the prices of sugar on these 40 shops. The variation between the prices of ten shops from each province is their internal variation and the totality of this internal variation is variability with in provinces. The variation between the 4 sample means is the variability between provinces.

The Utility and applications of ANOVA technique:

Utility:

The technique of studying the homogeneity of population by separating the total variation into its various components has a much wide scope. Now that was conceived by Prof. R.A. Fisher who used it in analyzing to agricultural data. Now - a - days it is used to handling the statistics of multiple groups in various other branches of study.

Applications:

The main application of ANOVA is to the homogeneity of the observations.

To test the significance of additional terms in a regression equation.

To test the curve linearity or no linearity of the fitted regression lines.

To test the significance in case of multiple regression.

14.2 One-Way Classification:

Let us purppose that N observations $x_{ij}, \left(i = 1,2,\cdots\cdots,k\ ;\ j = 1,2,\cdots\cdots\cdots,n_i\right)$ of a random variable x are grouped on some basis into k classes of sizes $n_1, n_2, \cdots\cdots\cdots, n_k$ respectively $\left(N = \sum_{i=1}^{k} n_i\right)$ as exhibited below:

		Mean	Total
x_{11}	$x_{12}, \cdots\cdots\cdots, x_{1x1}$	$\bar{x}_1$	T_1
x_{21}	$x_{22}, \cdots\cdots\cdots, x_{2x2}$	$\bar{x}_2$	T_2
.			
.			
.			
x_{i1}	$x_{i2} \cdots\cdots\cdots\cdots x_{ini}$	$\bar{x}_i$	T_i
.			.
.			.
.			.
x_{k1}	$x_{k2} \cdots\cdots\cdots\cdots x_{knk}$	$\bar{x}_k$	T_k

ANOVA Table for newly classification datat

Sources of Variation	Degrees of Freedom	Sum of Squares	mean sum of squares	E - Ratio
Treatment	f - 1	St^2	$st^2 = \frac{St^2}{k-1}$	$\frac{St^2}{se^2} = F_{(k-1)(N-k)}$
Error	N - K	SE^2	$se^2 = \frac{SE^2}{N-K}$	
Total	N-1	ST^2		

Conclusion:

If the calculated value of F is less than the tabulated value of F at $\alpha\%$ LOS then we accept our null hypothesis H_0 otherwise we reject H_0.

14.3 Problems - ANOVA One way classification:

Example 1: A test was given to 5 students taken at random from the 5^{th} class of 3 schools of a town. The individual scores are

School I	9	7	6	5	8
School II	7	4	5	4	5
School III	6	5	6	7	6

Carry out the analysis of variance and state your conclusion.

Solution:

For the given data, or null hypothesis is given by

H_0 : There is no significant difference between three schools.

And the alternative hypothesis is

H_1 : There is a significant difference between the schools.

						Total (T_i)	$\sum_i x_{ij}^2$
School I	9	7	6	5	8	35	255
School II	7	4	5	4	5	25	131
School III	6	5	6	7	6	30	182
						G = 90	$\sum_i \sum_j x_{ij}^2 = 568$

$$S_T^2 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 568 - \frac{90^2}{15}$$

$$= 568 - 540$$

$$= 28$$

$$S_{TR}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{35^2}{5} + \frac{25^2}{5} + \frac{30^2}{5} - \frac{90^2}{15}$$

$$= 245 + 125 + 180 - 540$$

$$= 550 - 540$$

$$= 10$$

$$S_E^2 = S_T^2 - S_{TR}^2$$

$$= 28 - 10$$

$$= 18$$

ANOVA TABLE				
Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Sum of Squares**	**F - Ratio**
Treatments	3 - 1 = 2	10	5	$F = \frac{5}{1 \cdot 5} = 3 \cdot 33 \sim F(2, 12)$
Errors	14 - 2 = 12	18	1.5	
Totals	15 - 1 = 14			

The tabulated value of F (2, 12) = 19.41 at 5% LOS

Conclusion:

Since the calculated value of F is less than the tabulated value of F at 5% level of significance, hence we accept H_0.

i.e., there is no significant difference between 3 schools.

Example 2: Three procesors A, B and C are tested to see whether their outputs are equivalent. The following observations of output are made.

A	10	12	13	11	10	14	15	13
B	9	11	10	12	12			
C	11	10	15	14	12	13		

Carry out the analysis of variance and state your conclusion.

Solution:

For the given data our null hypothesis is given by

H_0 : There is no significant difference between the three processors.

And the alternative hypothesis is given by

H_1 : There is a significant difference between the three processors.

Processors	**Observations**	Total (T_i)	$\sum_j x_{ij}^2$
A	10 12 13 11 10 14 15 13	98	1224
B	09 11 10 12 13	55	615
C	11 10 15 14 12 13	75	955
		G = 228	$\sum_i \sum_j x_{ij}^2 = 2794$

$$S_T^2 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 2794 - \frac{228^2}{19}$$

$$= 2794 - 2736$$

$$= 58$$

$$S_{TR}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{98^2}{8} + \frac{55^2}{5} + \frac{75^2}{6} - \frac{228^2}{19}$$

$$= 1200.5 + 605 + 937.5 - 2736$$

$$= 2743 - 2736$$

$$= 7$$

$$S_E^2 = S_T^2 - S_{TR}^2$$

$$= 58 - 7$$

$$= 51$$

ANOVA TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F - Ratio
Treatments	3 - 1 = 2	7	3.5	
Errors	18 - 2 = 16	51	3.1875	$F = \frac{3 \cdot 5}{3 \cdot 1875} = 1 \cdot 0980 \sim F(2, 16)$
Totals	19 - 1 = 18			

The tabulated value of F(2, 16) = 3.63 at 5% level of significance.

Conclusion:

Sincethe calculated value of F is less than the tabulated value of F at 5% level of significance, hence we accept our null hypothesis H_0.

i.e., there is no significant difference between the three processors A, B and C.

Example 3: Three varieties of coal were analyzed by four chemists and the ash content in the varieties were found to be

Varieties	Chemists			
	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Do the varieties differ significantly in their ash content?

Solution:

For the given data, our null hypothesis is

H_0 : There is no significant difference between three varieties of coal.

And the alternative hypothesis is

H_1 : There is a significant difference between three varieties of coal.

Varieties	Chemists					
	1	2	3	4	Total (T_i)	$\sum_j x_{ij}^2$
A	8	5	5	7	25	163
B	7	6	4	4	21	117
C	3	6	5	4	18	86
					G = 64	$\sum_i \sum_j x_{ij}^2 = 366$

$$S_T^2 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 366 - \frac{64^2}{12}$$

$$= 366 - 341 \cdot 3333$$

$$= 24.6667$$

$$S_{TR}^2 = \sum \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{25^2}{4} + \frac{21^2}{4} + \frac{18^2}{4} - \frac{64^2}{12}$$

$$= 156.25 + 110.25 + 81 - 341.3333$$

$$= 6.1667$$

$$S_E^2 = S_T^2 - S_{TR}^2$$

$$= 24.6667 - 6.1667$$

$$= 18.5$$

ANOVA TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F - Ratio
Treatments	3 - 1 = 2	6.1667	3.0834	
Errors	11 - 2 = 9	18.5	2.0556	$F = \frac{3\cdot0834}{2\cdot0556} = 1\cdot5 \sim F(2, 9)$
Totals	12 - 1 = 11			

The tabulated value of F (2, 9) = 4.26 at 5% level of significance

Conclusion:

Since the valculated value of 'F' is less than the tabulated value of 'F' at 5% level of significance hence we accept H_0.

i.e., There is no significant difference between the ash cotnent of three varieties of coal.

Example 4: The following data shows the lines in hours of four batches of electric lamps.

Batches								
1	1600	1610	1650	1680	1700	1720	1800	
2	1580	1640	1640	1700	1750			
3	1460	1550	1600	1620	1640	1663	1740	1820
4	1510	1520	1530	1570	1600	1680		

Perform an analysis of variance of these data and show that a significance test does not reject their homogeneity.

Solution:

For the given data our null hypothesis is

H_0 : There is no significant difference between the four batches

And the alternative hypothesis is

H_1 : There is a significant difference between four batches.

Since all the observations in the given data are very high for the sake of simplicity in the computation part we will use the technique of change of origin and scale.

Now shifting the origin to 1640 and then dividing by 10 the given data reduces to

Batches	**Observations**								**Total** (T_i)	$\sum_j x_{ij}^2$
1	-4	-3	1	4	6	8	16	-	28	398
2	-6	0	0	6	11	-	-	-	11	193
3	-18	-9	-4	-2	0	2	10	18	-3	853
4	-13	-12	-11	-7	-4	-4	-	-	- 43	515
									G = - 7	$\sum_i \sum_j x_{ij}^2 = 1959$

$$S_T^2 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 1959 - \frac{(-7)^2}{26}$$

$$= 1959 - 1{\cdot}8846$$

$$= 1957.1154$$

$$S_{TR}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{28^2}{7} + \frac{11^2}{5} + \frac{(-3)^2}{8} + \frac{(-43)^2}{6} - \frac{(-7)^2}{26}$$

= 112 + 24.2 + 1.125 + 308.1667 - 1.8846

= 443.6071

$$S_E^2 = S_T^2 - S_{TR}^2$$

= 1957.1154 - 443.6071

= 1513.5083

ANOVA TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F - Ratio
Treatments	4 - 1 = 3	443.6071	147.8690	$F = \frac{147 \cdot 8690}{68.7958} = 2 \cdot 1494 \sim F(3, 22)$
Errors	25 - 3 = 22	1513.5083	68.7958	
Totals	26 - 1 = 25			

The tabulated value of F(3, 22) = 3.05 at 5% level of significance.

Conclusion:

Since the calculated value of 'F' is less than the tabulated value of 'F' at 5% level of significance hence we accept our null hypothesis H_0.

i.e., There is no significant difference between the four batches.

14.4 ANOVA Two - Way Classification (With one observation per cell) Lay-Out:

Let us consider the case when there are two factos which may affect the variate values $x_{ij}, i = 1, 2, \cdots\cdots\cdots, k$ & $j = 1, 2, \cdots\cdots, h$. For example, the yield of milk may be affected by differences in treatments (rations) as well as the differences inv ariety i.e. breed and stock of the cows. Let us now suppose that 'N' cows are divided into 'k' different groups according to their breed and stock, each group contains 'h' cows and then let us consider the affect of 'k' treatmetns on the yield of milk $x_{ij}, i = 1, 2, \cdots\cdots, k$ & $j = 1, 2, \cdots\cdots, h$ of N = kh cows.

The yields may be expressed as variate values in the following k x h two way table.

Blocks Treatmetns	1	2	.....	j	.....	h	Total
1	x_{11}	x_{12}	⋯	x_{1j}	⋯	x_{1h}	T_1
2	x_{21}	x_{22}	⋯	x_{2j}	⋯	x_{2h}	T_2
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
i	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{ih}	T_i
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
k	x_{k1}	x_{k2}	⋯	x_{kj}	⋯	x_{kh}	T_k
Total	T_1	T_2	⋯	$T_{\cdot j}$	⋯	T_h	G

ANOVA TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F - Ratio
Treatments	k - 1	S^2_{TR}	S^2_{tr}	$F_{TR} = \frac{S^2_{tr}}{S^2_e}\left(S^2_{tr} > S^2_e\right) \sim F\left[k-1, (k-1)(h-1)\right]$
Blocks	h - 1	S^2_B	S^2_b	
Errors	$(k-1)(h-1)$	S^2_E	S^2_e	$F_B = \frac{S^2_b}{S^2_e}\left(S^2_b > S^2_e\right) \sim F\left[j-1, (k-1)(h-1)\right]$
Totals	N - 1			

Conclusion:

If the calculated values of 'F' are less than the tabulated values of 'F' at α % LOS then we accept ournull hypotheses otherwise we reject our null hypotheses.

14.5 Problems - Two Way Classification:

1. The following table gives quality rating of 10 service stations given by five professional raters.

Raters	Service Stations									
	1	2	3	4	5	6	7	8	9	10
A	99	70	90	99	65	85	75	70	85	92
B	96	65	80	95	70	88	70	51	84	91
C	95	60	48	87	48	75	71	93	80	93
D	98	65	70	95	67	82	73	94	86	80
E	97	65	62	99	60	80	76	92	90	89

Analyze the data and discuss whether there is any significant difference between ratings or between service stations.

Solution:

For the given data, our null hypothesis is defined as

H_0 : There is no significant difference between ratings as well as service stations

And the alternative hypothesis is

H_1 : There is a significant difference between ratings as well as service stations.

Raters	Service Stations										Total (T_i)	$\sum_j x_{ij}^2$
	1	2	3	4	5	6	7	8	9	10		
A	99	70	90	99	65	85	75	70	85	92	830	70266
B	96	65	80	95	70	88	70	51	84	91	790	64348
C	95	60	48	87	48	75	71	93	80	93	750	59166
D	98	65	70	95	67	82	73	94	86	80	810	66928
E	97	65	62	99	60	80	76	92	90	89	810	67540
Total $(T_{.j})$	485	325	350	475	310	410	365	400	425	445	G = 3990	$\sum_i \sum_j x_{ij}^2$ = 328248

$$S_T^2 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 328248 - \frac{3990^2}{50}$$

= 328248 - 318402

= 9846

$$S_{TR}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{830^2}{10} + \frac{790^2}{10} + \frac{750^2}{10} + \frac{810^2}{10} + \frac{810^2}{10} - \frac{3990^2}{50}$$

= 368

$$S_B^2 = \sum_i \frac{T_{\cdot j}^2}{n_{\cdot j}} - \frac{G^2}{N}$$

$$= \frac{485^2}{5} + \frac{325^2}{5} + \frac{350^2}{5} + \frac{475^2}{5} + \frac{310^2}{5} + \frac{410^2}{5} + \frac{365^2}{5} + \frac{400^2}{5} + \frac{425^2}{5} + \frac{445^2}{5} - \frac{3990^2}{50}$$

= 6608

$$S_E^2 = S_T^2 - S_{TR}^2 - S_B^2$$

= 9846 - 368 - 6608

= 2870

ANOVA TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F - Ratio
Treatments	4	368	92	$F_{TR} = \frac{92}{79 \cdot 7222} = 1 \cdot 1540 \sim F(4, 36)$
Blocks	9	6608	734.2222	
Errors	36	2870	79.7222	$F_B = \frac{734 \cdot 2222}{79 \cdot 7222} = 9 \cdot 2097 \sim F(9, 36)$
Totals	49			

The tabulated value of F (4, 36) is 2.63

The tabulated value of F(9, 36) is 2.15 at 5% level of significance.

