

Module-1: Nature & Scope of Research Methodology

Notes

Key learning Outcomes

At the end of this module, participant will be able to:

- Define research methodology
- Explain research process
- Identify different approaches of research
- Elaborate planning a research project
- Analyze Application of Research

Structure

Unit 1.1: Introduction to Research Methodology

- 1.1.1 What is research?
- 1.1.2 Characteristics of research
- 1.1.3 Differences between Research Methods and Research Methodology

Unit 1.2: Research Process

- 1.2.1 Research Process
- 1.2.2: Research Criteria
- 1.2.3 Different approaches of research

Unit 1.3: Research Approaches

- 1.3.1 Pure & Applied
- 1.3.2 Causal & Conceptual Research
- 1.3.3 Cross-sectional & Longitudinal Research
- 1.3.4 Experimental, Semi-experimental & Non-Experimental Research
- 1.3.5 Descriptive & Exploratory Research

Unit 1.4: Planning a Research Project

- 1.4.1 Identifying and Defining / Formulating the Research Problem
- 1.4.2 Factors Influencing the Complication of a Research Problem
- 1.4.3 Salient Features of Research Project
- 1.4.4 Formulation of a Research Project

Unit 1.5: Application of Research

- 1.5.1 Application of Research in Marketing
- 1.5.2 Application of Research in Finance

Notes

1.5.3 Application of Research in Human Resource Management

1.5.4 Application of Research in Production

1.5.5 Application of Research in Entrepreneurship

Amity University



Unit-1.1: Introduction to Research Methodology

Notes

Unit Objectives:

At the end of this unit, the participant will be able to:

- List various definitions of research
- Identify differences between research methods and research methodology
- Identify objectives, significance and types of research
- Definition and Objectives of Research

1.1.1 What is Research?

Authors and management gurus have defined research in different ways. Usually, a research is said to begin with a question or a problem. Research is defined as the generation of new concepts, methodologies, and understandings through the creation of new knowledge and/or the creative application of existing knowledge. This could include synthesising and analysing previous research to the point where it produces new and innovative results. By applying research we are able to find out the solutions of a problem with the application of systematic and scientific methods. You could talk about experimentation or innovation. You could use the word “risk” to describe the element of danger that comes with discovery. It is possible that investigation will lead to analysis. It is possible that you will conduct tests to prove your hypothesis. You could simply state that this work is unique and never seen. You could discuss what new knowledge will be gained because of your work.

You could talk about a new method or a new data source that will result in a breakthrough or a small improvement over current practise. You could state that it is a prerequisite for development in the sense of “research and development.”

Slesinger & Stephenson, Encyclopedia of Social Sciences:

“the manipulation of things, concepts or symbols for the purpose of generalizing to extend, correct or verify knowledge, whether that knowledge aids in the construction of theory or in the practice of an art”.

- **Redman & Mory:** “Systematized effort to gain new knowledge”. It is an academic activity and should be used in a technical sense.
- **Clifford Woody:** Research comprises “defining and redefining problems, formulating hypotheses or suggested solutions; collecting, organizing and evaluating data; making deductions and reaching conclusions; and finally, carefully testing the conclusions to determine whether they fit the formulated hypotheses”.

Research Objectives

General Objectives: General objectives, also known as secondary objectives, provide a detailed view of a study’s goal. In other words, by the end of your studies, you will have a general idea of what you want to accomplish. For example, if you want to investigate an organization’s contribution to environmental sustainability, your

Notes

broad goal could be to investigate sustainable practises and the organization's use of renewable energy.

Specific Objectives: Specific objectives define the primary aim of the study. In most cases, general objectives serve as the foundation for identifying specific goals. In other words, specific objectives are defined as general objectives that have been broken down into smaller, logically connected objectives. They assist you in defining the who, what, why, when, and how of your project. It's much easier to develop and carry out a research plan once you've identified the main goal.

Take, for example, a study of an organization's contribution to environmental sustainability. The specific goals will be as follows:

- Determine how the organisation has changed its practises and adopted new solutions throughout its history.
- To determine the impact of new practises, technology, and strategies on overall effectiveness.

1.1.2 Characteristics of Research:

- The research should concentrate on the most pressing issues.
- The investigation should be methodical. It emphasises the importance of following a structured procedure when conducting research.
- The research should follow a logical pattern. The scientific researcher cannot make much progress in any investigation without manipulating ideas logically.
- The study should be condensed. This means that a researcher's findings should be made available to other researchers so that they don't have to repeat the same research.
- The findings should be repeatable. This asserts that previous research findings should be able to be confirmed in a new environment and different settings with a new group of subjects or at a different time.
- The study should be fruitful. One of the most valuable characteristics of research is that answering one question leads to the generation of a slew of new ones.
- Action-oriented research is required. In other words, it should aim to find a solution that will allow its findings to be implemented.
- The research should take an integrated multidisciplinary approach, which means it will require research approaches from multiple disciplines.
- At all stages of the study, all parties involved (from policymakers to community members) should be invited to participate.
- The research must be straightforward, timely, and time-bound, with a straightforward design.
- The research should be as inexpensive as possible.
- The research findings should be presented in formats that are most useful to administrators, decision-makers, business managers, or members of the community.

1.1.3 Differences between Research Methods and Research Methodology

Although the names sound similar, both Research methods and Research methodology are different, as explained below:

Research Methods: The various procedures, schemes, steps, and algorithms used in research are known as research methods. The term “research methods” refers to all of the methods used by a researcher during a research study. They’re primarily planned, scientific, and value-agnostic. Observations, theoretical procedures, experimental studies, numerical schemes, statistical approaches, and so on are all examples of these. We can use research methods to collect samples, data, and come up with a solution to a problem. Business and scientific research methods, in particular, demand explanations based on collected facts, measurements, and observations, rather than solely on reasoning. They only accept explanations that can be verified through experiments.

Research Methodology: A systematic approach to solving a problem is known as research methodology. It is a science that studies how research should be conducted. Research methodology is essentially the procedures by which researchers go about their work of describing, explaining, and predicting phenomena. It can also be defined as the study of methods for gaining knowledge. Its goal is to provide a research work plan.

Notes

Notes**Unit-1.2: Research Process****Unit Objectives:**

At the end of this unit the participant will be able to:

- Identify and learn about process of research
- List different approaches of research: deductive, inductive, quantitative, qualitative, etc.
- Define hypothesis formulation and its types
- Describe various hypothesis errors

1.2.1 Research Process

In the early decades human inquiry was primarily based on the examination of one's own conscious thoughts and feelings that means the observation of any one and understanding through the logical discussion to seek the truth. This procedure was accepted for a millennium and was a well-established conceptual framework for understanding the world. The knowledge seeker was an integral part of the inquiry process. With time, this part was changed. The Scientific method introduced several major components in research procedure like Objectivity.

At every stage of the marketing research process, systematic planning is required. Each stage's procedures are methodologically sound, well documented, and, to the extent possible, planned of time. The scientific method is used in marketing research, in which data is collected and analysed to test preconceived notions or hypotheses.

Marketing research should be conducted impartially to provide accurate information that reflects the true situation. While the researcher's research philosophy will always influence the research, it should be free of the researcher's or management's personal or political biases.

For example, we may find out that our topic is too broad and needs to be narrowed, sufficient information resources may not be available, what we learn may not support our thesis or the size of the project does not fit the requirements.

There are main nine steps of research process that are followed at the time of designing a research project. They are as follows.

Step 1: Problem Definition

Step 2: Development of an Approach to the Problem

Step 3: Research Design Formulation

Step 4: Field Work or Data Collection

Step 5: Data Preparation and Analysis

Step 6: Report Preparation and Presentation

1.2.2: Research Criteria

- The research's purpose should be clearly defined, and common concepts should be used.
- The research procedure should be sufficiently described. detail to allow another researcher to continue the research for further advancement while maintaining the integrity of what has already been accomplished.
- The research's procedural design should be meticulously planned to produce objective results.
- The researcher should be completely honest about any flaws in the procedure design and estimate their impact on the findings.
- The data analysis should be sufficient to reveal its significance, and the analysis methods used should be appropriate. The data's validity and reliability should be double-checked.
- Conclusions should be limited to those that are supported by the research data and for which the data provide an adequate foundation.
- If the researcher is experienced, has a good research reputation, and is a person of integrity, greater trust in the research is warranted.

1.2.3 Different Approaches of Research:

Deductive: Deductive approach starts with developing a hypothesis based on existing theory, and then prepares a research strategy for testing the hypothesis. The process of going from particular to reasoning is called deductive . When a link or a casual relationship is implied by a particular theory or case example then it might be true in many cases. Deductive approach can be defined by the means of hypotheses or derived from the propositions of the theory. This approach is about conclusions being deduced from propositions or premises. This will start with a pattern "that is tested against observations", whereas induction "begins with observations and seeks to find a pattern within them".

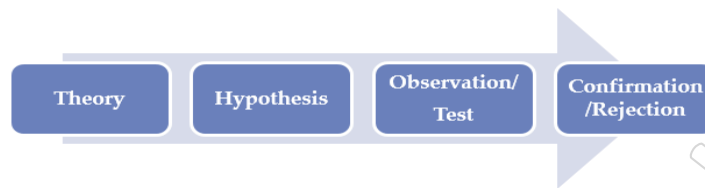
Advantages of deductive:

- Probability to clarify the relationships between variables and concepts.
- Probability to measure concepts in quantitative manner.
- Probability to generalize research findings to a certain extent.

Deductive approach mostly works in the given ways: Having hypothesis deduced from theory

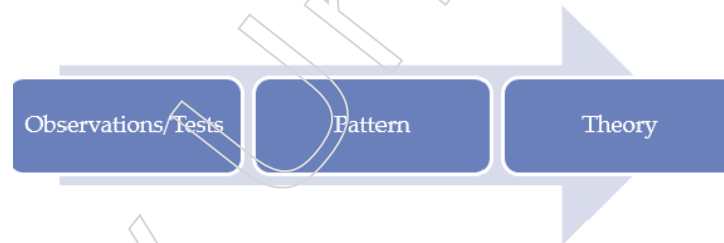
- Using operational terms to formulate hypothesis as well as suggesting relationships between two particular variables.
- Using relevant methods to test hypothesis such as quantitative methods, like regression and correlation analysis, along with the mean, mode and median etc.
- making decision to either confirm or reject on the basis of the result examined since it is essential to compare research findings against literature review findings.
- Theory modification when hypothesis cannot be confirmed.

Notes



Inductive: Inductive approach or inductive reasoning, starts with observations and theories are proposed which are related to the end of the research process as a result of the observations. It involves the search for pattern from observation and the development of explanations – theories – for those patterns through series of hypotheses. In this discipline of studies, at the start of research, both hypothesis and theories are not applicable. The researcher in this case, is free to make alterations in the study direction even if it is after the start of the research process.

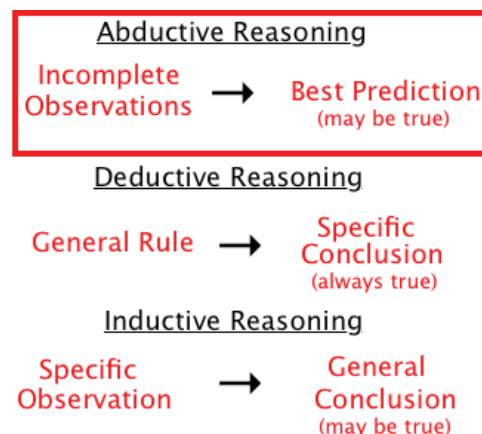
This approach doesn't disregard theories at the time of formulating questions and objectives for the research. Inductive approach helps to generate meanings from the data set collected in order to identify patterns and relationships to build a theory. This approach is mainly based on learning from experience. Previous patterns, resemblances and regularities are observed in order to reach conclusions or to generate theory.



Abductive Reasoning or Abductive Approach

Abductive reasoning is set to address weaknesses associated with deductive and inductive approaches, specifically for deductive reasoning so that 'how to select theory can be tested via formulating hypotheses' and for Inductive reasoning "no amount of empirical data will necessarily enable theory-building". It acts as a third alternative, overcomes these weaknesses via adopting a pragmatist perspective.

The figure below illustrates the main differences between abductive, deductive and inductive reasoning:



It is cleared that abductive reasoning is similar to deductive and inductive approaches in a way that it is applied to make logical inferences and construct theories.

Abductive approach starts with 'surprising facts' or 'puzzles' and the research process is devoted in their explanation. The 'Surprising facts' or 'puzzles' may emerge, when a researcher encounters empirical phenomena, which cannot be explained by the existing theories. In that approach, researcher searches for the 'best' explanation among many alternatives to choose. Researcher can combine both, numerical and cognitive reasoning for explaining 'surprising facts' or 'puzzles'.

Qualitative research: It is a non-statistical process of inquiry. It helps with in-depth understanding of problems or issues in their natural settings. It is highly dependent on the experience of the researchers and the questions used to probe the sample. The sample size is usually restricted in between a minimum of 6 and a maximum of 10 people. Open-ended questions work to (To get maximum information from a given sample) encourages answers which leads the researcher to another question or more questions. The below methods are used for qualitative research:

- One-to-one interview
- Focus groups
- Ethnographic research
- Content/Text analysis
- Case study research

Quantitative research: It is a structured process of data collection and analysis in order to draw conclusions. This method uses a computational and statistical process to collect and analyze data. Quantitative data is all about numbers. It involves a larger population as more people will bring more data to the table, which helps to obtain more accurate results. This research uses close-ended questions because the researchers are typically looking to gather statistical data. It involves use of data collection tools likes online surveys, questionnaires, and polls .There are various methods of deploying surveys or questionnaires. Online surveys helps surveyor to reach large number of people or smaller focus groups for different types of research that meet different goals.

Advocacy/participatory approach to research:

Sometimes researchers adopt an advocacy/participatory approach which do not respond to the needs or situation of people from marginalized or vulnerable groups. When researcher wants to bring about a positive change in the lives of the research subjects, it is sometimes described as emancipatory. It is not a neutral stance. Researchers want their research to directly or indirectly result in some kind of reform, for which they involve the group being studied in the research at all stages, so as to avoid further marginalizing them.

The researchers may adopt a less neutral position than that which is usually required in scientific research. This might involve interacting informally or even living amongst the research participants (the co-researchers). The searching of the research can be reported in more personal terms, often using the precise words of the research participants.

Notes

Unit-1.3: Research Approaches

Unit Objectives:

At the end of this unit the participant will be able to:

- Identify and learn about process

1.3.1 Pure & Applied

Applied Research: Applied research is a type of study that aims to solve a specific problem or offer novel solutions to issues that affect a person, a group, or a society. Because it involves the practical application of scientific methods to everyday problems, it is often referred to as a scientific method of inquiry or contractual research.

When conducting applied research, the researcher pays special attention to identifying a problem, developing a research hypothesis, and then conducting an experiment to test these hypotheses. In many cases, empirical methods are used in this research approach to solve practical problems.

Because of its direct approach to finding a solution to a problem, applied research is sometimes considered a non-systematic inquiry. It's a type of follow-up research that digs deeper into the findings of pure or basic research in order to validate them and use them to develop innovative solutions.

Applied Research Example in Business

- Applied research to improve the hiring process in a company.
- Applied research aimed at improving workplace efficiency and policies.
- Workplace skill gaps are being addressed through applied research.
- Applied Research Examples in Education
- An investigation into how to improve teacher-student engagement in the classroom.
- An investigation into how to improve a school's student readiness.
- A research project aimed at piquing students' interest in mathematics

Pure/Fundamental Research or Basic Research

A type of research approach aimed at gaining a better understanding of a subject, phenomenon, or basic law of nature is known as basic research. The goal of this type of research is to advance knowledge rather than to solve a specific problem.

Pure research or fundamental research are other terms for basic research. Between the late 19th and early 20th centuries, the concept of basic research arose as a means of bridging the gaps in science's societal utility.

Basic research can be exploratory, descriptive, or explanatory in nature; however, in many cases, it is explanatory. The primary goal of this research method is to collect data in order to improve one's understanding, which can then be used to propose solutions to a problem.

Basic Research Example in Education

- How does human retentive memory?
- How do different teaching methods affect students' concentration in class?

1.3.2 Causal & Conceptual Research:

Causal research is conducted to determine a cause and effect relationship between two variables.

Example: Effect of promotional events on sales

Correlational Research

Correlational research is a type of non-experimental research method in which a researcher measures two variables, understands and assesses the statistical relationship between them with no influence from any extraneous variable. Our minds can do some brilliant things.