Conclusion:

Since the calculated value of 'F' for treatmetns is less than the tabulated value of 'F' hence we accept our null hypothesis for treatments.

i.e, There is no significant difference between the ratings given be five professional raters.

Since the calculated value of 'F' for blocks is greater than the tabulated value of 'F' hence we reject our null hypothesis for blocks.

i.e, There is significant difference between ten service stations.

Example 6: Perform the analysis of variance for the following data with suitable technique.

Observer	**Consignment**					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	2	11	11	14	12	10

Solution:

For the given data, our null hypothesis is defined as

H_0 : There is no significant difference between observers as well as consignments.

And the alternative hypothesis is

H_1 : There is a significance difference between observers as well as consignmetns.

Observers	**Consignment**							
	1	**2**	**3**	**4**	**5**	**6**	**Total** (T_i)	$\sum_j x_{ij}^2$
1	9	10	9	10	11	11	60	604
2	12	11	9	11	10	10	63	667
3	11	10	10	12	11	10	64	686
4	2	11	11	14	12	10	60	686
Total $(T_{.j})$	34	42	39	47	44	41	G = 247	$\sum_i \sum_j x_{ij}^2 = 2643$

$$S_T^2 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 2643 - \frac{247^2}{24}$$

$$= 2643 - 2542.0417$$

$$= 100.9583$$

$$S_{TR}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{60^2}{6} + \frac{63^2}{6} + \frac{64^2}{6} + \frac{60^2}{6} - \frac{247^2}{24}$$

$$= 600 + 661.5 + 628.6667 + 600 - 2542.0417$$

$$= 2.125$$

$$S_B^2 = \sum_i \frac{T_{\cdot j}^2}{n_{\cdot j}} - \frac{G^2}{N}$$

$$= \frac{34^2}{4} + \frac{42^2}{4} + \frac{39^2}{4} + \frac{47^2}{4} + \frac{44^2}{4} + \frac{41^2}{4} - \frac{247^2}{24}$$

$$= 289 + 441 + 380.25 + 552.25 + 484 + 420.25 - 2542.0417$$

$$= 24.7083$$

$$S_E^2 = S_T^2 - S_{TR}^2 - S_B^2$$

$$= 100.9583 - 2.125 - 24.7083$$

$$= 74.125$$

ANOVA TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F - Ratio
Treatments	3	2.125	0.7083	$F_{TR} = \frac{4\cdot9417}{0\cdot7083} = 6\cdot9768 \sim F(15, 3)$
Blocks	5	24.7083	4.9417	
Errors	15	74.125	4.9417	$F_B = \frac{4\cdot9417}{4\cdot9417} = 1 \sim F(15, 5)$
Totals	23			

The tabulated value of F (15, 3) is 8.70

The tabulated value of F(15, 5) is 4.62 at 5% level of significance.

Conclusion:

Since the calculated values of 'F' for both treatments and blocks are less than the tabulated values of 'F' at 5% level of significance, hence we accept our null hypothesis.

i.e., There is no significant difference between the observers as well as consignments.

Example 7: Perform analysis of variance with a suitable technique for the following data and comment on your conclusions.

Varieties	Chemists			
	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Solution:

For the given data or null hypothesis is defined as

H_0 : There is no significant difference between the varieties as well as chemists.

And the alternative hypothesis is

H_1 : There is a significant difference between the varieties as well as chemists.

Varieties	Chemists				Total (T_i)	$\sum_j x_{ij}^2$
	1	2	3	4		
A	8	5	5	7	25	163
B	7	6	4	4	21	117
C	3	6	5	4	18	86
$T_{.j}$	18	17	14	15	G = 64	$\sum_i \sum_j x_{ij}^2 = 366$

$$S_T^2 = \sum_i \sum_j x_{ij}^2 - \frac{G^2}{N}$$

$$= 366 - \frac{64^2}{12}$$

$$= 366 - 341.3333$$

$$= 24.6667$$

$$S_{TR}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{G^2}{N}$$

$$= \frac{25^2}{4} + \frac{21^2}{4} + \frac{18^2}{4} - \frac{64^2}{12}$$

$$= 156 \cdot 25 + 110 \cdot 25 + 81 - 341 \cdot 333 = 6 \cdot 1667$$

$$S_B^2 = \sum_j \frac{T_j^2}{n_j} - \frac{G^2}{N}$$

$$= \frac{18^2}{3} + \frac{17^2}{3} + \frac{14^2}{3} + \frac{15^2}{3} - \frac{64^2}{12}$$

$$= 108 + 96.3333 + 65.3333 + 75 - 341.3333$$

$$= 3.3333$$

$$S_E^2 = S_T^2 - S_{TR}^2 - S_B^2$$

$$= 24.6667 - 6.1667 - 3.3333$$

$$= 15.1667$$

ANOVA TABLE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F - Ratio
Treatments	2	6.1667	3.0834	$F_{TR} = \frac{3\cdot0834}{2\cdot5278} = 1\cdot2198 \sim F(2, 6)$
Blocks	3	3.3333	1.1111	
Errors	6	15.1667	2.5278	$F_B = \frac{2\cdot5278}{1\cdot1111} = 2\cdot2750 \sim F(6, 3)$
Totals	11			

The tabulated value of F(2, 6) = 5.14

The tabulated value of F(6, 3) = 8.94 at 5% level of significance

Conclusion:

Since the calculated values of 'F' for both the treatments and blocks are less than the tabulated values of 'F' at 5% level of significnce hence we accept our null hypothesis.

i.e., There is no significant difference between the varieties as well as chemists.

Example 8: Apply Two-way ANOVA on the row means 39, 41, 48 and the column means 38, 21, 69, 42 with 3 rows and 4 columns and also given $\sum_i \sum_j x_{ij}^2 = 25,944$.

Solution:

For the given data, null hypothesis is given by

H_0 : There is no significant difference between 3 rows as well as 4 columns.

And the alternative hypothesis is given by

H_1 : There is a significant difference between 3 rows as well as 4 columns.

Columns/Rows	1	2	3	4	Means	Total (T_i)
1					39	156
2					41	164
3					48	192
Means	38	21	69	42		512
Total $(T_{.j})$	114	63	207	126	510	

Since the grand total for rows is not equal to the grand total for columns from the above table, hence the given information is not correct.

14.6 Summary:

In this lesson, we have explained the concept of Analysis varience and also computational of one way and two way classification.

14.7 Exercises:

1. Explain ANOVA and compare the ANOVA one way classification with that of two way classification with the assumptions.
2. Explain what do you understand by analysis of variance and give the ANOVA one-way classification.
3. Writethe statistical analysis of one-way classification of ANOVA.
4. The following data refers to the output of three machines of the same make by each of the four operators. Perform the analysis of variance for one way classification and find whether the operators produce differently or can the variation between the operator's out-put be attributed to chance errors ?

	Operators			
	A	B	C	D
Machine 1	174	173	171.5	173.5
Machine 2	173	172	171	171
Machine 3	173.5	173	173	172.5

5. State the mathematical model used in analysis of variance is a two-way classification. State the basic assumptions. Discuss the advantages of this method over one - way classification if any.
6. Write about the analysis variance for one way classifications mathematical model, and assumptions of the model.
7. Write the statistical analysis of one - way classification.

8. Describe the technique of ANOVA write down the ANOVA table for one-way layout.

9. The following data obtained from a completely randomized design with four treatmetns analysse the given data and draw conclusion about the equality of treatmetn effects.

Treatments

T_1	T_2	T_3	T_4
20.9	23.7	13.2	5.8
12.4	14.4	10.2	6.1
10.1	9.0	5.1	4.8
4.2			1.5

14.8 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer

Dr. K. CHANDAN

Lesson - 15

THEORY OF PROBABILITY

Objective:

After going through this lesson, you will learn:

- The concept of probability.
- To estimate the probability under different approaches.
- To learn the rules of probability when events are independent and iterdependent.

Structure:

15.1 Introduction:

Probability is a part of every day life. In personal and managerial decisions, we face uncertainty and we use probability theory whether or not we admit the use of something so sophisticated. Managers who deal with inventories of High Styled Women's Clothing, must wonder about the chances that sales will reach or exceed a certain level.

15.2 What is Probability:

It is difficult to give a precise definition of probability. In simple terms 'probability means the number of occasions that a particular event is likely to occur in a large proportion of events'. From a mathematical standpoint, probability is nothing more than an assignment of numbers to the subsets of a universal set. The probability of a particular outcome is regarded as the long - run relative frequency of that outcome over many repetitions of the experiment.

15.3 Certain Technical Terms:

Event:

In probability theory, an event is one or more of the possible outcomes of doing something. If we toss a coin getting a 'tail' would be an event, and getting a 'head' would be another event. Similarly, if we are drawing from a 'deck of cards' selecting the 'ace of spades' would be an event.

Random Experiment of Trial:

A random experiment or trial is one which when conducted repeatedly under essentially homogeneous conditions, the result is not unique but may be any one of the various possible out comes.

For Example:

- Tossing a fair coin is an experiment.
- Rolling an unbiased dice is an experiment.
- Drawing a card from a pack of cards etc.

The term random is being used here to note the unbiasedness.

Sample Spaces:

The activity that produces an event is referred to in probability theory, as an experiment. Using this formal language we could ask the question, "In a coin toss experiment, what is the probability of the event Head?" And, of course, if it is a fair coin with an equal chance, the answer would be $1/2$. The set of all possible outcomes of an experiment is called the sample space for the experiment.

In a coin - toss experiment, the sample space is

$$S = \{\text{Head}, \text{Tail}\}$$

Mutually Exclusive Events:

Events are said to be mutually exclusive if one and only one of them can take place at a time. It means that the happening of any one of them excludes the happening of all others in the same experiment.

For example, in the toss of a coin the events hed or tail are mutually exclusivne because if head comes, we cannot get tail and if tail comes, we cannot get head.

Similarly in the throw of a dice, the six faces numbered 1, 2, 3, 4, 5 and 6 are mutually exclusive. Thus events are said to be mutually exclusive, if no two or more of them can happen simultaneously.

Collectively Exhaustive Cases:

The total number of possible outcomes of a random experiment is called 'collectively exhaustive cases' for the experiment. For example, in toss of a single coin, exhaustive number of cases is 2, in a throw of a dice exhaustive number of cases is 6 and in case of throw of two dice, exhaustive number of cases are $6^2 = 36$.

Equally likely cases:

The outcomes are said to be equally likely or equally probable if none of them are expected to occur in preference to the other. In tossing of a coin, all the outcomes of head or tail are equally likely if the coin is not biased.

15.4 Types of Probability and Approach to Estimate:

There are two broad ways in grouping the probability. One is objective probability and another is subjective probability.

15.4.1 Objective Probability:

The objective probability is based on certain laws of nature which are undisputed. This is not based on the 'impressions' as in the case of subjective probability. The basic approaches for estimating the objective probability are as follows:

15.4.1.1 Classical Approach:

The classical approach is based on certain prior laws of nature. If random experiment results in 'n' exhaustive, mutually exclusive and equally likely outcomes (cases) out of which 'm' are favourable to the happening of an event A, then the probability of occurrence of A, usually denoted by P(A) is given by

$$\text{Probability of an event} = \frac{\text{Number of outcomes facourable to the events}}{\text{Total number of possible outcomes}}$$

(or)

$$P(A) = \frac{m}{n}$$

Probability, lies between "Zero" and '1'. If P(A) = 0, then (A) is called an 'impossible event' and if P(A) = 1, then A is called an event which is certain.

Thus the classical definition of probability does not require the actual experimentation, i.e., no experimental data are needed for its computation nor it is based on previous experience. It enables us to obtain probability by logical reasoning even without conducting actual trials and hence it is called as "a prior" or mathematical probability.

Example 1:

(a) What is the probability of getting head in a throw of a coin?

(b) If two coins are tossed once, what is the probability of getting

(i) both heads?

(ii) at least one head?

Solution: (a) When one coin is tossed, the outcomes are two

Head or Tail

n = 2

Outcome 'head' is favourable (m) = 1

Hence, P (Head) = 1/2

(b) In case of two coins possible outcomes are

(i) Two Heads (H , H)

(ii) One Head, One Tail (H, T)

(iii) One Tail in one coin and Head in second coin (T, H)

(iv) Two Tails (T, T)

So 'n' = 4

Favourable cases are both Heads in only one case (m = 1)

Hence P (H, H) = 1/4

(c) In case H, H, H, T, T, H we get at least one head

P (at least one head) = 3/4

or = 1 - 1/4

Example 2:

If two dices are thrown

(i) what is probability of throwing two sixes?

(ii) What is the probability of throwing a total of 9?

(iii) What is the probability of not throwing a total of 9?

Solution:

Total number of outcomes in a throw of two dice are 36

1,1	2,1	3,1	4,1	5,1	6,1
1,2	2,2	3,2	4,2	5,2	6,2
1,3	2,3	3,3	4,3	5,3	6,3
1,4	2,4	3,4	4,4	5,4	6,4
1,5	2,5	3,5	4,5	5,5	6,5
1,6	2,6	3,6	4,6	5,6	6,6

(i) In a throw of two dices 6, 6 will come only once.

Total No.of outcomes are = 36

P (getting 6, 6) = 1/36

(ii) In a throw of two dices a total of 9 comes

$$\begin{bmatrix} 3, & 6 \\ 4, & 5 \\ 5, & 4 \\ 6, & 3 \end{bmatrix}$$

Hence,

P (of getting a total of 9) = 4/36 = 1/9

(iii) Probability of not getting a total of 9

$$= 1 - \frac{1}{9} = \frac{8}{9} \quad (\text{or}) \quad \frac{32}{36} = \frac{8}{9}$$

Example 3:

A ball is drawn at random from a box containing 6 white, 8 red and 10 green balls. Determine the probability of a ball drawn. (i0 white, (ii) red, (iii) green, (iv) not red, (V) red or green

Solution:

Total number of balls = 6 + 8 + 10 = 24

n = 24

(i) Probability of drawing a white ball = 6/24 = 1/4

(ii) Probability of drawing a red ball = 8/24 = 1/3

(iii) Probability of drawing a green ball = 10/24 = 5/12

(iv) Probability of drawing a ball which is not red = 16/24 = 2/3

(V) Probability of drawing red or green ball

$$\frac{8+10}{24} = \frac{18}{24} = \frac{6}{8} = \frac{3}{4}$$

Example 4:

A card is drawn from a pack of cards. Find the probability of it being (a) red card (b) a club (c) an ace (d) a red ace (e) ace of spades (f) not a spade and (g) a king or a queen.

Solution:

(a) As there are 28 red cards, probability of a red card is

$$P(\text{Red Card}) = \frac{28}{56} = \frac{1}{2}$$

(b) There are 13 clubs in a pack (m = 13)

$$P(\text{Club}) = \frac{13}{52} = \frac{1}{4}$$

(c) There are four aces in all, one in each suit

$$P(\text{Ace}) = \frac{4}{52} = \frac{1}{13}$$

(d) The ace of diamonds and the ace of hearts are two red aces

$$P(\text{Red Ace}) = \frac{2}{52} = \frac{1}{26}$$

(e) The ace of spade is only 1 card

$$P(\text{Red Card}) = \frac{1}{56}$$

(f) In 52 cards, 13 ace spades and the remaining 39 are not spade (m = 39)

$$P(\text{Ace}) = \frac{39}{52} = \frac{3}{4}$$

(or) $$= 1 - \frac{1}{4} = \frac{3}{4}$$

(g) There are 4 kings and 4 queens and we may select any one of the 8 cards to get a king or a queen. Therefore m = 8.

$$P(\text{a King or Queen}) = \frac{8}{52} = \frac{2}{13}$$

15.4.1.2 Empirical Approach (or) Relative Frequency of Occurrence Approach:

Under this approach the probability of an event represents the proportion of times under identical circumstances, the outcome can be expected to occur. The value is the relative frequency of occurrence.

If the experiment is repeated for a large number of times under essentially identical conditions, the limiting value of the ratio of the number of times the event 'A' happens, is called the probability of the occurrence of "A".

Example 5:

The following data relates to the length of life of wholesale grocers in a particular city.

Length of Life (Yrs)	Percentage of wholesalers
0 - 5	65
5 - 10	16
10 - 15	9
15 - 25	5
25 and above	5
	100

(i) During the period studies, what is the probability that an entrant to this profession will fail within five years?

(ii) That he will survive at least 25 years?

(iii) How many years would he have to survive to be among 10 percent longest survivers?

Solution:

Total number of cases = 100

(i) The number of favourable cases to the condition that an entrant to the profession will fail within five years = 65. Hence

P (Entrant will fail in the profession within 5 years)

$$= \frac{65}{100} = 0 \cdot 65$$

(ii) The number of favourable cases to the condition of above 25 years of life is 5 and hence

P (Survival after 25 years) $= \frac{5}{100} = 0 \cdot 05$

(iii) The number of favourable cases for a wholesaler with more than 15 years age is 5 + 5 = 10, hence

$$\text{P (Survival after 15 years)} = \frac{10}{100} = 0 \cdot 01$$

Example 6:

The following table gives a distribution of monthly wages of 2000 employees of a firm.

Wages (in Rs.)	No. of Workers
Below - 280	18
280 - 320	236
320 - 360	956
360 - 400	400
400 - 440	284
440 - 480	70
480 - above	36

If an individual is selected at random from the above groups, what is the probability that his wages are (i) under Rs. 320? (ii) above Rs. 400? and (iii) between Rs. 320 and Rs. 400?

Solution:

(i) Total number of wage earners = 2000

Total wage earners below Rs. 320 = 18 + 236 = 254

$$\text{P (Individual selected is under Rs. 320)} = \frac{254}{2000} = \frac{127}{1000}$$

(ii) No.of wage earners earning wages of Rs. 400 and above per month

= 284 + 70 + 36 = 390

$$\text{P (individual getting wages above Rs. 400)} = \frac{390}{2000} = \frac{39}{200}$$

(iii) No.of wage earners in the wage group of Rs. 320 of Rs. 400 are $956 + 400 = 1356$.

$$\text{P (individual in the range of Rs. 320 to 400)} = \frac{1356}{2000} = \frac{339}{500}$$

15.4.2 Subjective Probability:

Subjective probabilities are based on the beliefs of the persons making the probability assessment. In fact, subjective probabilities can be defined as the probabilities assigned to an event by an individual, based on whatever evidence is available. This evidence may be in the form of relative frequency of past occurences, or it may be just a guess work.

15.5 Concepts of Permutations, Combinations And Their Use In Probability:

Permutations:

Permutations refer to separate arrangement of different objects contained in a set of elements. For example, if seven alphabets A, B, C, D, E, F, G are to be arranged by taking two letters at a time without containing same letter (like AA, BB etc...) the following permutations are possible.

AB	AC	AD	AE	AF	AG
BA	BC	BD	BE	BF	BG
CA	CB	CD	CE	CF	CG
DA	DB	DC	DE	DF	DG
EA	EB	EC	ED	EF	EG
FA	FB	FC	FD	FE	FG
GA	GB	GC	GD	GE	GF

Hencethere are 7 x 6 = 42 permutations.

The number of different permutations of 'n' different objects taken 'r' at a time without repetition is:

$$^{n}P_{r} = n\left[(n-1)(n-2)\cdots\cdots(n-r+1)\right]$$

For Example:

$$^{3}P_{2} = 3\times 2 \qquad = \text{ 6 ways}$$

$$^{4}P_{3} = 4\times 3\times 2 \qquad = \text{ 24 ways}$$

$$^{7}P_{2} = 7\times 6 \qquad = \text{ 24 ways}$$

In terms of factorial symbols:

$$^{n}P_{r} = \frac{n\times(n-1)\times(n-2)\cdot\ \cdot\ \cdot(n-r+1)(n-r)}{(n-r)!}$$

$$= \frac{n!}{(n-r)!}$$

Note: The product of first n natural numbers viz., 1, 2, 3, ...n is called factorial n or n factorial and is written as n! or Ln.

Combinations:

A Combination is a selection of objects considered without regard to their arrangements. in a permutation the order of the grouped items is important : in Combination the order does not matter. Combinations are arrangements of items wherein order is not important and duplication of components is in admissible.

For example, if letters, A, B, C, D, E are to be arranged in rows, but the same letters are not to be used at the same time.

The permutaions will be $n \text{ X } (n-1) \cdot \cdot \cdot = 5 \times 4 = 20$

AB	AC	AD	AE
BA	BC	BD	BE
CA	CB	CD	CE
DA	DB	DC	DE
EA	EB	EC	ED

In case of combinations

AB	AC	AD	AE
BC	BD	BE	
CD	CE		
DE			

A combination of 'n' different objects taken 'r' at a time, denoted by ${}^{n}C_{r}$ or $\binom{n}{r}$ a selection of only 'r' objects out of 'n' objects, without any regard to the order of arrangements.

$${}^{n}C_{r} = \frac{n!}{r!(n-r)!} = \frac{5!}{2!(5-2)!} = \frac{5\times4\times3\times2\times1}{2\times1\times3\times2\times1} = 10 \qquad \text{(or)}$$

$${}^{n}C_{r} = \frac{{}^{n}P_{r}}{r!} = \frac{20}{2\times1} = 10$$

Example 7:

Out of 4 officers and 10 clerks in a business firm, a committee consisting of 2 officers and 3 clerks is to be formed. In how many ways can this be done if.

(a) any officer and any clerk can be included

(b) one particular clerk must be on the committee

(c) two particular officers cannot be on the committee

Solution:

(a) 2 officers out of 4 can be selected in:

$= {}^{4}C_{2}$ ways

3 clerks out of 10 can be selected at

$= {}^{10}C_{3}$ ways

Total number of possible selections

$= {}^{4}C_{2} \times {}^{10}C_{3} = 720$

(b) One particular clerk must be in the selection committee

2 officers out of 4 can be selected

$= {}^{4}C_{2}$ ways

Out of the '9' remaining clerks

2 additional clerks out of 9 can be selected

$= {}^{3}C_{2}$ ways

Total number of possible selections

$= {}^{4}C_{2} \times {}^{3}C_{2} = 120$ ways

Example 8:

A bag contains 4 white, 5 red and 6 freen balls. Three balls are drawn at random. What is the chance that a red and a green balls are drawn?

Solution:

Total number of balls in a bag

$= 4 + 5 + 6 = 15$ balls

3 balls can be drawn out of 15 in

$= {}^{5}C_{3}$ ways

Posibility of drawing a white ball $= {}^{4}C_{1}$

Posibility of drawing a red ball $= {}^{5}C_{1}$

Posibility of drawing a green ball = ${}^{6}C_{1}$

The total number of favourable cases for drawing three balls subsequently is

$${}^{4}C_{1} \times {}^{5}C_{1} \times {}^{6}C_{1}$$

P (drawing red and green balls in three draws) = $\frac{{}^{4}C_{1} \times {}^{5}C_{1} \times {}^{6}C_{1}}{{}^{15}C_{3}}$

15.6 Theory of Sets : Rules of Probability:

a) Elementary Event:

Each one of the possible outcomes in a single trial or experiment is called an "Elementary Event".

Eg:

1. In an experiment of possing a coin elementary events - head or tail.
2. In case of throw of a dice the possible elementary events are 1, 2, 3, 4, 5 and 6.

b) Sample Space:

A set representing all possible outcomes of a random experiment is called the Sample Space. The outcomes of the experiments are also known as same points.

$$S = \{e_1, e_2, e_3, \cdots, e_n\}$$

Eg:

When a coin and dice are tossed together, the total number of sample points and sample space are as follows:

$$S = \begin{Bmatrix} (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6) \\ (H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6) \end{Bmatrix}$$

$$n(S) = 12$$

(c) Event Set:

If all the possible outcomes in the sample space of a random experiment, some outcomes satisfy a specified description, which are called event set.

Eg:

In a toss of two coins, the number of cases favourable to the "two heads" is one. viz., $[H, H]$.

(d) Union of two events:

Union of two event set A and B denoted $\{A \cup B\}$ $\{A \text{ Union } B\}$ is the event set which consists of all the outcomes (Sample Points) that belong either to A or to B or both.

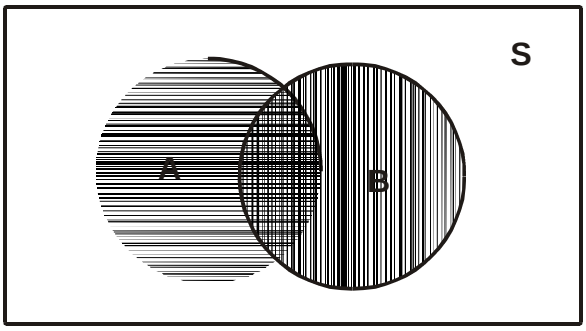
For example:

If A is the event that a Commerce graduate joins ICWAI course. If B is the event that a Commerce Graduate joins the M.Com course.

Then $A \cup B$ means the event that the student joins either ICWAI or M.Com. or both.

The total shaded region is known as

$$\mathbf{A \cup B \ (or) \ B \cup A}$$

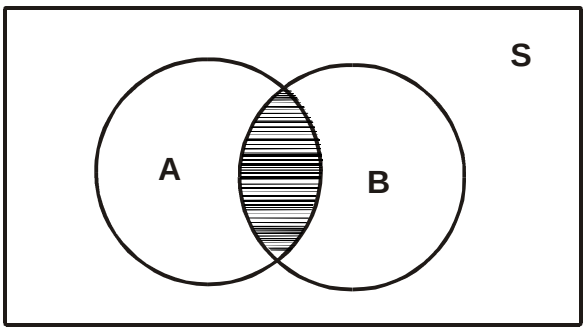


e) Intersection of two events:

The intersection of two event sets A and B is denoted as $A \cap B$ (A intersection B) is the event set which consists of all the outcome (Sample Points) which the two event. Sets A and B have in common.

In earlier example is the event that the person will join both CA and ICWAI.

The Shaded area is $\mathbf{A \cap B}$

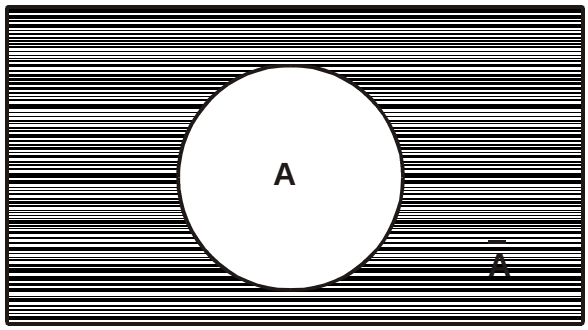


f) Complement of an event set:

The complement of an event set A is the set of all sample points of the sample space 'S' that are not contained in A Complement is denoted as A or $\bar{A}$.

Complement of a set

$\bar{A} = S - \bar{A}$: (shaded area)

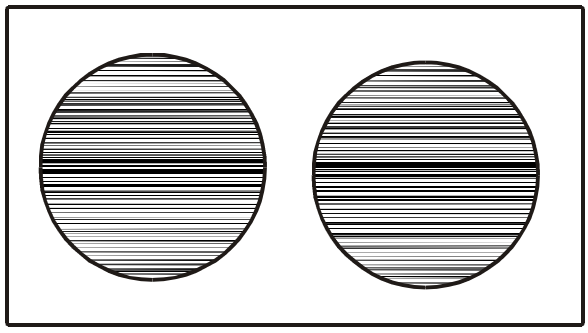


g) Mutually Exclusive Events:

Two or more events are considered mutually exclusive if the events cannot occur together, i.e., the occurrence of any one of them precludes the occurrence of the other.

A and B events are mutually exclusive or disjoint.

Venn Diagram



h) Dependent or Independent Events:

Two or more events are considered independent if the occurrence of one event in no way affects the occurrence of the other. The question of dependence or independence of events is relevant when experiments are consecutive and not simultaneous.