Examples of Causal Research (Explanatory Research)

- To determine the effects of foreign direct investment on Taiwanese economic growth.
- To investigate the impact of rebranding initiatives on customer loyalty.
- To determine the nature of the impact of work process re-engineering on employee motivation levels.

Did u know? Conceptual Research is related to certain abstract ideas or theories that are often applied by philosophers to develop new concepts or to rework on the existing ones.

1.3.3 Cross-sectional & Longitudinal Research:

Cross-sectional study

The longitudinal and cross-sectional studies are both observational studies. This means that researchers record data about their subjects without tampering with the research environment. We would simply measure the cholesterol levels of daily walkers and non-walkers, as well as any other.

Characteristics of interest to us, in our study. We would not persuade non-walkers to start walking or advise daily walkers to change their habits. In a nutshell, we'd try not to get in the way.

A cross-sectional study is distinguished by the ability to compare different population groups at a single point in time. Consider it like taking a photograph. The findings are based on whatever fits into the frame.

To return to our previous example, we could compare cholesterol levels in daily walkers in two age groups, over 40 and under 40, to cholesterol levels in non-walkers in the same age groups. We might even create gender subgroups. We would not, however, consider past or future cholesterol levels because they would be outside the scope. We'd only examine cholesterol levels at a single point in time.

Notes

A cross-sectional study design has the advantage of allowing researchers to compare multiple variables at once. We could look at age, gender, income, and educational level in relation to walking and cholesterol levels, for example, with little or no extrapolation.

Cross-sectional studies, on the other hand, may not provide conclusive evidence of cause-and-effect relationships. This is because such studies provide a snapshot of a single moment in time and do not consider what occurs before or after the snapshot. As a result, we can't say for sure whether our daily walkers had low cholesterol levels before starting their exercise routines or if the daily walking behaviour helped to lower cholesterol levels that were previously high.

Longitudinal study

A longitudinal study is observational, just like a cross-sectional one. As a result, researchers do not interfere with their subjects once more. A longitudinal study, on the other hand, involves researchers making multiple observations of the same subjects over a long period of time, sometimes many years.

A longitudinal study has the advantage of allowing researchers to detect changes in the characteristics of the target population at both the group and individual level. The important thing to remember is that longitudinal studies go beyond a single point in time. As a result, they can create event sequences.

To return to our example, we could examine the change in cholesterol levels among women over 40 who have walked every day for the past 20 years. The longitudinal study design would take into account cholesterol levels at the start of a walking programme and as the programme progressed. As a result of its scope, a longitudinal study is more likely than a cross-sectional study to suggest cause-and-effect relationships.

In general, the design should be driven by the research. However, the progression of the research can sometimes aid in determining which design is best. Longitudinal studies take longer to complete than cross-sectional studies.

1.3.4 Experimental, Semi-experimental & Non-Experimental Research:

Experimental: Experimental research is a type of study that employs a scientific approach to manipulate one or more control variables of the research subject(s) and then measure the impact of the manipulation on the subject. It is well-known for allowing the manipulation of control variables.

Even though it can be difficult to execute, this research method is widely used in a variety of physical and social science fields. They are far more common in information systems research than in library and information management research within the information field.

When the goal of the research is to trace cause-and-effect relationships between defined variables, experimental research is usually used. The type of experimental research chosen, on the other hand, has a significant impact on the experiment's outcomes.

Semi Experimental: The prefix quasi means “similar to.” As a result, quasi-experimental research is research that resembles experimental research but isn’t actually experimental. Participants are not randomly assigned to conditions or orders of conditions, despite the fact that the independent variable is manipulated (Cook & Campbell, 1979). 1st The directionality problem is eliminated in quasi-experimental research because the independent variable is manipulated before the dependent variable is measured. However, because participants are not assigned at random, there is a chance that other differences exist between conditions. Thus, quasi-experimental research does not eliminate the problem of confounding variables.

Example: Non-equivalent groups design just like we hypothesized a new after-school program which assured the students of higher grades. We divided two similar groups of children who attend different schools, one of which implements the new program while the other does not.

In field settings where random assignment is difficult or impossible, quasi-experiments are most likely to be conducted. They’re frequently used to assess the efficacy of a treatment, such as psychotherapy or an educational intervention. There are numerous types of quasi-experiments, but we will focus on a few of the most common here.

Non-Experimental: Non-experimental research is defined as research in which no control or independent variable is manipulated. Researchers in non-experimental research measure variables as they occur naturally, without any further manipulation.

When the researcher doesn’t have a specific research question about a causal relationship between two variables and manipulating the independent variable is impossible, this type of research is used. They’re also useful for:

- It is impossible to assign subjects to conditions at random.
- The subject of the study is a causal relationship, but the independent variable cannot be changed.
- The study is broad and exploratory in nature.
- The study focuses on a variable-to-variable non-causal relationship.
- Only a limited amount of information about the research topic is available.

1.3.5 Descriptive & Exploratory Research:

Descriptive research

Descriptive research is defined as a research method that describes the characteristics of the population or phenomenon studied. This methodology focuses more on the “what” of the research subject than the “why” of the research subject.

The descriptive research method primarily focuses on describing the nature of a demographic segment, without focusing on “why” a particular phenomenon occurs. In other words, it “describes” the subject of the research, without covering “why” it happens.

For example, surveys related to frequency of shopping, food habits, product preference, etc. are examples of descriptive research. Example: A group of market researchers aims at identifying the impulse buying trends among various households

Notes

pan India. The researchers would focus on collecting data related to “what is the impulse buying pattern of Indian consumers” and the scope of their research would be limited to that. The research does not explain the underlying reasons behind such impulse buying practices or “why” such buying pattern exists. Here, the scope of the research is just to report the existence of such buying trends and not why do people resort to impulse buying. This is, hence, an ideal example of descriptive research.

Exploratory research

Exploratory research is the investigation of a problem that has not previously been studied or thoroughly investigated. Exploratory research is usually done to gain a better understanding of the problem at hand, but it rarely yields a conclusive result.

Exploratory research is used by researchers when they want to learn more about an existing phenomenon and gain new insights into it in order to formulate a more precise problem. It starts with a broad concept, and the research findings are used to uncover related issues to the research topic.

The process of exploratory research varies depending on the discovery of new data or insight. The results of this research, also known as interpretative research or grounded theory approach, provide answers to questions like what, how, and why.

Exploratory Research Example on Product Research

When developing a new product or service, companies conduct two types of research. The first is carried out prior to the development of the product, while the second is carried out after it has been developed.

The exploratory research conducted after product development will be the focus of our attention. It's known as the beta testing stage of product development for tech products.

For example, if a new feature is added to an existing app, product researchers will want to see how well the feature is received by users. The research is not exploratory if the feature added to the app is something that already exists.

If Telegram adds a status feature to its app, for example, the app's beta research stage is not exploratory. This is because this feature is already available, and they can easily obtain sufficient information from WhatsApp.

When it comes to a new feature, such as Snapchat filters when they first launched, the research is instructive. A focus group of beta testers is used to conduct exploratory research in this case.

Unit-1.4: Planning a Research Project

Notes

Unit Outcomes:

At the end of this unit, participants will be able to:

- List the methods of identifying and defining the research problem
- Identify the factors influencing the complication of a research problem
- Identify the salient features of a research project

1.4.1 Identifying and Defining / Formulating the Research Problem

Prior to any stage, it is important to go through selection and then defining the problem in a research process. The researcher needs to identify the problem in order to have it formulated, and then make it suitable to research. Usually, research problem is an unanswered question that is encountered by the researcher in regards to a practical or theoretical situation, for which he needs a solution. Kothari states research problem to exist if any of the given is noticed:

- There is an organization or an individual (X) who is facing the problem. The organization or the individual has an environment (Y) and is affected with variables that are beyond control (Z).
- There needs to be two courses of action that need pursuing (the least) which are (A1 and A2). These are defined by one or sometimes more values related to the controlled variables.
- The above mentioned courses of action need to have two alternative and possible outcomes at the least (B1 and B2). One of these will be preferred more than the other, which is what the researcher wants and this becomes the objective.
- The possible courses of action that are available must yield a way to the researcher to have the objective achieved but not the exact chance. So, if $P(B_j / X, A, Y)$ represents the probability of the occurrence of an outcome B_j when X selects A_j in Y, then $P(B_1 / X, A_1, Y) \neq P(B_1 / X, A_2, Y)$. From this we get that the choices must not have equal efficiencies for the desired outcome.