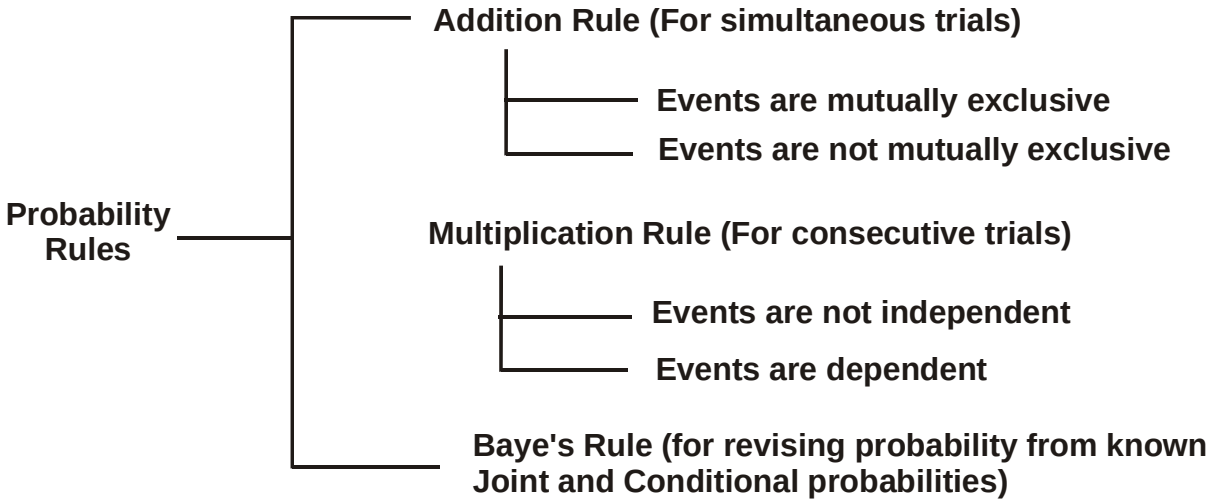
Eg:

1. In tossing of a coin later trial is not affected by the result of the previous trial - they are independent events.

2. Out of 52 cards in a pack, if one is drawn, 51 are left unless the card is replaced, the composition changs and hence the probability of the second card is affected. It is a dependent event.

Probability Rules:

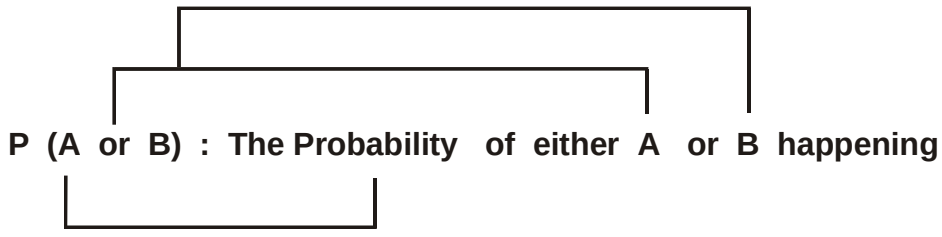
The solutions to many problems involveing probabilities require a through understanding of some basic rules which govern the manipulations of probabilities. They are generally called probability rules.



(i) Addition Rule for Mutually Exclusive Events:

Often, we are interested in the probability of one thing or the another to occur. If these two events are mutually exclusive we can express this probability using the addition rule for mutually exclusive events.

This rule is expressed symbolically as:



and it is calculated as follows:

$$P(A \text{ or } B) = P(A) + P(B)$$

In other words:

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

Example 9:

A card is drawn at random from an ordinary pack of 52 playing cards, Find the probability that a card drawn is either a Spade or the Ace of Diamonds.

Solution:

Total number of Spades = 13

Probability of drawing a Spade $= \frac{13}{52}$

Probability of drawing an Ace of Diamonds $= \frac{1}{52}$

Probability of drawing a Spade or an Ace of Diamond $= \frac{13}{52} + \frac{1}{52} = \frac{14}{52} = \frac{7}{26}$

Illustration:

Find the probability of getting a total of either 7 or 11 in a single throw of two dice.

A total of 7 can come in 6 different ways $= \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix}$

A total of 11 can come in 2 different ways $= \begin{bmatrix} 5 & 6 \\ 6 & 5 \end{bmatrix}$

The probability of getting a total $7 = \frac{6}{36}$ or $\frac{1}{6}$

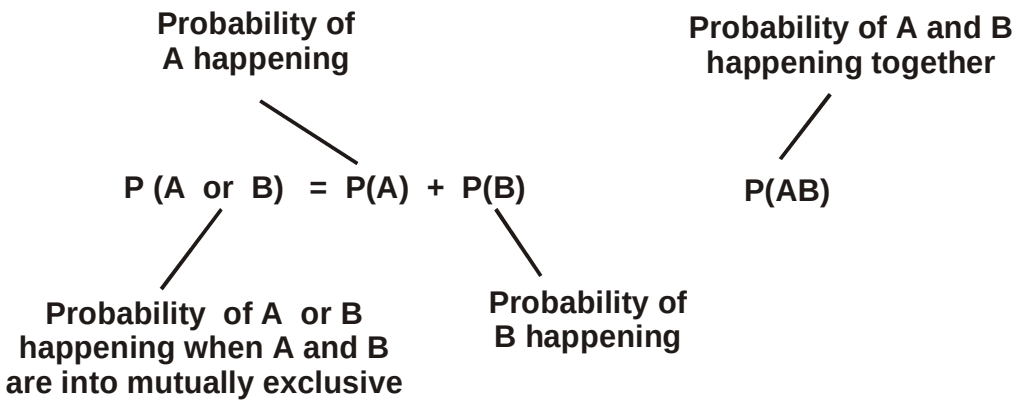
The probability of getting a total $11 = \frac{2}{36}$ or $\frac{1}{18}$

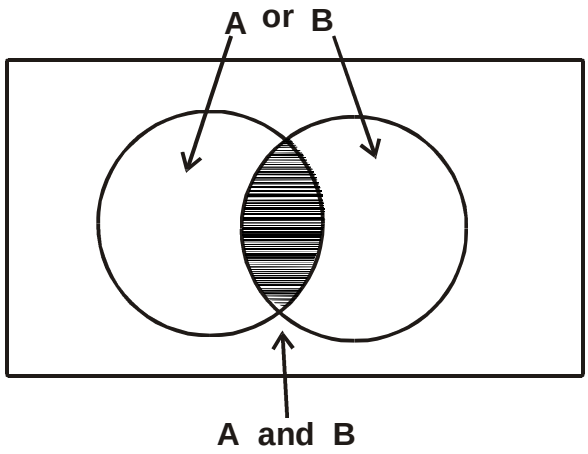
The probability of getting either 7 or $11 = \frac{1}{6} + \frac{1}{18} = \frac{4}{18}$ or $\frac{2}{9}$

(ii) Addition Rule For Events that are not Mutually Exclusive:

If two events are not mutually exclusive, it is possible for both the events to occur. In these case our addition rule must be modified. For example. What is the probability of drawing either an Ace or Heart from a deck of Cards? Obviously, the events Ace and Heart can occur together because we would draw the Ace of Hearts.

Thus Ace and Heart are not mutually exclusive evetns. W emust adjust the earlier equation to avoid double counting i.e. we have to reduce the probability of drawing either an Ace or Heart by the chance that twe could draw both of them together. As a result, the correct equation for the probability of one or more events that are not mutually exclusive is





To determine the probability of drawing either an Ace or a Heart, we can calculate.

P (Ace or Heart) = P (Ace) + P (Heart) - P (Ace and Heart)

$$= \frac{4}{52} + \frac{13}{52} = \frac{1}{52}$$

$$= \frac{16}{52} \text{ or } \frac{4}{13}$$

Example 10:

A bag contains 25 balls, numbered from 1 to 25. One ball is drawn at random. Find the probability that the number of the drawn ball will be a multiple of 5 or 7. Also find out the probability of the number being a multiple of 3 or 5.

Solution:

The probability of the number being multiple of 5 (5, 10, 15, 20, 25)

$$= \frac{5}{25}$$

The probability of the number being multiple of 7 (7, 14, 21)

$$= \frac{3}{25}$$

Thus the probability of the number being a multiple of 5 or 7 will be

$$= \frac{5}{25} + \frac{3}{25} = \frac{8}{25}$$

Probability of number being multiple of 3

(3, 6, 9, 12, 15, 18, 21, 24) $= \frac{8}{25}$

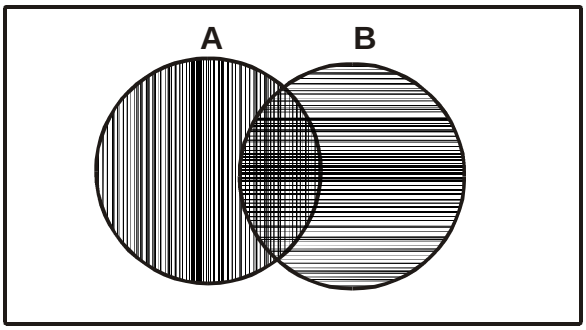
Probability of number being multiple of 3 or 5

joint probability $= \frac{8}{25} + \frac{5}{25} = \frac{13}{25}$ but this is to be adjusted as '15' is not mutually exclusive

$$= \frac{8}{25} + \frac{5}{25} - \frac{1}{25} = \frac{12}{25}$$

Generalisation:

$$\mathbf{A \cap B}$$



In this venn diagram $A \cap B$ can be expressed as the union of 3 disjoint sets

$$A \cup \bar{B} \; ; A \cap B \; ; \; \bar{A} \cup B$$

The event A can be expressed as the unions of two disjoint event sets. It is as follows:

$$P(A) = P\left(A \cup \bar{B}\right) + P\left(A \cap B\right)$$

$$P(B) = P\left(A \cap B\right) + P\left(\bar{A} \cup B\right)$$

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

Example 11:

The probability that a contractor will get a plumbing contract is 2/3 and the probability that he will not get an electric contract is 5/9. If the probability of getting at least one contract is 4/5 what is the probability that will be get both?

Solution:

Let A - denote the event that the contractor will get the plumbing contract

Let B - denote the event that the contractor will get the electric contract

Then,

$$P(A) = \frac{2}{3} \qquad P(\bar{B}) = \frac{5}{9}$$

$$P(B) = 1 - P(\bar{B}) = \frac{4}{9}$$

The probability that the contractor gets at least one contract

$$= \frac{4}{5} \qquad \text{(or)}$$

$$P(A) + P(B) - P(A \cap B) = \frac{4}{5}$$

By addition rule of probability

$$= \frac{2}{3} + \frac{4}{9} - P(A \cap B) \Rightarrow \frac{4}{5}$$

$$P(A \cap B) = \frac{2}{3} + \frac{4}{5} - \frac{4}{5} = \frac{14}{45}$$

Hence the probability that the contractor will get both the contacts is 14/15.

(iii) Probability under conditions of Statistical Independence:

When two events happen, the outcome of the first event may or may not have an effect on the outcome of the second event. That is, the events may be either dependent or independent. There are 3 types of probability under statistical independence.

1) Marginal 2) Joint 3) Conditional

a) Marginal Probability under Statistical independence:

A marginal or unconditional probability is the simple probability of the occurrence of an event.

In a fair coin toss P(H) = 0.5 and P(T) = 0.5 Every toss stands alone and is in no way connected with any other toss. Thus the outcome of each toss of a fair coin is statistically independent event.

b) Joint Probabilities under statistical independence:

The probabilities of two or more independent events occuring together or in succession is the product of their marginal probabilities.

$$P(AB) = P(A) \times P(B)$$

Probability of events A, B occuring together i.e., joint probability.

In terms of the fair coin example, the probability of the heads on two successive tosses is the probability of head on the first toss (H_1) times the probability of head on the second toss (H_2).

i.e., $P(H_1\ H_2) = P(H_1) \times P(H_2)$

The events are statistically independent, as the probability of any later outcome is not affected by an preceding outcome.

Therefore the probability of Head on any toss is 0.5 and

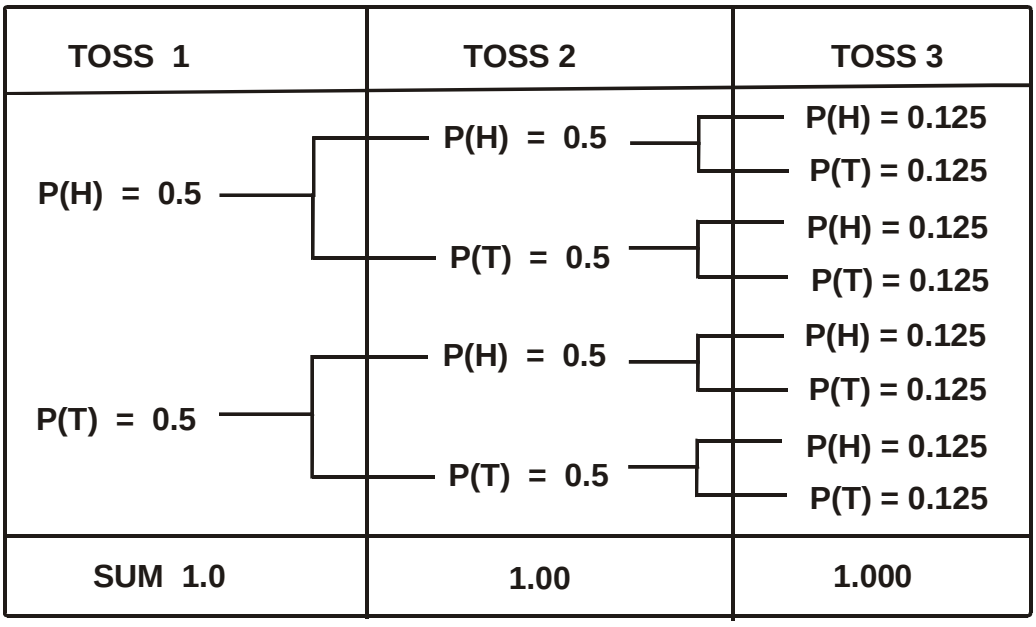
$$P(H_1\ H_2) = 0{\cdot}5 \times 0{\cdot}5$$

$$= 0.25$$

Similarly, the probability of getting 3 heads on the three successive tosses:

$$P(H_1\ H_2\ H_3) = 0{\cdot}5 \times 0{\cdot}5 \times 0{\cdot}5$$

$$= 0{\cdot}125$$



Example 12:

A bag contains 8 red and 5 white balls, two successive drawing of 3 balls are made such that (i) balls are replaced before the second trial (ii) the balls are not replaced before the second trial. Find the probability that the first draw will give 3 white and the second 3 red balls.

Solution:

(i) Draw with Replacement:

Let A : drawing 3 white balls in the first draw

Let B : drawing 3 red balls in the second draw

In case of replacement two draws are independent

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

First Draw:

Total number of possible ways:

$$= 3 \text{ balls out of 13 total balls} = {}^{3}C_{3}$$

In case of white balls, the possible ways are $= {}^{5}C_{3}$

$$P(A) = \frac{{}^{5}C_{3}}{{}^{13}C_{3}}$$

Second Draw:

As balls drawn in first draw are replaced, the bag contains 13 balls before the second draw,

Probability of drawing 3 red balls

$$P(A) = \frac{{}^{8}C_{3}}{{}^{13}C_{3}}$$

Joint Probability

$$P(A \cap B) = \frac{{}^{5}C_{3}}{{}^{13}C_{3}} \times \frac{{}^{8}C_{3}}{{}^{13}C_{3}}$$

(ii) Draws without replacement:

If balls are not replaced the events A and B are not independent. Then the required probability is

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

$\uparrow$

B given A

As worked out earlier

$$P(A) = \frac{^5C_3}{^{13}C_3}$$

Second Draw:

As balls are not replaced, then total balls before the second draw is only 13 - 3 = 10.

$P(B/A)$ is the probability of drawing 3 red balls from the bag containing 10 balls out of which 2 are white and 8 are red.

Hence $$P(B/A) = \frac{^8C_3}{^{10}C_3}$$

Therefore, $$P(A \cap B) = \frac{^5C_3}{^{13}C_3} \times \frac{^8C_3}{^{10}C_3}$$

c) Conditional Probability under statistical independence:

Conditional probability is the probbaility of an event in the sub - population or a reduced sample size. The probability of A, given B stated as, A/B is called as 'Conditional Probability of A' (Subject to the condition that has happened).

To understand the conditional probability under statistical independence look at the following illustration.

What is the probability of a fair coin to result in second head, given that head already resulted in the first toss? Symbolically, this is written as $P(H_2/H_1)$.

In case of independent events, the result of the first toss has absolutely no effect on the result of the second toss. Since the probabilities of heads and tails are identical for every toss the probability of head on thesecond toss is 0.5. Then $P(H_2/H_1) = 0.5$

Summary of probability in case independent events:

Types of Probability	Symbol	Formula
Marginal	P(A)	P(A)
Joint	P(AB)	P(A) x P(B)
Conditional	P(B/A)	P(B)

d) Probabilities under conditions of Statistical dependence:

Statistical dependence exists when the probability of some event is dependent upon or affected by the occurrence of some other event. The probabilities are as follows:

(i) Conditional Probability Under Statistical Dependence:

The conditional probability under statistical dependence is

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

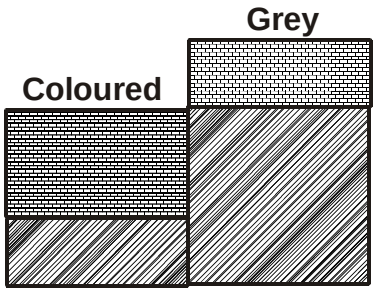
Example 13:

A box contains ten balls, of which 3 are coloured and dotted. '1' is colored and striped, '2' are grey and boxes and 4 are grey and striped.

Solution:

Now the probability of drawing any one ball from the box is 0.10

If a coloured ball is drawn, what is the probability that it is boxes. What is the probability that is striped? Symbolically $P(D/C)$ = conditional probability that this ball is boxes given that is coloured.



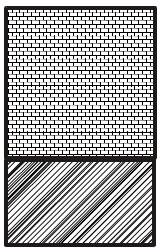
3 balls are coloured and boxes 1 ball is coloured and striped

2 balls grey and boxes

4 balls grey and striped

In the box there are a total of 4 coloured balls and one of which is striped. Therefore 3 are dotted.

$$P(D/C) = \frac{3}{4} = 0 \cdot 75$$



But to calculate the conditional probability of dotted given coloured $P(D/C)$, divide the probability of coloured and dotted balls (3 out of 10 or 0.3 by the probability of coloured balls 4 out of 10 or 0.4)

$$P(D/C) = \frac{P(DC)}{P(C)}$$

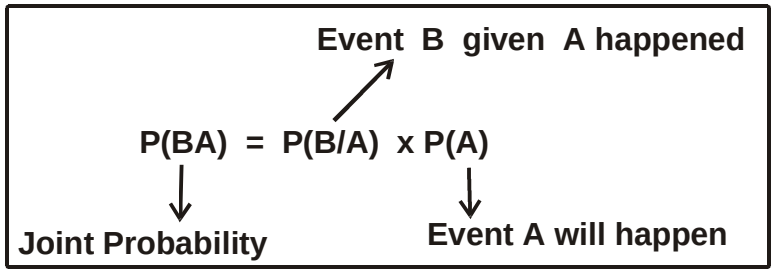
That is the formula for conditional probability.

(ii) Joint probability under statistical dependence:

When the formula for conditional probability under conditions of statistical dependence is

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

We can get the joint probability P(BA) by simple cross multiplication.



In the earlier illustration.

$$P(CD) = P(C/D) \times P(D)$$

or $P(CD) = 0 \cdot 6 \times 0 \cdot 5 = 0 \cdot 3$

Three balls out of 10 are coloured and dotted.

(iii) Marginal Probabilities under statistical dependence:

Marginal probabilities under statistical dependence are computed by summing up the probabilities of all the joint evetns in which the sample event occurs. In the example, the marginal probability of the two joint events in which coloured ball occurs.

$$P(C) = P(CD) \times P(CS)$$

$$= 0 \cdot 3 + 0 \cdot 1 = 0 \cdot 4$$

Summary:

Type of Probability	Symbol	Formula under Statistical dependence
Marginal	P(A)	Sum of joint events in which A occurs
Joint	P(AB)	$P(A/B) \times P(B)$
Conditional	$P(B/A)$	$P(BA)/P(A)$

Example 14:

The Personnel Department of a Comapany has records which show the following analysis of its 200 engineers.

Age	Bachelor's Degree	Master's Degree	Total
Under 30	90	10	100
30 - 40	20	30	50
Above 40	40	10	50
	150	50	200

If one Engineer is selected at random, find

(i) The probability that he has only a Bachelor's degree.

(ii) The probability that he has a Master's Degree given that, he is over 40.

(iii) The probability that he is under 30, given that he has only Bachelor's degree.

Solution:

Let A : Engoneer with Bachelor's degree only

Let B : Engineer with Master's degree

Let C : Engineer with an age below 30 years

Let D : Engineer above 40 yers

(i) $P(A) = \frac{150}{200}$

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

$= \frac{10/200}{50/200} = \frac{1}{5}$

(iii) $P(C/D) = \frac{P(C \cap A)}{P(A)}$

$= \frac{90/200}{150/200} = \frac{3}{5}$

15.7 Summary:

In this lesson we have explained the concept of probability and also estimate the probability under different approaches. Concept of permutation, combinations and their use in probability are also explained. Further, theory of sets and rules of probability are discussed.

15.8 Exercises:

1. (i) A bag contains 3 white and 5 black balls. If a ball is drawn at random find the probability for it to be black.

   (ii) A bag contains 5 white, 7 black and 4 red balls. 3 balls are drawn from it at random. Find the probability that all the 3 balls are white.

2. A book containing 100 pages is opened at random. find the probability that on the page (i) a doublet is found, (ii) a number whose sum of the digits is 10.

3. A four digit number is formed with the digits 1, 2, 3, 4, 5 with no repetitions of any digit. Find the chance that the number is divisible by 5.

4. Find the probability that in a random arrangement of the letters of the word UNIVERSITY, the two I's do not come together.

5. 10 books are placed at random in a shelf, find the probability that a particular pair of books shall be (i) always together, (ii) never together.

6. Three cards are drawn from a pack of cards. find the chance that hey are an ace, a king and a queen.

7. Three fair dice are thrown. What is the probability of getting a sum 6 or less on the three dice?

8. Six boys and six girls sit in a row at random. What is the probability that (i) The six girls sit together (ii) The boys and girls sit alternately.

9. Five tickets are draw N at random from a bag cotnaining 50 tickets numbered 1, 2, 3,.....,50. The tickets are arranged in asscending order of magnitude $\left(x_1 < x_2 < x_3 < x_4 < x_5\right)$. Find the probability that $x_3 = 30$.

10. From 6 gentlemen and 4 ladies a committee of 5 is to be formed. find the probability that this can be done so as to always include at least one lady.

11. A bag contains 2 white, 3 black and 4 green balls. The first ball is taken at random from the bag. The second ball is taken at random from the remaining balls. What is the probability that the first ball is white and the second black.

12. Mr. X is selected for interview for 3 parts. For the first post there are 5 condidates, for the second there are 4 and for the third there are 6. If the selection of each condidate is equally likely, find the chance that Mr. X will be selected for atleast one post.

13. An urn A contains 8 black balls and 5 white balls. A second urn B contains 6 black and 7 white balls. Find the probability that blind folded person in one draw shall obtain a white ball from one urn.

14. A husband and wife appear in an interview for the same post. The probability of husband's selection is $\frac{1}{7}$ and that of wife's selection is $\frac{1}{5}$. What is the probability that (a) both of then will be selected (b) none of them will be selected (c) only one of them will be selected.

15. An anti - air craft gun can take a maximum of 4 shots at an enemy plane moving away from it. The probability of hitting the plane at the first, second, third and fourth shots are 0.4, 0.3, 0.2 and 0.1 respectively. What is the probability that the gun hits the plane.

16. Two persons A and B loss a dice. The person who first throws 4 or 5 wins. A starts the game Show that the probability of A's and B's wimming are in the ration 3 : 2.

17. Box A contains 5 red and 3 white morbles and Box B contains 2 red and 6 white morbles. (a) If a marble is drawn from each box, what is the probability that they are both of the same colour? (b) If 2 morbles are drawn from each box. What is the probability that all four marbles are of some colour.

18. A die is tassed. If the number is odd, what is the probability that it is prime.

19. In a certain town 40% have brown hair 25% have brown eyes and 15% have both brown hair and brown eyes. A person is selected at random from the town

    (1) If he has brown hair, what is the probability that he has brown eyes also?

    (2) If he has brown eyes, determine the probability that he does not have brown hair.

    (3) Determine the probability that he has reither brown hair nor brown eyes.

20. Determine whether sex and blood group are independent from the following table:

Blood Group	Male	Female	Total
0	113	113	226
A	103	103	206
B	25	25	50
AB	10	10	20
Total	251	251	502

21. Box I contains 1 white, 2 red, 3 green balls Box II contains 2 white, 3 red, 1 green balls, Box III contains 3 white, 1 red, 2 green balls. Two balls are drawn from a box choosen at random. These are found to be one white and one red. Determine the probability that the balls so drawn came from box II.

15.9 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer

Dr. K. CHANDAN

Lesson - 16

BAYE'S THEOREM

Objective:

After going through this lesson, you will learn:

- The concept of Baye's theorem and computation procedure.

Structure:

16.1 Introduction:

The joint and marginal probabilities can be used to revise the probability of a particular event in the light of available additional information. For example, if an event has occured through one of the various mutually disjoint events or reasons, then the unditional probability that it has occured due to a particular event or reason is called its reversed or posterior probability. These probabilities are computed by Baye's rules, named after the British mathematician Thomas Bayer who had propounded it in 1763.

16.2 Baye's Theorem:

$E_1, E_2, \cdots\cdots\cdots, E_n$ are n mutually exclusive and exhaustive events such that $P(E_1) > 0 \; (i = 1, 2, \cdots\cdots\cdots, n)$ in a sample space S and A is any other event in S intersecting with every E_i (i.e., A can only occur in combination with every one of the events $E_1, E_2, \cdots\cdots\cdots, E_n$) such that $P(A) > 0$.

If E_i is any of the events of $E_1, E_2, \cdots\cdots\cdots, E_n$ where $P(E_1), P(E_2), \cdots\cdots\cdots P(E_n)$ and $P\left(\frac{A}{E_1}\right) \cdot P\left(\frac{A}{E_2}\right) \cdots\cdots\cdots\cdots P\left(\frac{A}{E_n}\right)$ are known then

$$P\left(\frac{E_k}{A}\right) = \frac{P(E_k) \cdot P(A/E_k)}{P(E_1) \cdot P(A/E_1) + P(E_2) \cdot P(A/E_2) + \cdots\cdots\cdots + P(E_n) \cdot P(A/E_n)}$$

Example 1:

In a certain college 25% of boys and 10% of girls are studying mathematics. The girls constitute 60% of the student body. (a) What is the probability that mathematics is being studied? (b) If a student is selected at random and is found to be studying mathematics, find the probability that the student is a girl? (c) a boy?

Solution:

Given P (Boy) $= P(B) = \frac{40}{100} = \frac{2}{5}$

$P(G) = \frac{60}{100} = \frac{3}{5}$

Probability that mathematics is studied given the student is a boy $= P(M/B) = \frac{25}{100} = \frac{1}{4}$

Probability that mathematics is studied given the student is a girl $= P(M/G) = \frac{10}{100} = \frac{1}{10}$

(a) Probability that maths is studied

$= P(M) = P(G)\ P(M/G) + P(B)\ P(M/B)$

$\therefore$ By total probability theorem

$P(M) = \frac{3}{5} \cdot \frac{1}{10} + \frac{2}{5} \cdot \frac{1}{4} = \frac{4}{25}$

(b) By Baye's theorem probability of maths student is a girl

$P(G/M) = \frac{P(G)P(M/G)}{P(M)} = \frac{3/5 \cdot 1/10}{4/25} = \frac{3}{8}$

(c) Probability of maths student is a boy

$= P(B/M) = \frac{P(B)\ P(M/B)}{P(M)} = \frac{3/5 \cdot 1/4}{4/25} = \frac{5}{8}$

Example 2:

The chance that doctor A will diaginse a disease x correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40% and the chance of death by worng diagnosis is 70%. A patient of doctor A, who had disease x died. What is the chance that his disease was diagnised correctly.

Solution:

Let E_1 be the event that "disease x is diagnised correctly by doctor A" and E_2 be the event that "a patient of doctor A who has disease x died". What is the chance that his disease was diagnised correctly.

Solution:

Le t E_1 be the event that "disease x is diagnised correctly be doctor A" and E_2 be the event that "a patient of doctor A who has disease x died".

Then $P(E_1) = \frac{60}{100} = 0\cdot 6, \quad P\left(\frac{E_2}{E_1}\right) = \frac{40}{100} = 0\cdot 4$

$$P(\bar{E}) = 1 - 0\cdot 6 = 0\cdot 4,\ P\left(\frac{E_2}{\bar{E}_1}\right) = \frac{70}{100} = 0\cdot 7$$

By Baye's theorem $P\left(\frac{E_1}{E_2}\right) = \frac{P(E_1)\cdot P(E_2/E_1)}{P(E_1)\cdot P(E_2/E_1) + P(\bar{E}_1)\cdot P(E_2/\bar{E}_1)}$

$$= \frac{0\cdot 6 \times 0\cdot 4}{0\cdot 6 \times 0\cdot 4 + 0\cdot 4 \times 0\cdot 7} = \frac{6}{13}$$

Example 3:

A bag A_1 contains 2 white and 3 red balls and a bag B_1 contains 4 white and 5 red balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that the red ball drawn is from bag B_1 .