When taking the above into consideration, the individual or organization may reach the research problem only if, X has no idea of the best course of action. In other words, X should have a doubt about the solution.

An individual or a group of persons face a problem if there is more than one desired outcome. It is required to have two or more alternative courses of action, which have some but not equal efficiency. This is needed for probing the desired objectives, if there is doubt that the best course of action is yet to be taken. Research problem components can be summarized as:

- A group or an individual facing some problem.
- One objective at least that should be pursued otherwise there cannot be a problem.

Notes

- Alternate options of object pursuing should be met which will allow the researcher to have more than one alternative. Otherwise without the choice of alternative options, there won't be a problem for the researcher. T
- The researcher needs to have doubt on the alternative means and making a selection. This means that the researcher needs to have answer for relative frequency or suitability question, pertaining to its alternatives that are possible.
- A context is needed attributed to the difficulty faced. T

Thus, identification of a research problem is something that happens even prior to conducting research. Research problem requires a researcher to look up for the best available solution to the given problem. This means researcher needs to find out the best course of action through which the research objective may be achieved optimally in the context of a given situation.

1.4.2 Factors Influencing the Complication of a Research Problem:

There are factors that can complicate a research problem.

- Changes in environment which affect the efficient alternative courses of action taken or the quality of the outcomes.
- Available alternative courses may be a lot and the person who isn't involved in decision making might get affected with the environment change. His reaction can be favourable or unfavourable.
- There are different similar factors that may cause these changes related to the research context. All of these can be thought of and considered from the point of view of a research problem.

1.4.3 Salient Features of Research Project

A research project can have the following noticeable features:

- Forming a plan that identifies the types and sources of information that research problem needs.
- It strategises in a specific manner all the methods of data collection and analysis which would be adopted.
- It also mentions the time period of research along with monetary budget needed in conducting the study, which include the two major constraints of undertaking any research.

1.4.4 Formulation of a Research Project

Post research problem is defined, it essential to prepare the design of the research project, popularly termed as 'research design'. Its function is to decide upon issues like what, when, where, how much, by what means, etc., when it comes to an enquiry or a research study. In this conditions are arranged to help towards collection and analysis of data to facilitate the combining of relevance to the research purpose with economy in procedure.

Seltiz and others stated that, this is the conceptual structure within which research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data. We can conclude that research design offers an outline of what the researcher plans to execute in terms of framing the hypothesis, its operational implications and the final data analysis. Particularly, the research design highlights decisions which include:

- The nature of the study
- The purpose of the study
- The location of the study that is to be conducted
- The nature of data required
- Source of data that is to be collected
- Time period of the study
- Sample design type that can be used
- Data collection techniques that are usable
- Data analysis methods that can be applied
- Structure of the report

Taking into consideration the research design decisions, the overall research design may be divided into the following (Kothari 1988):

- The sampling design that is related to the method of selecting items to be taken into observation for selected study.
- The observational design that one can relate to conditions under which the observations are to be made.
- The statistical design that regarding the question of how many items are to be observed, also the manner in which, information and data gathered are to be analyzed.

The operational design which is about the techniques, using which the procedures mentioned in the sampling, statistical and observational designs can be carried out.

Notes

Unit-1.5: Application of Research

Unit Objectives

At the end of this unit, the participants will be able to:

- Identify various field of application of research
- Describe how to help for building a new product.
- Explain how to promote a product
- Analyze how to help entrepreneurs.

1.5.1 Application of Research in Marketing

Research may be used in the area of marketing

Product development and distribution issues are discussed, as well as marketing institutions, marketing policies and practises, consumer behaviour, advertising and sales promotion, sales management, and after-sales service. One of the most popular and well-established areas is marketing research. Market potentials, sales forecasting, product testing, sales analysis, market surveys, test marketing, consumer behaviour studies, and marketing information systems are all examples of marketing research.

1.5.2 Application of Research in Finance

Here we discuss the various application of research in modern finance. Applying research we can develop our competency in applying statistical and various econometric techniques to solve various problems in finance. Research is much more essential if we are working or planning to work, in the finance sector. It can also be helpful if we work, or intend to work, outside of the finance domain such as conducting academic research in finance. Research basically takes responsibilities on to

- Finding out suitable sources of finance
- Making effective investment decisions on company's behalf
- Preparing budget
- Forecasting costs and profits
- Framing dividend policies

1.5.3 Application of Research in Human Resource Management

In HRM, Research is used to evaluate HR practices and performance. By using research we can analyses the collected information and drawing conclusions for decision-making. Sometimes the research could also be advanced, hoping on sophisticated designs and statistics. But whether information is rigorous or not, research seeks to boost the performance. Here is the subsequent area of HRM, where the research technique is applicable. By using research technique research team helps the organization-

- To stay updated on:
- latest labour laws
- wage rates
- employment trends and best practices

To study:

- Incentive schemes
- Cost of living
- Employee turnover rates
- Performance appraisal techniques

Planning manpower and utilising human resources effectively

- Framing human resources policies for the organisation
- Compares its organization / division with another organization / division to uncover areas of poor performance that need to improve

Relies on the expertise of a consultant to diagnose the causes of problems

With the help of existing records generates statistical standards against which activities and programs are evaluated

- With the human resource information system taken care of laws and company policies or procedures.
- MBO (management by objectives) is applied to compare between the actual results and stated objectives.

1.5.4 Application of Research in Production

Planning, organising, staffing, communicating, coordinating, motivating, and controlling are all management functions. Research has resulted in a variety of motivational theories. Production (also known as manufacturing) research is more concerned with materials and equipment than with human factors. It covers a wide range of topics, including developing new and better ways to produce goods, developing new technologies, lowering costs, and improving product quality.

1.5.5 Application of Research in Entrepreneurship

Inventors and entrepreneurs who are just starting out are frequently preoccupied with the details of developing their new product or service. Market research is a crucial component that allows them to take a step back and consider how their product might fit into the marketplace. Entrepreneurs gain valuable information about industry trends, who their true competitors are, and which consumers they should target and how through market research. Market research aids start-up entrepreneurs in developing, fine-tuning, and improving their specific product or service, which leads to increased revenue from new customers.

Notes

Summary:

At the end of this module, the participants have covered:

- Defining research methodology
- Explaining research process
- Identifying different approaches of research
- Elaborating planning a research project
- Analyzing Application of Research

Exercise:

1. The purpose of research is to find solutions through the application of and different methods.
 - a) Synthesizing and Analyzing
 - b) Applying and interpreting
 - c) Both and b
 - d) none of the above
2. Which of the following scopes of research is related to human resource development?
 - a) Projecting demand
 - b) Studying performance appraisal techniques
 - c) Cost budgeting
 - d) Measuring effectiveness of promotional activities
3. Which of the following scopes of research is NOT exclusively related to the framing of government policies?
 - a) Evolving the union finance budget
 - b) Modifying the five-year plan
 - c) Revising fiscal policies
 - d) Revising monetary policies
4. _____ is a crucial component that allows Inventors and entrepreneurs to take a step back and consider how their product might fit into the marketplace
 - a) Market research
 - b) Product research
 - c) Demand research
 - d) none of the above
5. Planning, organising, staffing, communicating, _____, _____ and _____ are all management functions
 - a) coordinating
 - b) motivating

Business Research Methods

21

- c) controlling
- d) all of the above

Answers:

1. a) Synthesizing and Analyzing
2. b) Studying performance appraisal techniques
3. d) Revising monetary policies
4. a) Market research
5. d) all of the above

Notes



Amity University

Module-2: Research Methods & Data Collection Techniques

Key learning outcomes

At the end of this module the participant will be able to:

- Analyze Research Modelling
- Define Data Collection and its Methods
- Explain Questionnaire Designing
- Describe Measurement and Scaling
- Analyze Sampling

Structure

Unit 2.1: Research Modelling

- 2.1.1 Types of Research Models
- 2.1.2 Importance of Research Model
- 2.1.3 Types of Research Models
- 2.1.4 Stages of a Research Model
- 2.1.5 Heuristic Research Model
- 2.1.6 Simulation Research Modelling

Unit 2.2: Data Collection and its Methods

- 2.2.1 Introduction to Data Collection
- 2.2.2 Types of Data Collection Methods
- 2.2.3 Tabulating and Validating the Collected Data