Solution:

Let E_1, E_2 be the events of drawing from bag A_1, bag B_1 respectively.

$$\therefore P(E_1) = \frac{1}{2} \text{ and } P(E_2) = \frac{1}{2}$$

Let A be event of drawing a red ball from any of the bags.

Probability of drawing a red ball from bag $A_1 = P(A/E_1) = \frac{3}{5}$

Probability of drawing a red ball from bag $B_1 = P(A/E_2) = \frac{5}{9}$

$\therefore$ The probbaility that a red ball is drawn from bag B_1

$$P(E_2/A) = \frac{P(E_2)\cdot P(A/E_2)}{P(E_1)\cdot P(A/E_1) + P(E_2)\cdot P(A/E_2)} \quad \text{(By Baye's Theorem)}$$

$$= \frac{1/2 \cdot 5/9}{1/2 \cdot 3/5 + 1/2 \cdot 5/9} = \frac{5/9}{52/45} = \frac{25}{52}$$

Example 4:

First box contains 2 black, 3 red 1 white balls second box contains 1 black 1 red 2 white balls and third box contains 5 black 3 red, 4 white balls of these a box is selected at random from it a red ball is randomly drawn if the ball is red, find the probability that is from second box.

Solution:

Let x, y, z be the 1st second and third boxes

$$P(x) = \frac{1}{3},\ P(y) = \frac{1}{3},\ P(z) = \frac{1}{3}$$

Let R be the event of drawing a red ball from a box

So, $P(R/x) = \frac{3}{6},\ P(R/y) = \frac{1}{4},\ P(R/z) = \frac{3}{12}$

By Baye's theorem the required probability

$$= P(y/R)$$

$$= \frac{P(y)\cdot P(R/y)}{P(x)\cdot(R/x) + P(y)\cdot P(R/y) + P(y)\cdot P(R/z)}$$

$$= \frac{1/3 \times 1/4}{1/3 \times 3/6 + 1/3 \times 1/4 + 1/3 \times 3/12} = \frac{1}{4}$$

Example 5:

Suppose 5 men out of 100 and 25 women out of 10,000 are color blind. A color blind person is choosen at random. What is the probability of the person being a male (Assume male and female to be in equal numbers).

Solution:

Given that 5 men out of 100 and 25 women out of 10000 are color blind.

A color blind person is choosen at random.

The probability that the choosen person is male $P(M) = \frac{1}{2}$

Similarly the probability that the choosen person is female $P(W) = \frac{1}{2}$

Let B represent a blind person then

$$P(B/M) = \frac{5}{100} = 0 \cdot 05$$

$$P\left(\frac{B}{W}\right) = \frac{25}{10000} = \frac{1}{400} = 0 \cdot 0025$$

The probability that the choosen person is male

$$P\left(\frac{M}{B}\right) = \frac{P(B/M)\,P(M)}{P(M) \cdot P\left(\frac{B}{M}\right) + P(W) \cdot P\left(\frac{B}{W}\right)}$$

$$= \frac{0 \cdot 05 \times 0 \cdot 5}{(0 \cdot 05 \times 0 \cdot 5) + (0 \cdot 5 \times 0 \cdot 0025)} = 9 \cdot 5$$

Example 6:

In a bolt factory machines A, B, C manufacture 20%, 30% and 50% of the total of their output and 6%, 3% and 2% are defective. A bolt is drawn at random and found to be defective. Find the probabilities that is manufactured from (i) Machine A, (ii) Machine B (iii) Machine C.

Solution:

Let P(A), P(B), P(C) denote the probabilities of the events "bolts are manufactured by the machines A, B, C". Then by data

$$P(A) = \frac{20}{100} = \frac{1}{5}\ ,\ P(B) = \frac{30}{100} = \frac{3}{10},\ P(C) = \frac{50}{100} = \frac{1}{2} \cdot$$

Let D denote that the bolt is defective

$$P(D/A) = \frac{6}{100}\ ,\ P(D/B) = \frac{3}{100}\ ,\ P(D/C) = \frac{2}{100}$$

(i) If bolt is defective, then the probability that it is from machine $A = P(A/D)$

$$= \frac{P\left(\frac{D}{A}\right) \cdot P(A)}{P\left(\frac{D}{A}\right) \cdot P(A) + P\left(\frac{D}{B}\right) \cdot P(B) + P\left(\frac{D}{C}\right) \cdot P(C)} = \frac{12}{31}$$

Similarly

(ii) $P(B/D) = \frac{9}{31}$

(iii) $P(C/D) = \frac{10}{31}$

Example 7:

Of the three men, the chances that politician, a business man or an academician will be appointed as a vice - chancellor (V.C) of a University are 0.5, 0.3, 0.2 respectively. Probability that research is promoted by these persons if they are appointed as V.C. are 0.3, 0.7, 0.8 respectively.

(i) Determine the probability that research is promoted

(ii) If research is promoted, what is the probability that VC is an academician.

Solution:

Let A, B, C are the events that a politician, businessmen or an academician will be appointed as V.C. of the three men.

$$P(A) = 0\cdot5,\ P(B) = 0\cdot3,\ P(C) = 0\cdot2$$

$$= \frac{5}{10} \qquad = \frac{3}{10} \qquad = \frac{2}{10}$$

The probbailities that research is promoted if they are appointed as V.C.s are

$$P(R/A) = 0\cdot3 = \frac{3}{10}$$

$$P(R/B) = 0\cdot7 = \frac{7}{10}$$

$$P(R/C) = 0\cdot8 = \frac{8}{10}$$

(i) The probability that the research is promoted

$$= P(R/A) = P(R/B) + P(R/C) = \frac{3}{10} + \frac{7}{10} + \frac{8}{10} = 1\cdot8$$

(ii) The probability that if research is promoted that theV.C. is an academician

$$= \frac{P\left(\frac{R}{C}\right)\cdot P(C)}{P\left(\frac{R}{C}\right)\cdot P(C) + P\left(\frac{R}{B}\right)\cdot P(B) + P\left(\frac{R}{A}\right)\cdot P(A)} = \frac{4}{13}$$

Example 8:

A business man goes to hotels X, Y, Z 20% , 50%, 30% of the time respectively. It is known that 5%, 4%, 8% of the rooms in X, Y, Z hotels have faulty plumbings. What is the probability that business man's room having faulty plumbing is assingned to hotel z.

Solution:

Let the probabilities of business man going to hotels X, Y, Z are respectively $P(X), P(Y), P(Z)$.

$$P(X) = \frac{20}{100} = \frac{2}{10}$$

$$P(Y) = \frac{50}{100} = \frac{5}{10}$$

$$P(Z) = \frac{30}{100} = \frac{3}{10}$$

Let E = the event that the hotel room having faulty plumbing.

The probabilities that X, Y, Z hotels have faulty plumbing are

$$P(E/X) = \frac{5}{100} = \frac{1}{20}$$

$$P(E/Y) = \frac{4}{100} = \frac{1}{25}$$

$$P(E/Z) = \frac{8}{100} = \frac{2}{25}$$

The probability that the business man's room having faulty plumbing is assigned to hotel Z.

$$= P\left(\frac{Z}{E}\right) = \frac{P\left(E/Z\right) \cdot P(Z)}{P\left(E/Z\right) \cdot P(Z) + P\left(E/Y\right) \cdot P(Y) + P\left(E/X\right) \cdot P(X)}$$

$$= \frac{2/25 \times 3/10}{2/25 \times 3/10 + 1/25 \cdot 5/10 + 1/20 \cdot 2/10} = \frac{4}{9}$$

Example 9:

There are two boxes in box I - 11 cards are there numbered 1 to 11 and in box II - 5 cards numbered 1 to 5. A box is chosen and a card is drawn. If the card thows an even number then another card is drawn from the same box. If card show's an odd number another card is drawn from the other box. Find the probbaility that (i) both are even (ii) both are odd (iii) if both are even what is the probability that they are from box I.

Solution:

Number of cards in box I = 11

Number of cards with even number = 5

Number of cards with odd number = 6

Number of cards in box II = 5

Number of cards with even number = 2

Number of cards with odd number = 3

The probability of choosing any one box = $\frac{1}{2}$

(i) Let E = The event that both the cards are even.

For this a box is choosen and a card is picked, if the first card is even then the second card is also picked from the same box and that card is also even.

Let E_1 = both the cards are from box I

$$\therefore P(E_1) = \frac{1}{2}\left(\frac{5}{11}\right)\left(\frac{4}{10}\right) = \frac{1}{11}$$

Let E_2 = both the cards are from box II

$$P(E_2) = \frac{1}{2}\left(\frac{2}{5}\right)\left(\frac{1}{4}\right) = \frac{1}{20}$$

$$\therefore P(E) = P(E_1) + P(E_2) = \frac{1}{11} + \frac{1}{20} = \frac{31}{220}$$

(ii) Let E = both the cards are odd.

Then a box is choosen. !st card is odd and second card is picked from another box and that is also odd.

E_1 = 1st card is odd from box 1 and second card is odd from box 11

$$P(E_1) = \frac{1}{2}\left(\frac{6}{11}\right)\left(\frac{3}{5}\right) = \frac{9}{55}$$

E_1 = 1st card is odd from box II and 2nd card is odd from box 1

$$P(E_2) = \frac{1}{2}\left(\frac{3}{5}\right)\left(\frac{6}{11}\right) = \frac{9}{35}$$

$$\therefore P(E) = P(E_1) + P(E_2) = \frac{9}{35} + \frac{9}{55} = \frac{18}{55}$$

(iii) The Probability that both cards are even and from box I

$$= \frac{1}{2}\,\frac{5}{11}\,\frac{4}{10} = \frac{1}{11}$$

The probability that both cards are even and from box 11

$$= \left(\frac{1}{2}\right)\left(\frac{2}{5}\right)\left(\frac{1}{4}\right)$$

Probability that if both cards are even then they are from box I

$$= \frac{1/11}{1/2 \cdot 5/11 \cdot 4/10 + 1/2 \cdot 2/5 \cdot 1/4} = \frac{20}{3} = \frac{1/11}{1/11 + 1/20} = \frac{20}{31}$$

Example 10:

The bolts are drawn from a box containing 4 good and 6 bad bolts. Find the probability that the second bolt is good if the first one is found to be bad.

Solution:

Let G = Probability of getting a good bolt

B = Probability of getting a bad bolt

P(B) = Probability of getting first bold bad

$$P(B) = \frac{6}{10}$$

$P(G/B)$ = Probability of getting second bolt good given the first bolt is bad.

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

$$P(G \cap B) = \frac{6}{10} \times \frac{4}{9} = \frac{4}{15}$$

$$P\left(\frac{G}{B}\right) = \frac{4/15}{6/10} = \frac{4}{9}$$

Example 11:

In a factory, machine A produces 40% of the output and machine B produces 60%. On the average, 9 items in 1000 produced by A are defective and 1 item in 250 produced by B is defective. An item drawn at random from a day's output is defective. what is the probability that it was produced by A or B?

Solution:

Output produced by A = 40%

$$\therefore P(A) = 0 \cdot 4$$

Output produced by B = 60%

$$\therefore P(B) = 0 \cdot 6$$

$P\left(\frac{D}{A}\right)$ = Probability that items produced by A are defective $= \frac{9}{1000} = 0 \cdot 009$

Similarly $P\left(\frac{D}{B}\right) = \frac{1}{250} = 0 \cdot 004$

$P\left(\frac{A}{D}\right)$ = Probability of manufacturing the defective bolt by machine A

$$= \frac{P(A) \times P\left(\frac{D}{A}\right)}{P(A) \times P\left(\frac{D}{A}\right) + P(B) \times P\left(\frac{D}{B}\right)} = \frac{0 \cdot 4 \times 0 \cdot 009}{0 \cdot 4 \times 0 \cdot 009 + 0 \cdot 6 \times 0 \cdot 004}$$

$$= \frac{0 \cdot 0036}{0 \cdot 0036 + 0 \cdot 0024} = \frac{0 \cdot 0036}{0 \cdot 006} = 0 \cdot 6$$

$P\left(\frac{B}{D}\right)$ = Probability of manufacturing the defective bolt by machine B

$$= \frac{0\cdot 6 \times 0\cdot 004}{0\cdot 006} = \frac{0\cdot 0024}{0\cdot 006} = 0\cdot 4$$

$$P\left(\frac{A}{D}\right) = 0\cdot 6$$

Then we have $P\left(\frac{B}{D}\right) = 0\cdot 4$

16.3 Summary:

In this lesson, we have explained the Baye's Theorem and their applications.

16.4 Exercise:

1. A test of multiple choice questions with four choices is held. A candidate appearing for the test either guesses or recalls or computes the answer for any question. The probability that he makes a guess as $\frac{1}{6}$ and the probability that he recalls the answer is $\frac{1}{3}$. The probability that his answer is correct, given that he recalls it is $\frac{1}{8}$. Find the probability that his answerto the question, given that he answered it correctly.
2. A box contains 4 balls. Two balls are drawn from it and are found to be white. Find the probability that all the balls in the bag are white.
3. In a certain college, 40% of men and 10% of women are taller than 1.8 metres. Further more in the college 60% of students are women. If a student is selected at random and is taller than 1.8 metres find the probability that the selected student is a woman.
4. Companies B_1, B_2, B_3 produce 30%, 45% and 25% of the cars respectively. It is known that 2%, 3% and 2% of the cars produced from B_1, B_2 and B_3 are defective.
   (a) What is the probability that a car purchased is defective?
   (b) If acar purchased is found to be defective what is the probability that this car is produced by company B_3?
5. Suppose three companies X, Y, Z produce T.V.'s X produce twice as many as Y while Y and z produce the same number. it is known that 2% of X, 2% of Y and 4% of y are defective. All the T.V.'s produced and put into one shop and then one T.V. is choosen at random.

(a) What is the probbaility that the TV is defective?

(b) Suppose a T.V. choosen is defective, what is the probability that this T.V. is produced by company X?

6. Box I contains 1 white, 2 red, 3 green balls, Box II contains 2 white, 3 red, 1 green balls, Box III contains 3 white, 1 red, 2 green balls. Two balls are drawn from a box chosen at random. These are found to be one white and one red. Determine the probability that the balls so drawn come from box II.

7. A shopkeeper buys a particular type of Electric bulbs from three manufacturers M_1, M_2 and M_3. He buys 25% of his requirement from M_1 – 45% from M_2 and 30% from M_3. Based on the past experience he found that 2% of type M_3 bulbs are defecive where as only 1% of type M_1 and type M_2 are defective. If a bulb chosen by him at random is found defective let us find the probability that it was of type M_3.

8. Suppose that an urn B_1 contains 2 white and 3 black balls and another uron B_2 contains 3 white and 4 black balls. One uron is selected at random and a ball is drawn from it. Ifthe ball drawn is found black. Let us find the probability that the uron chosen was B_1.

9. A speaks truth in 75% of the cases and B in 80% cases. What is the probability that their statements about an incident do not match.

10. A problem in calculus is given to two students A and B whose chances of solving it are $\frac{1}{3}$ and $\frac{1}{4}$. Find the probability of the problem being solved if both of them try independently.

11. A and B toss a coin 50 times each simolfaneouusly. Find the probability that both of them will not got tails at the same toss.

12. Three boxes B_1, B_2 and B_3 contains balls with different colours as shown below:

	white	black	Red
B_1	2	1	2
B_2	3	2	4
B_3	4	3	2

A die is thrown B_1 is chosen if either 1 or 2 terms up. B_2 is chosen if 3 or 4 terms up and B_3 is choosen if 5 or 6 terms up. Having chosen a box in this way a ball is chosen at random from this box. if the ball is down is found to be red find all probability that it is drawn from box B_2.

8. An uron contains w white ball s and b black balls. Two players Q and R alternatively draw a ball with replacement from the uron. the player that draws a white ball first wins the game. If Q begins the game, find the probability of hs winning the game.

16.5 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer

Dr. K. CHANDAN

Lesson - 17

RESEARCH REPORT WRITING

Objective:

After going through this lesson, you will learn:

- The concept of report writing, report characteristic, communication, guidelines for report writing and audiovisual techniques.

Structure:

17.1 Introduction:

A report is a very formal document that is written for a variety of purposes in the sciences, social sciences, engineering and business disciplines, Generally, findings pertaining to a given or specific task are written up into a report. It should be noted that reports are considered to be legal documents in the workplace and thus, they need to be precise, accurate and difficult to misinterpret.

(a) Types of Reports:

Laboratory reports, research reports, field study reports, proposals, progress reports, technical reports, financial reports, health and safety reports, case study reports, cost - benefit analysis reports, comparative advantage reports, feasibility studies, instruction manuals, and on it goes.

(b) Report differ from the structure of an essay:

Reports are organized into separate sections according to the specific requirements of the given task. While it is important that paragraphs are structured and there is unity, coherence and logical development to the report, it is not a continuous piece of writing like an essay. Each type of report serves a very specific purpose and is aimed at a very particular audience.

Report writing may seem respectively to us, but this is because reports are not usually read from cover - to - cover by one person. For example, a manager may read only the synopsis or abstract and act on the advice it contains while a technical officer may read only the section that explains how things work. On the other hand, a personnel officer may look at only the conclusions and recommendations that directly affect his or her working area.

(c) Report Includes:

Because there is such a wide range of reports that serve different purposes, your faculty will generally have guidelines that they want you to follow. As a general rule, however, the following should give you some indication of what to include in a formal report.

(d) Purpose of Report:

The research report requirement allows you to refine and demonstrate your research skills. These skills are essential for conducting sound studies for employers, for consulting work, and for Ph.D. level graduate work. The report project will enable you to demonstrate your ability to:

- Conduct a literature search on a selected topic
- Critically review the literature
- Compare, contrast and integrate the findings
- Identify areas for further research and
- Design a (hypothetical) study to investigate at least one of those areas

17.2 General Report Characteristics:

The report consists of two distinct parts: (a) The literature review and (b) a proposal for conducting a study that extends beyond there viewed literature. Ideally, the finished product should approach the quality of published literature reviews (e.g., those appearing in Annual Reviews of Psychology, Academy of Management Review, and Psychological.)

1. **Literature Review:** This is the major part of the report. The review should conclude with a discussion of areas for further research that follows logically from the reviewed literature.
2. **Study Proposal:** This part of the report should be considerably shorter than the literature review. Typical proposals range from 3 to 6 double - spaced typed pages. The aim is to demonstrate your ability to design a psychological study. The proposal focuses upon an area for potential further research by
   - discussing the research question and its significance,
   - defining research variables (i.e., dependent and if appropriate, independent ones),

- stating a theory or hypothesis based on those variables (with justification) and
- detailing the design and implementation of a study that could be used to test the theory or hypothesis.

You are not expected to actually conduct the study or to collect data for the research report, you are expected to think out the design of the study carefully, including the statistical methods of data analysis and experimental conditions, if applicable. Essentially, the last step above entails writing the equivalent of a METHOD section. Good models for METHOD sections can be found in most empirical journals (e.g., Journal of Applied Pshychology, Academy of Management journal, Journal of Applied Behavioral Science).

3. **Writing Style:** The report should be written in the third person. In addition, note that it is inappropriate for a research report to present advice for readers, individuals, or organizations (e.g., all large organizations should have stress management programs) or to state conclusions or speculations that extend far beyond researched samples (e.g.) turnover is probably greatest among uneducated employees. Further more, unsubstantiated and overly general statements about human nature, managerial work organizations etc., should be avoided (e.g., most managers are not concerned with setting long - term performance objectives). Together with unsubstantiated statistics about populations (e.g., two - thirds of the managers in the united states are over forty years old) AS general rules,

   - try to cite references to back up major assertions,
   - be cautions and logical about making extensions beyond published research findings and
   - use words that denote judgement (e.g., should, must, good, bad, etc.,)

   very cautiously if at all.

4. **Selecting a Tentative Topic:** Invest considerable care in choosing your topic. One that is too broad will overwhelm you (and the reader) one that is too narrow will not provide you with enough literature to review. To sustain your efforts it is also critical that you have a strong interest in your topic. This is your opportunity to become a specialist and to build a knowledge area that can enhance your career prospects.

5. **Focusing your topic and locating literature:** Unless you are already an expert on the topic, focusing your report may require several iterations. Here are some general guidelines.

6. **Read Widely:** Be sure your proposed topic is clearly and read as much as you can about it. Seek out recent literature (especially review articles, if possible) on your topic or on topics closely related to yours. (Computer searches in the library, based on keywords, are an excellent way to start). Consult prior course readings. Seek out as many empirical (data -oriented) articles as possible before looking for practitioner - oriented articles. For some topics, business or educational journals may be appropriate to consult. Whenever you locate a relevant article, check its References list for additional publications that may pertain to your topic. One sign of through ness in searching the literature is when most of the entries in references lists point to sources you have already investigated.

7. **Take Stock of what you have found:** You should be able to evaluate the practicality of your topic after you have read widely, Ideas from the literature may very well give you a new or revised perspective. If you can't find enough articles (e.g., on "career development of sanitation engineers".) You may need to broaden or change your topic area. If you find an overwhelming amount of literature (e.g. on motivation in the workplace) try to identify a facet of your original topic that is more limited in scope but still addressed sufficiently by the literature (e.g., on motivating workers through pay for performance systems). Another way to limit your topic is to concentrate upon recent theories, research and / or developments in rapidly changing areas. Overall, the conent of your research report should represent your original thinking and should not overlap substantially with existing review articles or reports.

8. **Keep track of what you have read:** Many writors prefer to photocopy the articles they review so that they can be highlighted and marked up. If you copy an article, be sure that the full journal citation is clearly identified on the copy. For some individuals, taking notes on index cards works well.

9. **Seek input from your report advisor:** It might be appropriate to discuss the feasibility of your initial ideas with him/her before reviewing a great deal of literature. However, your report advisor cannot be expected to be an expert on your topic nor to be a source of topic ideas and literature references. You are expected to take independent initiative to generate potential topics and to become knowledgeable with what has been written about them. Which advisor you work will depend partly upon your preference, your proposed topic, the advisor's areas of interest, and his/her schedule. Generally, you are expected to do some preliminary reading, literature searching and hard thinking on your proposed topic before selecting and meeting with an advisor. In some situations, it may be appropriate to have a backup topic that you have also investigated.

10. **Organizing and Refining Your Literature Review:** Good organization is one of the keys to a successful research project. Here are some suggestions:

(a) Create an outline: Reading several published review articles will give you a good sense for how literature reviews are organized. It is best to outline your review before writing it. Most reviews are organized with headings and sub headings. Start by determining the headings of the major sections of your review. Then create subheadings under each headings that list the major concepts or ideas to be addressed. Under each major concept or idea, you can then identify specific articles or finer points that are to be explored.

(b) Refine your ideas: You may have to rearrange headings or information in your outline in order for ideas to flow smoothly and to be integrated. Seeking additional references may be appropriate for categories that appear to be incomlete. When you draft your review based on the outline, you do not have to devote equal space to each sub topic; it is common for reviews to initially overview several broad areas before developing into a logical subset of them in more detail. It is a good idea to briefly review your outline with your advisor before actually drafting the review.

(c) Create a rough draft first: Most sound reviews are written by successively improving upon earlier drafts. Your first draft should aim to capture your basic ideas without emphasizing find wording, Later, read over and edit your draft, focusing upon the readability

and logic of what you have written. It is also essential to check carefully for spelling, grammatical errors, and types before you hand in anything. In addition, you will need to verify that (a) every in text reference citation is backed up by a matching entry in the references list and (b) every entry in the reference list is cited in at least one place in your text.

(d) Use a word processor: It is virtually a necessity to use a word processor for this project so that alterations can be made without extensive retyping of drafts.

(e) Show insight: Most published reviews go beyond simply reporting research results. Selectively, they also

- Compare / Contrast findings from different studies,
- Compare / Contrast methodologies used to arrive at those findings,
- Critique the methodologies nothing important strengths and / or weaknesses,
- Suggest extensions of studies and / or,
- Combine results or findings from multiple studies into an integrative picture or pattern.

These same approaches are crucial for creating a successful research report. The integration stressed in the last point above is especially important since literature reviews aim to crystallize reader's thinking about a broad topic.

(f) Submit a refined draft to your advisor: Once you have created a near - final draft, you may submit it to your advisor for review, Although you can expect a significant number of constructive suffusions to be made, you will be graded only on your final report. You are expected to respond to suggested improvements in your next draft. When you resubmit a draft, attach the previous marked up draft to it. Two or three submissions to your advisor may be necessary to achieve a polished product.

(g) Writing the study proposal: The study proposal section is placed after the literature review and just prior to your references list. The intent is to propose an empirical study and describe applicable methodology, normally the study will not be carried out unless the research report is used as a basis for a master's thesis. Here is a partial list of things to be discussed in the proposal:

(i) Research Question: Select an idea (or closely related set of ideas) for further research based upon the suggestions you have discussed near the end of your in terature review, discuss the research question, why it is important to investigate and the potential implications from having the research question answered.

(ii) Research Variables: Define study variables that can logically be used to investigate your research question. At the very least, you need to clearly identify one or more dependent or outcome measures. You may need to create operational definitions of constructs (e.g. motivation, career progress, turnover , performance etc). For most studies you will also need to identify predictors or independent variables that you would expect to influence the outcome measures (e.g. gender, salary, years of experience amount of training personal or situational characteristics etc.)

(iii) Population: Specifically describe the population to which the proposed study applies.

(iv) Hypotheses: Based on your research variables state one or more hypotheses about how the dependent or outcome measures should relate to (a) the predictors of independent variables or (b) each other be specific (e.g., involuntary disfunctional turnover will be lower when workers are paid on a piece - rate system than when they are paid at a fixed hourly rate). Justify your hypotheses by

- citing research you have included in your literature review,
- citing specific theories or frameworks by others and / or,
- creating a theory or model of your own of the mechanisms by which the outcome measures may be influenced.

(v) Subjects: Describe the number of subjects you propose to use, how they will be selected (including important subject screening characteristics) and how they will be recruited. Do they represent a sample or a population?

(vi) Design: Describe the analytical design of the study (e.g., A 4 x 2 (Training method x user friendliness) factorial analysis of variance (ANOVA) would be employed, with 10 subjects randomly assigned to each condition). You need not propose a classical between subjects experimental design within - subject, correlational, factor analytic, survey based and case study designs are also permissible if they lend themselves to the research question. You also need to describe the quantitative or statistical methods that would be applied to the data to test your hypotheses. If you propose an experiment, describe the levels. If the independent variable(s). Consult one or more statistics texts if you are unsure about design considerations.

(vii) Procedure: In a systematic way describe how your data will be collected. What stimulus materials will be used? How will they be administered? What constraints or safeguards are important? How long will the procedure take? Who will collect the data and what training or orientation do they need? If you propose an experiment, describe any unique administrative procedires within each condition. Also, describe how subjects will be briefed.

17.3 Oral vs Written Communication:

Though oral communication is similar to writing in some ways, it also has important differences.

Like a written paper, oral presentations should be prepared with the purpose, audience and occasion in mind. Because time is limited, the purpose may need to be more limited in scope. Also, because of the specific occasion for the speech, the tone and style can fit the particular audience even more specifically than most witting.

The organization of an oral report should be clear and logical, as any paper would be. But because listeners cannot go back and reread previous material, the structure needs to be especially obvious. The introduction should provide a clear overview of the reports purpose and structure. The body of the paper should have a limited number of sub - points, so that each one can be well supported in the time available. The conclusion needs so review the main points of the presentation and make evident their importance to the audience.

The delivery of the oral report has special significance. Visuals should be used and used well. They should be appropriate, clearly visible, simple to read and easy to use. Presenters should dress professionally have confident posture and make eye contact with the audience. The voice needs to be loud enough to be comfortably heard (use a microphone in a large room). Also presenters should vary their pitch and pacing and articulate clearly.

Finally, presenters need to know that to the audience, nothing they can say is important enough to merit going overtime. They need to plan ways to shorten or lengthen their material to fit time constraints necessary.

Because of these important differences, presenters need to prepare carefully and rehearse their presentations to perfect their delivery.

17.4 Statistical Abstract:

The statistical abstract is used when the issue is less complex and does not have the long range implications associated with a statistical report. The statistical abstract is shorter and less formal that the report form. Unlike the statistical report, the statistical abstract is seldom accompanied by an executive summary. The less complex nature of the issue the abstract is to address makes such a formal summary unnecessary.

Other than the executive summary, the abstract contains essentially the same features as the report. However, the components parts of the abstract are much less detailed and shorter in length. The statistical abstract can sometimes be presented in a single page. The following discussion of the abstract's main components reveals that each resembles those found in the statistical report, but in some what abbreviated form.

Introduction:

The introduction is a brief statement describing the motivation for the study. It explains what problem or concerns prompted the study and why the study is important. Little or no reference is made to histroical developments as was the case with the report form.