Unit 2.3: Questionnaire Designing

- 2.3.1 Introduction to Questionnaire
- 2.3.2 Steps to be followed for constructing a Questionnaire
- 2.3.3 Types of Questions to be asked in a Questionnaire
- 2.3.4 Format of a Questionnaire

Unit 2.4: Measurement and Scaling

- 2.4.1 Introduction to Measurement and Scaling Techniques
- 2.4.2 Types of Scaling Techniques
- 2.4.3 Attitude Measurement Scales

Unit 2.5: Sampling

- 2.5.1 Introduction to Sampling

- 2.5.2 Sampling Plan and Sampling Frame
- 2.5.3 Steps involved in Sampling Process
- 2.5.4 Simple Random Sampling
- 2.5.6 Sampling and Non-Sampling Errors

Notes

© Amity University

Notes

Unit-2.1: Research Modelling

Unit Objectives:

At the end of this unit, you will learn:

- Use of Research Models
- Importance of Research Models
- Types of Research Models
- Stages of Research Model
- Heuristic Research Model
- Simulation Research Model
- Data Considerations while analyzing Data for a Research

2.1.1 Types of Research Models

Research Models are classified broadly into two types as mentioned below:

- Qualitative Research Model
- Quantitative Research Model

Qualitative Research Model

It involves non-numerical data collection and analysis in order to understand concepts, opinions and experiences. This helps to gather in-depth insights into a problem or develop new ideas for research. Qualitative research finds its use mostly in the humanities and social sciences, in subjects such as anthropology, sociology, education, health sciences, history, etc. Qualitative research helps to visualize how people can experience the world. While there are many approaches to qualitative research, they are less desirable as they are flexible and focus on retaining rich meaning when interpreting data.

Quantitative Research Model

It is about collecting and analysing numerical data. Used for locating and defining patterns and averages, this research model can make predictions, test causal relationships, and help to generate results to wider populations. Quantitative research finds a wide use in the natural and social sciences: biology, chemistry, psychology, economics, sociology, marketing, etc.

2.1.2 Importance of Research Model:

The importance of using a research model is highlighted below:

- Model building is an integral part of the research design because models guide both theory development and research design.
- Models seem appropriate to the worlds of computers, biotechnology, and automation, and they have conferred new status on the scientist in government, industry, and the military.

- Models are also very important to social scientists because they provide a framework through which important questions are investigated.

2.1.3 Types of Research Models

Research Models are classified broadly into two types as mentioned below:

- Qualitative Research Model
- Quantitative Research Model

Qualitative Research Model

It involves non-numerical data collection and analysis in order to understand concepts, opinions and experiences. This helps to gather in-depth insights into a problem or develop new ideas for research. Qualitative research finds its use mostly in the humanities and social sciences, in subjects such as anthropology, sociology, education, health sciences, history, etc. Qualitative research helps to visualize how people can experience the world. While there are many approaches to qualitative research, they are less desirable as they are flexible and focus on retaining rich meaning when interpreting data.

Quantitative Research Model

It is about collecting and analysing numerical data. Used for locating and defining patterns and averages, this research model can make predictions, test causal relationships, and help to generate results to wider populations. Quantitative research finds a wide use in the natural and social sciences: biology, chemistry, psychology, economics, sociology, marketing, etc.

2.1.4 Stages of a Research Model

These steps are: (1) choosing a topic, (2) defining the problem, (3) reviewing the literature, (4) formulating a hypothesis, (5) selecting a research method, (6) collecting data, (7) analysing the results, and (8) sharing the findings.

Other authors may identify more or fewer steps, but the fundamental model remains the same. Validity and reliability are two important aspects of research. Validity refers to whether or not the research actually measures what it claims to. The degree to which research produces consistent or dependable results is referred to as reliability.

Sociologists use six different research methods to conduct their studies: (1) surveys, (2) participant observation, (3) secondary analysis, (4) documents, (5) unobtrusive measures, and (6) experiments. Resources, access to subjects, the purpose of the research, and the researcher's background all play a role in how sociologists choose their research methods.

2.1.5 Heuristic Research Model

There are a wide range of qualitative research models available and one of the lesser-known models is the Heuristic research model. This research model was developed by Clark Moustakas (an American psychologist and researcher). The name, Heuristic was derived from the Greek work 'Heuriskein' (which means discover, find). The research model has six phases

Notes

- Initial engagement
- Immersion
- Incubation
- Illumination
- Explication
- Creative synthesis

Shelly Chaiken developed the heuristic-systematic model of information processing (HSM), which attempts to explain how people receive and process persuasive messages. Individuals can process messages in one of two ways, according to the model: heuristically or systematically. Heuristic processing, on the other hand, entails the use of simplifying decision rules or “heuristics” to quickly assess the message content, whereas systematic processing entails the careful and deliberate processing of a message. This model’s guiding belief is that people are more likely to use heuristics instead of cognitive resources, which affects message intake and processing. The elaboration likelihood model, or ELM, is very similar to the HSM. Both models were developed primarily in the early to mid-1980s, and they share many of the same concepts and ideas.

2.1.6 Simulation Research Modelling

By using statistical descriptions of the activities involved, stimulation models attempt to replicate the workings and logic of a real system. For example, a line might produce 1000 units per hour on average. If we assume this is always the case, we lose sight of what happens when there is a breakdown or a stoppage for routine maintenance, for example. When we consider the effect on downstream units, the effect of such a delay may be amplified (or absorbed).

‘Entities’ (e.g. machines, materials, people, etc.) and ‘activities’ are two types of entities in a simulation model (e.g. processing, transporting, etc.). It also includes an explanation of the logic that governs each activity. A processing activity, for example, can only begin when a certain quantity of working material, a person to operate the machine, and an empty conveyor to transport the product are all available. Once an activity has begun, the time it will take to complete it is calculated, which is frequently done using a sample from a statistical distribution.

Unit-2.2: Data Collection and its Methods

Notes

Unit Outcomes

At the end of this unit, you will learn:

- Introduction to Data Collection
- Types of Data Collection Methods
- Tabulating and Validating the Collected Data

2.2.1 Introduction to Data Collection

The researcher should know data sources that he/she requires for all purposes. Data or information is of two types:

- Primary Data
- Secondary Data

Information gathered through original or first-hand research is referred to as primary data. Surveys and focus group discussions, for example. Secondary data, on the other hand, is information that has already been gathered by someone else. For instance, internet research, newspaper articles, and company reports.

Any study's goal determines whether primary or secondary data will be collected. For example, if a company wants to enter the women's apparel market in India and wants to know the size of the market, it can use secondary data such as industry reports and newspaper articles, whereas if it wants to learn about consumer preferences for a new type of fabric or style, it must conduct primary research.

Primary data collection is usually more expensive and time-consuming than secondary data, but it serves a specific purpose and helps to eliminate biases.

2.2.2 Types of Data Collection Methods

There are various methods to collect the two sources of data (Primary and Secondary) as mentioned and explained below:

Primary data is gathered from first-hand experience and has never been used before. The data gathered through primary data collection methods is highly accurate and specific to the research's purpose.

Quantitative and qualitative data collection methods are the two types of primary data collection methods.

Quantitative Methods:

Time Series Analysis: A time series is a sequential order of values of a variable at equal time intervals, also known as a trend. An organisation can forecast demand for its products and services for the future using patterns.

Smoothing Techniques: Smoothing techniques can be used when the time series lacks significant trends. They get rid of the random variation in historical demand. It aids in the identification of patterns and demand levels in order to forecast future demand.

Notes

The simple moving average method and the weighted moving average method are the two most common methods for smoothing demand forecasting techniques.

Barometric Method: Researchers use this method, also known as the leading indicators approach, to predict future trends based on current events. When past events are used to forecast future events, they are referred to as leading indicators.

Qualitative Methods:

Surveys: Surveys are used to gather information about the target audience's preferences, opinions, choices, and feedback on their products and services. Most survey software allows you to choose from a variety of question types.

You can also save time and effort by using a pre-made survey template. By changing the theme, logo, and other elements, online surveys can be tailored to fit the brand of the company. They can be distributed via a variety of channels, including email, website, offline app, QR code, social media, and so on. You can choose the channel based on the type and source of your audience. Survey software can generate various reports and run analytics algorithms to uncover hidden insights once the data has been collected. A survey dashboard can show you statistics such as response rate, completion rate, demographic filters, export and sharing options, and so on. Integrating survey builder with third-party apps can help you get the most out of your online data collection efforts.

Polls: One single or multiple choice questions is asked in a poll. You can use polls when you need to get a quick pulse on the audience's feelings. It is easier to get responses from people because they are short in length.

Online polls, like surveys, can be integrated into a variety of platforms. After the respondents have responded to the question, they can see how their responses compare to those of others.

Interviews: The interviewer asks the respondents questions either face-to-face or over the phone in this method. In face-to-face interviews, the interviewer asks the interviewee a series of questions in person and takes notes on the answers. If meeting the person is not possible, the interviewer can conduct a telephonic interview. When there are only a few respondents, this method of data collection is appropriate. If there are many participants, repeating the same process is too time-consuming and tedious.

Delphi Technique: Market experts are given the estimates and assumptions of forecasts made by other industry experts in this method. Based on the information provided by other experts, experts may reconsider and revise their estimates and assumptions. The final demand forecast is based on the consensus of all experts on demand forecasts.

Focus Groups: A focus group is a small group of people (around 8-10 members) who meet to discuss the problem's common areas. Each person expresses his or her viewpoint on the subject at hand. The discussion among the group members is moderated by a moderator. The group comes to an agreement at the end of the discussion.

Questionnaire: A questionnaire is a printed set of open-ended or closed-ended questions. The respondents must respond based on their knowledge and experience

with the topic at hand. The survey includes the questionnaire, but the questionnaire's end-goal may or may not be a survey.

Sources of Secondary data:

The various sources for secondary data collection may be classified into two broad categories:

- Published Sources
- Unpublished Sources

Published Sources:

International, governmental and local agencies are the ones to publish statistical data, among which the following are important: T

- **International Publications:** We have international institutions and bodies like I.M.F, I.B.R.D, I.C.A.F.E and U.N.O who occasionally publish on occasional reports on statistical and economic matters.
- **Official Publications of Central and State Governments:** Reports on different subjects are published by several departments of the Central and State Governments regularly. They collect all the additional information. Important publications among these are: The Reserve Bank of India Bulletin, Census of India, Statistical Abstracts of States, Agricultural Statistics of India, Indian Trade Journal, etc.
- **Semi-Official Publications:** Example: Municipal Corporations, District Boards, Panchayats, etc. that will publish reports relating to different matters of public concern.
- **Publications of Research Institutions:** Indian Statistical Institute (I.S.I), Indian Council of Agricultural Research (I.C.A.R), Indian Agricultural Statistics Research Institute (I.A.S.R.I), etc., publish the findings of their research programs.
- **Publications of various Commercial and Financial Institutions**
- **Reports of various Committees and Commissions appointed by the Government:** Such as the Raj Committee's Report on Agricultural Taxation, Wanchoo Committee's Report on Taxation and Black Money, etc. are also important sources of secondary data.
- **Journals and Newspapers:** Journals and News Papers are the powerful sources from where data is obtained. Current and important materials on statistics and socio-economic problems are provided by journals and newspapers like Economic Times, Commerce, Capital, Indian Finance, Monthly Statistics of trade etc.
- **Unpublished Sources:** There are different examples of these source of data like records maintained by various government and private offices, the theses of the numerous research scholars in the universities or institutions etc.,
- **Precautions to be taken in the use of Secondary Data:** As secondary data is obtained already, it is better to scrutinize it to ensure its accuracy. The investigator needs to be more careful when using this type of data. I Prof. Bowley is right to say, "Secondary data should not be accepted at their face

Notes

value.” This data can be erroneous in different respects due to biases and prejudiced mindset of the information collectors along with the sample size being inadequate, mistakes in definition, mathematical errors and substitution issues. Even without error, such data still can be unsuitable for enquiry purpose. According to Prof. Simon Kuznet’s (which is of importance), “the degree of reliability of secondary source is to be assessed from the source, the compiler and his capacity to produce correct statistics and the users also, for the most part, tend to accept a series particularly one issued by a government agency at its face value without enquiring its reliability”.

Thus we need to follow some of the given factors:

- **The Suitability of Data:** This is possible by judging the scope and nature of the present enquiry with the original one. For example, if we are looking for trend in retail prices while the data provided is meant for wholesale prices, then it is of no use.
- **Adequacy of Data:** Once it is ensured that the data is suitable for investigation, it should be checked for the purpose of present analysis. Geographical area for the original enquiry can be studied in this respect along with the time for which we are getting the data. In the above example, if we want to study the retail price trend of india, and acquired data will cover only the retail price trend in the state of UP, then it would not serve the purpose.
- **Reliability of Data:** This issue concerns whether research findings can be applied to a larger group than those who participated in the study. To put it another way, would similar results have been obtained if a different group of respondents or a different set of data points had been used? Is the information obtained from these 40 people sufficient to conclude how the entire sales force feels about company policies, for example, if 40 salespeople out of a 2,000-person corporate sales force participate in a research study focusing on company policy? Would the results be the same if the study was repeated with 40 different salespeople?

The main goal of reliability is to ensure that the data collection method produces consistent results. This can be measured in some types of research by having different researchers use the same methods to see if the results can be replicated. If the results are similar, the data collection method is most likely reliable. The scientific research method includes ensuring that research can be replicated and produces similar results.

While editing primary data, the following considerations should be borne in mind:

- The data should be complete in all respects
- The data should be accurate
- The data should be consistent
- The data should be homogeneous

For data to possess the above-mentioned characteristics, they have to undergo the following editing types:

2.2.3 Tabulating and Validating the Collected Data

Meaning of Tabulation: Tabulation is a systematic and logical presentation of

numeric data in rows and columns to facilitate comparison and statistical analysis. It facilitates comparison by bringing related information close to each other and helps in further statistical analysis and interpretation.

In other words, the method of placing organized data into a tabular form is called as tabulation. It may be complex, double or simple depending upon the nature of categorization.

Tabulation of Data Collected

The objectives of tabulation of collected data are as follows:

- Orderly arrangement of data in columns and rows
- To bring out essential features of the data collected
- To simplify the data collected
- Conserves space
- Ease of comparison
- Summation of items
- Enables easy detection of errors and omissions
- Facilitates Statistical computations

Validating Data

Definition of Data Validation: As defined by United Nations Economic Commission of Europe (UNECE 2013), data validation is an activity aimed at verifying whether the value of a data item comes from the given (finite or infinite) set of acceptable values.

Data validation means checking the accuracy and quality of source data before using, importing or otherwise processing data. Different types of validation can be performed depending on destination constraints or objectives.

Data validation is a form of data cleansing.

For example, an email question will automatically check if the data entered is a valid email. A phone number question can check whether the phone number has the right number of digits, based on its country code.

Reasons for performing Data Validation:

The following are the reasons for validating data:

- It is cost-effective because the collection of datasets saves the right amount of time and money.
- Because it removes duplicates from the entire dataset, it is simple to use and compatible with existing processes.
- With improved information collection, data validation can directly aid in business improvement.
- It's made up of a data-efficient structure that provides information from a standard database as well as a cleaned dataset.

Notes

Types of Validity

1. **Content validity:** The extent to which the items' content adequately represents the universe of all relevant items under investigation. Is it true that samples are representative of the population/universe?
2. **Criterion Validity:** The extent to which each criterion can be measured correctly. For instance, consider a family's income.
3. **Construct Validity:** The construct validity of a scale or test refers to how well it measures the construct.

For example, a doctor might assess the effectiveness of a painkiller. Each day, he tries to assess the level of pain by asking his patients to rate pain on a 1-10 scale. Whether its pain or numbness, he's measuring it.

Unit-2.3: Questionnaire Designing

Notes

Unit Objectives

At the end of this unit, participants will be able to:

- Analyze the importance of Questionnaire
- List the Steps to be followed for constructing a Questionnaire
- Identify the types of questions to be asked in a Questionnaire
- The Format of a Questionnaire

2.3.1 Introduction to Questionnaire

These days, questionnaire is widely used for data collection in research. It is a reasonably fair tool for gathering data from large, diverse, varied and scattered social groups. The questionnaire is the media of communication between the investigator and the respondents. According to Bogardus, "a questionnaire is a list of questions sent to a number of persons for their answers and which obtains standardized results that can be tabulated and treated statistically".

The Dictionary of Statistical Terms defines it as a, "group of or sequence of questions designed to elicit information upon a subject or sequence of subjects from information".

2.3.2 Steps to be Followed for Constructing a Questionnaire

Step 1: Determine what data is required.