Methodology:

As with the report form, the methodological statement contained in the abstract describes in some technical detail the statistical tools and techniques that will be used to complete the study. This is perhaps the most technical component of the abstract. A brief description of the population and the manner in which the s ample was taken is customary.

Findings:

This section includes the actual statistical computaions and implements the statistical tools described in the methodology section. Due to the less involved, less complex nature of the problem, this section may consist of only a few calculations, which will serve as the basis for the study's conclusion. Brief commentary is provided regarding the outcome of the computations.

Discussion and Interpretation:

Relying on the findings in the previous section, the researcher presents a discussion of the study's findings and offers an interpretaion. This interpretation translates the technical findings for those who are less trained in statistical procedures.

Conclusion and Recommendation:

The abstract may be completed without a conclusion or any statement regarding recommendations. The study may have been requested by a superior who simply requires more information to make his or her own managerial decision. This superior may consider a recommendation for action as a usurpation of his or her administrative power, Remember, the abstract is used when the decision to be made is of lesser consequence; the decision can often be administered by a single authority. For this reason, a recommendation is not usually offered unless specifically requested.

17.5 Guidelines for Report Writing and Presentation:

1. Report Format:

According to the guidelines

- Line Spacing: 1.5
- Font: 12pt
- Typing: Back to Back

2. Report Contents:

The report should have the following:

(a) Inner cover page

(b) Certificate (signed by supervisor (s) begin numbering pages with this page as number (i)),

(This is required for only final report and not for mid-term report)

(c) Acknowledgment (S)

(d) Abstract (150 words) & Key words (max 6) [one page]

(e) Table of contents [include above items, title of all chapters, references, appendices, drawings, program listings]

(f) Nomenclature & Abbreviations [in alphabetical order followed by Greek symbols, superscripts, subscripts, underlined quantities etc.]

(g) Body of the report: The body should contain

Chapter 1: Introduction (Begin numbering pages with this page as number 1).

Chapter 2: Review of Literature and statement of problem.

Chapter 3: (Work done covering **Analytical Modelling, Employment of Software package[s] & other Computational algorithms, Equipment design, Simulation, Experimental verification, Equipment building** and any other aspect of the work you decide to mention should be given in this and subsequent chapters)

Chapter [last one]: Conclusions/ concluding remarks and scope for future work.

(h) References : and

(i) Appendices

3. References:

In the text, the references should be given in one of the following ways: Author's last name (if only one author) or both authors last names (if only two authors) or by first authors last name followed by et.al. The name(s) should be followed by the year in the brackets. In the list of references, the references should belisted in (a) alphabetical order of the author's name or (b) in chronological order and alphabetical order for each year. Some examples are given below: (a) Full reference:

(a) Kumar A, 1994, "Studies on Water Sprays "B.Tech project report Mech.Engg. Dept., IIT Delhi In the Text it should be mentioned as" : Kumar (1994)

(b) Full Reference:

Prasad, A.B., Kumar C.D., Johnes E.F., Chiu S.H., and Frost P. 1992, "Some studies in Engineering," J. Hypothetical Technology, V 72, N. 2, pp 82 - 90.

In Text : Prasad et.al. (1992)

(c) Full Reference:

Raman, A and Bashyam, T.C.A., 1991, Dynamics of M.Tech Theses Tata Mcgraw-Hill, New Delhi.

In Text: Raman and Bashyam (1991)

[Note that for books, publisher's name and place must be included]

4. Program Listings:

All computer programs developed should be presented in the Report in the following manner:

(a) Source Code listing with complete coding

(b) Program source code, executable files, data files etc.

(c) User's manual giving details of how to use the programs [including flow charts] and the system requirements to be bound with the report.

5. Drawings:

(a) All engineering drawings must confirm to the requirements of Bureau of Indian Standards Publication SP - 46/1998 Engineering Drawing Practice for Schools and Colleges (available at Bureau of Indian Standards, Bahadur Shah Jafar Marg, Delhi).

(b) If drawings are large they be included at the back of the report in a separate pocket.

(c) In case drawings are made using CAD pacakges, CD rom should be included which contains all the files and details of the packages used.

(d) All drawings must have the title block (see SP - 46/1968).

6. General:

(a) Each equation must be numbered and numbering should be sequential.

(b) Sketches, drawings, graphs and photographs should have a fig. Number and title.

(c) Page numbers at bottom centre

(d) All figure pages should also be numbered

(e) Figures to be preferable included where their reference occurs

(f) Figures and plates and equations to be numbered chapter wise refer to figure as "figure" notas fig... equation as equation and not "eq". Examples figure 2.3 shows..... in equation (2.3) .....

(g) Each table should be numbered.

(h) Margins: All text, drawings tables etc must be positioned on an A4 sheet with 25mm margin on the top, bottom and right side and 37mm margin on the left side.

(i) Make sure proper units SI as far as possible, appear wherever required.

(j) The report, which is submitted must be complete, error free and referable in future by academia. Use of spelling and grammar software is strongly r3ecommended. Read the draft of the report carefully before submission. All references, Figures, Tables, Equations, etc... Which are referenced in the text should be present in the Report and with the same numbering orreferencing. Conversely all References, Figures, Tables, Equations etc. Must be crossreferenced in the text. e.g. there should be no Figure in the report which is not referenced in the text. Spell check cannot identify correctly spelt words in the wrong context, e.g. If you have typed "he" in place of "the" then spell check will not indicate any mistake but them earning of the sentence would change drastically.

Some Guidelines for the Presentation of dissertation:

Prefer to make your presentation in Power Point as you would be able to demonstrate animations. Alternatively if advised by the program coordinator for making presentation on an OHP following guidelines may help:

1. Text of transparencies should be > = 18pt in black (other colours should be used only for highlighting).

2. Have all transparencies [preferably not more than 10] organized well in advance of your presentation.

3. Focus your presentation on the following:

   - Begin with a sheet showing sequence & contents of your presentation
   - Engineering problem formulation and your approach to solving it

- Highlights of the results and inferences of analytical modelling validations, employment of software packages[s] other computations, equipment design, simulation, experimental verification, equipment building and any other aspect of the work you decide to mention.
- Last sheet showing conclusion/ concluding remarks including utility, significance and major achievement of your work.

4. Use consistent units in your transparencies and talk.

5. Rehearse well before hand to finish latest by the time schedyled for you.

17.6 Audiovisual Techniques:

1. **CHALKBOARDS AND OVERHEAD TRANSPARENCIES** are the two most accessible audiovisual tools. Departmental offices usually have a supply of chalk and transparency film that can be used in any copier. Students can see transparencies better than chalkboards in large classrooms.

2. **FLIPCHARTS or EASELS** are an under - used audiovisual support for small classes or discussion groups. Flipcharts can be prepared ahead or used to record classroom discussion and are easily referred to again. They are less formal than overheads and can be torn off and posted so that results from several class discussions can be viewed at once. They are particularly useful for students to report results from small group discussions to the class as a whole.

3. **COMPUTERS, VIDEODISCS AND INTERACTIVE VIDEODISC TECHNOLOGY** are used in many courses to enhance student learning. In the classroom, portable computers linked to an LCD projection panel allow students to see the computer's display on a large screen.

17.7 Audiovisual Preparation:

Since chalkboard and transparencies are the most commonly used audiovisual tools, we will briefly address tips on using these resources well.

CHALKBOARD

K.I.L.L. (Keep It Large and Legible). Audiovisual tools are of no use if all students can't see them. Write legibly. Test by going to the back of the room to look.

PLAN AHEAD: When preparing a class it is useful to plan out what you will want to write on the chalkboard. This saves time and is generally clearer for the students. One way to do this it to outline information that goes on the board in the class notes.

MAKE EYE CONTACT FREQUENTLY An teacher writing on the board for extended periods of time has his or her back to the students. This behavior can result in losing control of the class. Instructors need to regularly observe student nonverbal behavior and use eye contact to keep students involved.

OVERHEAD TRANSPARENCIES:

LIMIT INFORMATION ON EACH TRANSPARENCY Since we can process no more than seven bits of information at a time, a single transparency should contain no more than four to six major points. Charts or tables with a great deal of information should be broken into smaller pieces using an enlarging copier. Small areas should be highlighted or colored to help students focus their attention.

GET EQUIPMENT AND TRANSPARENCIES PREPARED A teacher using overheads should plan to arrive early enough to get the equipment set up and focus the overhead. Transparencies should be arranged in order and numbered so that there is no need to fumble with them during the presentation.

BE A WARE OF STUDENT'S VIEW It can be hard to remember not to stand in front of the image and block student's view. A transparency students can't see is frustrating and distracting.

DON'T READ THE TRANSPARENCIES Talk about the material using the transparency as a launching pad rather than writing out the entire thought on it and reading it to the audience. Too much information on a single transparency is distracting and if students can simply read it, they do not need to pay attention to the teacher.

Students take very seriously whatever is written on the board or presented on transparencies. Be sure to allow them time to copy it all down, Be alert to occasions where you have given them too much to copy.

17.8 Making an Audio Visual Presentation:

Making Your Message Visual:

In today's visual society of TV computers and films, visuals are essential if you wish to make an impact. Mohamed Ali once said 'One in the eye is worth two in the ear'. While he was referring to boxing, this also applies to making presentations: A picture is worth of thousand words:

Designing Visuals:

Good design aids your credibility and helps with understanding. If you do not have the time or the skills to create your own audio - visual aids get somebody to do it for you. Use the communications department, a colleagure or a design agency. Keep all visual aids simple and uncluttered. Always take alonga series of low complexity aids (such as handouts) as a backup. Equipment can fail, so you might want to be prepared to go without audio-visual aids at all.

Whiteboards and Flipcharts:

Filpcharts are excellent to use when you need to record lists of ideas or need to record comments from the audience. Pages can be torn from filipchart and stuck around the room to create an expanding display.

The main drawback of using a whiteboard is the lack of any permanent record of what has been written, unless you are using the electronic whiteboard.

Overhead Projectors (OHPs):

The OHP was once widely used in business presentations but has gradually been replaced by computer - based displays. The main advantage of using an OHP is its ease of useit requires no warm - up time, there is little or no noise and the only searching is when the presenter looks for the side he or she wants.

Using Computer - based Displays:

Laptop presentations with data projectors are the norm these days. Although the equipment is more complex than an OHP and hence there is more to go wrong. They look professional and modern. They allow a smooth and imaginative transition between sides. The visuals can be sophisticated and incorporate the use of sound and video footage. The presenter can also check the visuals by looking at the laptop and not over his or her shoulder to the screen.

17.12 Effective Presentation Skills : Opening and Closing:

When you stand in front of an audience you have five seconds to get their attention and thirty seconds to develop interest and curiosity. How you open is critical to the success of your presentation. In those opening words you must hook your audience, establish report set the mood, demonstrate your credibility and introduce your topic. Yes, all this in seconds. You cannot afford to waste a moment or wing your opening.

Establishing Report:

Establish Report with your audience. Let them know you understand how they are thinking and feeling. For example : Some of you may be thinking you haven't got the time to be sitting here listening to me, but today provides real opportunites.....

Being Credible:

Your credibility is likely to be less to do with your academic qualifications and professional experiences and more to do with having a strong posture, quality eye contact and being enthused about your message.

Tips for closing:

Not only is the opening critical to the success of your presentation, so is your conclusion. Prepare your closing and know exactly what you are going to say and do. Create something memorable for the audience to take away with them a present, if you like. It needs to be purposeful and memorable and linked to your objectives, to what you want your audience to think, feel and do.

The Things to Do:

1. Feel - good ending:

Aim for something catchy - a story a phrase, a thought an image that will continue to play in people's minds for hours days or weeks. This is the time to deliver your take - home message straight to them eye to eye and person to person.

2. The closing summary:

The closing summary is useful when the presentation is intended to convey information and not a call to action. So we have seen that and this means .....

3. A call to action:

Remind your audience of the benefits of taking action and stir their emotions. Remember, feelings are a catalyst to action.

Beware of padding out your speech to the allotted time. If you finish early most people will consider this a bonus ! Rather than overstay your welcome, leave them wanting more.

Remember:

- Always plan your closing
- Make it catchy, brief and to the point
- Link to the main points of your talk
- Stand confidently and look directly at your audience
- Leave your audience wanting more
- Summarize the main points and answer the question now what ?

If your presentation is to be followed by a question and answer session, the impact of your final sentences can be diluted. You can counter this by a second very brief closing after accepting a series of questions.

How you end is how people will remember you. The lasting impression is formed from your final words, be they uplifting and motivating or empty and wishy - washy. Your closing is your signature. You might want to leave your audience feeling upbeat, needed and special.

17.10 Presentation Tips - Do's and Don'ts

How to Get Your Audience Wanting More:

1. Believe in your message.
2. Believe in yourself.
3. Open with impact.
4. Close on a positive note.
5. Let your personality shine through.
6. Flex to the style of your audience.
7. Answer the question your audience will be asking: What's in this for me?
8. Involve your audience and arouse their curiosity.
9. Use picture language.

10. Add variety and be creative.
11. Have an unusual slant or angle to your topic.
12. Spend as much time thinking about delivery and performance as you do about content.
13. Keep the focus on what you want rather than on what you don't want.

Ways to Destroy a Presentation:

1. Be unprepared
2. Relate to your material more than your audience
3. Apologies for yourself.
4. Repeat yourself.
5. Overload with information.
6. Tell a bawby joke.
7. Have titlle variety.
8. Read your oresentation.
9. Ignore time constratints.
10. Use slang or speak technical jargon.
11. Learn how the equipment works in front of the audience.
12. Make yourself so important the audience feels inadequate.
13. Direct your presentation at one or two people

17.11 Summary:

In this lesson we have explained the concept of report writing, report characteristics, oral as written communication, Guideliness for report writting, Audio visual techniques and effective presentation, Reports are considered to be legal documents in the work place and thus they need to be precise, accurate and difficult to misintepet.

17.12 Exercises:

1. What is a report writing? Explains with an example.
2. What are general report characteristics.
3. Compare oral vs written communication.
4. What are the guidliness for report writing.
5. Explain the audio visual techniques.
6. Give some important presentation tips Dos and Don'ts.

17.12 Reference Books:

1. Gupta S.P. and M.P. Gupta, 1988 Business Statistics, Sultan Chand & Sons., New Delhi.
2. Moskowitz H and G.P. Wright, 1985 Statistics for Management and Economics Charles E. Merill Publishing Company.
3. Gupta & Kapoor 1990 Fundamentals of Mathematical Statistics, New Delhi., S. Chand & Co.

Lesson Writer

Dr. K. CHANDAN