The researcher should begin by reviewing the proposal and brief and making a list of all of the objectives as well as the information needed to achieve them.

Step 2: Make a list of the questions you'll be asked.

A list of all the questions that could be included in the questionnaire is now being compiled. The goal at this point is to be as thorough as possible with the listing and not to be concerned with the wording of the questions. That'll be the next step.

Step 3: Improve the wording of the question

The questions must now be refined to the point where they make sense and generate the correct responses.

Step 4: Create a format for your response.

Every question necessitates an answer. This could be a pre-coded list of responses or an open-ended question to collect verbatim feedback. It's just as important to think about the responses as it is to get the questions right. In fact, thinking about the answers will help you answer the questions correctly.

Step 5: Arrange the questionnaires in the correct order.

The order in which the questions are asked is crucial because it gives the interview logic and flow. Typically, the respondent is eased into the task with relatively simple

Notes

questions, with the more difficult or sensitive ones being saved until the respondent has warmed up. Unprompted questions about brand awareness are asked first, followed by prompted questions.

Step 6: Complete the questionnaire layout.

The questionnaire must now be fully formatted, including a powerful introduction, routings, and probes, as well as clear instructions for the interviewer. There must be enough space to write in answers, and the response codes must be well separated from one another so that the wrong one is not circled.

Step 7: Practice and rehearse

The questionnaire must then be tested. Because the goal of a pilot is to ensure that it works, rather than to obtain pilot results, it is usually not necessary to conduct more than 10 to 20 interviews. The questionnaire should theoretically be piloted with the interviewing method that will be used in the field (over the phone if telephone interviews are to be used; self completed if it will be a self completion questionnaire). Because time and money may prevent a proper pilot, it should be tested on one or two colleagues for logic, flow, and clarity of instructions at the very least. The entire point of the test is to see if any changes are required before final revisions can be made. When conducting the pilot, it is best to go over the questionnaire with the guinea pig respondent and then ask for each question, "What went through your mind when you were asked this question?"

2.3.3 Types of Questions to be asked in a Questionnaire

Open-ended questions

With this question, you can start a conversation. These are good survey questions to get more meaningful responses from because people can provide additional feedback via a text box. You'll need to use a closed-end question if you're looking for a yes/no response.

Open-ended question examples:

- What are you wearing today?
- How did you meet your best friend?
- What is it like to live in Barcelona?

Closed-ended questions

Some questions only require a single word answer. Yes, I agree. Yes or no. You can use them to get some quick tidbits of information, then segment your survey-fillers based on that information.

Closed-ended questions examples:

- Did you order the chicken?
- Do you like learning German?
- Are you living in Australia?

Rating questions

Strive for the moon and stars. Alternatively, the hearts. Alternatively, smiles. Send a rating question to your survey participants to see how they would rate something. It's a great question to ask because it allows you to gauge people's opinions across the board.

Rating questions examples:

- How would you rate our service out of 5?
- How many stars would you give our film?
- Please, rate how valuable our training was today.

Likert scale questions

Likert scale questions are useful in surveys to determine what people think about certain topics. They usually come in five, seven, or nine-point scales, and you've probably used one before.

Likert scale questions examples:

- Do you agree that channel 5 offers more comedy than channel 6?
- How satisfied are you today with our customer service?
- Do you feel affected by the recent changes in the office?

Multiple choice questions

Do you want to send out a test or a quiz? Multiple-choice questions are your best pal. You can give a few answers while keeping the true answer hidden. Multiple-choice questions are also useful for determining time periods or dates for an event. Plus, you can group them all together in a dropdown menu.

Multiple choice questions examples:

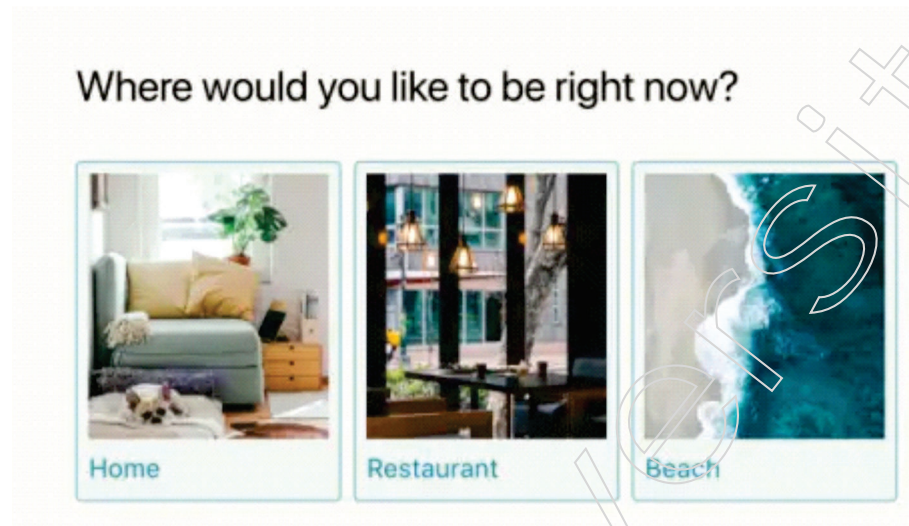
- Facebook was launched in... 2003 | 2004 | 2005 | 2006
- How many of our restaurants have you visited? 1 | 2 | 3 | 4+
- What is the capital of Scotland? Perth | Glasgow | Aberdeen | Edinburgh

Picture choice questions

A thousand words are painted by a single image. But in a poll? It accomplishes a great deal more. Make your survey even more interactive by including a picture choice question. Show rather than tell when telling a story.

Notes

Picture choice questions example



Demographic questions

Questions in demographic surveys are a mix of different types of questions. Whether you use a dropdown or an open-ended question with them is entirely up to you. Take note that they all discuss topics that could be considered sensitive.

Multiple choice questions examples:

How old are you?

What's your gender?

Which industry do you work in?

2.3.4 Format of a Questionnaire

Sample Questionnaire

Section I: Personal Information

1. In what age group are you?

- 19 and under
- 20 - 29
- 30 - 39
- 40 - 49
- 50 - 59
- 60 +

2. Gender:

- Male
- Female

3. In terms of your current occupation, how would you characterize yourself?

- Writer
- Administrative Assistant
- Journalist
- Secretary
- Academic
- Professional
- Technical expert
- Student
- Designer
- Administrator/Manager
- Other, please specify:

Part-2: To be completed during and/or after software use

1. With respect to the version of Microsoft Word currently installed on your machine, please indicate the extent to which you agree or disagree with the following statements:

- SD = Strongly Disagree
- D = Disagree
- N = Neutral
- A = Agree
- SA = Strongly Agree

This software is easy to use.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
I am in control of the contents of the menus and toolbars.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
I will be able to learn how to use all that is offered in this software.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
Navigating through the menus and toolbars is easy to do.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
This software is engaging.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
The contents of the menus and the toolbars match my needs.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
Getting started with this version of the software is easy.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
This software is flexible.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA
Finding the options that I want in the menus and toolbars is easy.	<input type="checkbox"/> SD	<input type="checkbox"/> D	<input type="checkbox"/> N	<input type="checkbox"/> A	<input type="checkbox"/> SA

Notes

Part-3: To be Completed once both versions of Microsoft Word have been used by the subject.

1. If you could choose only one of the versions to continue using, which would it be?
 - Microsoft Word 2000
 - Microsoft Word Personal
2. What particular aspect(s) of Microsoft Word 2000 did you like?
3. What particular aspect(s) of Microsoft Word 2000 did you dislike?
4. What particular aspect(s) of Microsoft Word Personal did you like?
5. What particular aspect(s) of Microsoft Word Personal did you dislike?
6. There are a number of criteria listed below. Please select the version that would be your 1st choice according to each of the criteria. If you really cannot make a choice for a given criteria please select "Equal".

2000 = Microsoft Word 2000

Personal = Microsoft Word Personal

Equal = 2000 and Personal satisfy this criteria equally

Criteria	1st Choice		
This software is easy to use.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
I am in control of the contents of the men-us and toolbars.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
I will be able to learn how to use all that is offered in this software.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
Navigating through the menus and toolbars is easy to do.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
This software is engaging.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
The contents of the menus and the toolbars match my needs.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
Getting started with this version of the software is easy.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
This software is flexible.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
Finding the options that I want in the men-us and toolbars is easy.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
It is easy to make the software do exactly what I want.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
Discovering new features is easy.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
I get my word processing tasks done quickly with this software.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal
This software is satisfying to use.	<input type="checkbox"/> 2000	<input type="checkbox"/> Personal	<input type="checkbox"/> Equal

Unit-2.4: Measurement and Scaling

Notes

Unit Objectives

At the end of this unit the participants will be able to:

- Introduction to Measurement and Scaling Techniques
- Types of Scaling Techniques
- Attitude Measurement Scales and its Types

2.4.1 Introduction to Measurement and Scaling Techniques

Definition of Measurement

Measurement is the process of observing and recording the observations that are collected as part of research. "Process of mapping aspects of a domain onto other aspects of a range according to some rule of correspondence"

By C.R.Kothari—

The process of describing some property of a phenomenon of interest, usually by assigning numbers in a reliable and valid manner, is known as measurement. The numbers provide details about the object being measured. When numbers are used, the researcher must follow a set of rules for assigning a numerical value to an observation in a way that is accurate.

It's a broad concept that encompasses a group of objects, attributes, events, or processes.

Definition of Scaling

Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. In other words, the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned is called as scaling.

Levels of Measurement Scale

The level of measurement refers to the relationship among the values that are assigned to the attributes, feelings or opinions for a variable.

Typically, there are four levels of measurement scales or methods of assigning numbers:

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

Nominal Scale

This is the crudest among all measurement scales but it is also the simplest scale. In this scale the different scores on a measurement simply indicate different categories.

Notes

The nominal scale does not express any values or relationships between variables. The nominal scale is often referred to as a categorical scale.

The assigned numbers have no arithmetic properties and act only as labels. The only statistical operation that can be performed on nominal scales is a frequency count. We cannot determine an average except mode.

Example: Labelling Apples as 1 and Oranges as 2 for data recording does not mean Apples are tastier than Oranges.

Ordinal Scale

A system for assigning numbers and symbols to events in chronological order, but not according to any interval rule.

Places events in order of importance, from highest to lowest, for example. Exam results are ranked by students. The first-place finisher is not three times better than the third-place finisher. He only outperforms the second and third-placed students.

Interval Scale

This is a scale in which the numbers are used to rank attributes such that numerically equal distances on the scale represent equal distance in the characteristic being measured. An interval scale contains all the information of an ordinal scale, but it also allows to compare the difference/distance between attributes. Interval scales may be either in numeric or semantic formats.

The interval scales allow the calculation of averages like:

- Mean
- Median
- Mode
- Dispersion like Range and Standard Deviation.

Example: Measuring temperature is an example of interval scale. But we cannot say 40°C is twice as hot as 20°C.

A couple of examples of Interval Scale in Numeric format and Semantic format have been given below:

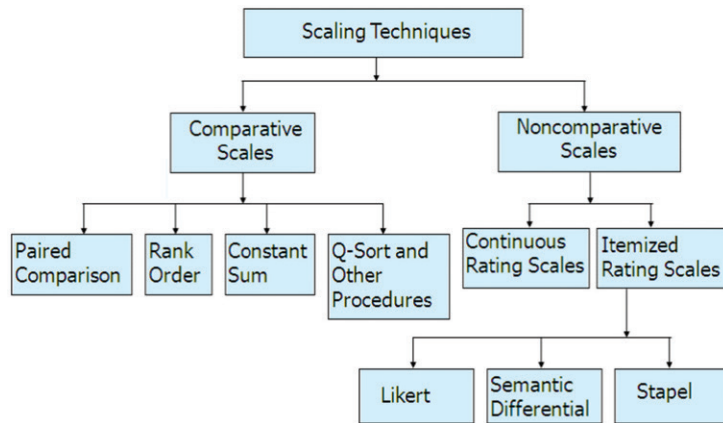
Ratio Scale

This scale is the highest level of measurement scales. This has the properties of an interval scale together with a fixed (absolute) zero point. The absolute zero point allows us to construct a meaningful ratio.

Ratio scales permit the researcher to compare both differences in scores and relative magnitude of scores. Examples of ratio scales include weights, lengths and times.

Example: The number of customers of a bank's ATM in the last three months is a ratio scale. This is because you can compare this with previous three months.

2.4.2 Types of Scaling Techniques



Notes

Comparative Scales

In comparative scales, the respondent is asked to compare one object with another.

Comparative scales can be further classified into the following types of scaling techniques:

- Paired Comparison Scale
- Rank Order Scale
- Constant Sum Scale
- Q - sort Scale

Paired Comparison Scale

The respondent must choose a preferred object from several pairs of objects based on some property, which results in object rank ordering. Respondents found it time-consuming and exhausting.

E.g. Give your preference for bikes-

- | | |
|----------|--------------|
| Pulsar | CBZ |
| Shine | Passion Plus |
| Activa | Vigo |
| Pleasure | Pep Plus |

For 8 bikes we will have 56 pairs for comparison

Rank Order Scale

A scale where the participant orders several objects or properties of objects.

E.g. Rank the bikes in your order of preference. Place the number 1 to the most preferred, 2 by the second choice, and so forth.

- ___ Pulsar
- ___ CBZ

Notes

- ___ Passion plus
- ___ Shine
- ___ Activa
- ___ Vigo
- ___ Pleasure

Constant Sum Scale

Respondent allocates points to more than one attribute or property, such that they total a constant sum, usually 10 or 100.

E.g. Success of 'Golmal'

Story: ___

Music: ___

Songs: ___

Casting: ___

Total: 100

Q - Sort Scale

The systematic study of participant viewpoints is known as Q-methodology (also known as Q-sort). By having participants rank and sort a series of statements, the Q-methodology is used to investigate the perspectives of participants who represent different stances on an issue.

2.4.3: Attitude Measurement Scales

1. Non-Comparative Scales

A non-comparative scale is used to evaluate a product's or object's performance across a variety of parameters. Some of the most common types are as follows:

Continuous Rating Scales (CRS) are a type of rating scale that is used

It's a graphical rating scale in which respondents can place the object in any position they want. It's done by picking a point on a vertical or horizontal line that falls between two extreme criteria and marking it.

2. Scale of Itemized Ratings

It focuses on the respondents' selection of a specific category from among the many options presented to them. The following are the three most commonly used itemised rating scales:

Likert Scale: In a Likert scale, the researcher presents some statements to the respondents and asks them to indicate their level of agreement or disagreement with these statements by selecting one of the five options from a list of five.

- **Semantic Differential Scale:** A bi-polar seven-point non-comparative rating scale in which the respondent can mark on any of the seven points for each of the object's attributes based on personal preference.
- **Stapel Scale:** A Stapel scale is an itemised rating scale that uses a unipolar rating to measure the respondents' response, perception, or attitude toward a specific object. A Stapel scale has a range of -5 to +5, so it excludes 0 from the equation.

Continuous Rating Scale

The respondents use a comparative scale to compare two or more variables. The various types of comparative scaling techniques are as follows:

1. Paired Comparison

A paired comparison denotes a situation in which the respondent must choose one of two variables.

When comparing more than two objects, such as P, Q, and R, compare P with Q first, then the superior one (i.e., the one with a higher percentage) with R.

2. Rank Order

The respondent must rank or arrange the given objects according to his or her preference in rank order scaling.

3. Sum Constant

It's a method of scalability in which the features, attributes, and values are assigned a constant sum of units such as dollars, points, chits, chips, and so on. The respondents place a high value on a specific product or service.

4. Scaling by Q-Sort

Q-sort scaling is a method for selecting the most appropriate objects from a large set of variables.

Notes

Unit-2.5: Sampling

Unit Objectives:

At the end of this unit, participants will be able to learn:

- Describe sampling
- Analyze sampling plan and sampling frame
- List steps involved in sampling process
- Identify different sample selection methods
- Describe probability and non-probability sampling techniques
- Identify sampling and non-sampling errors

2.5.1 Introduction to Sampling

Sample-

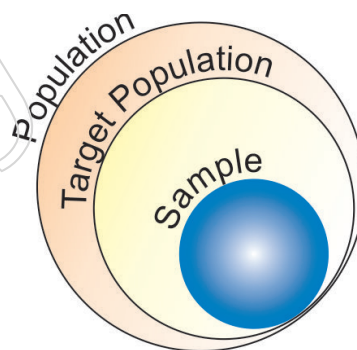
It is a subset of population

"Segment of population that is selected for investigation."

Bryman & Bell-

"Some of the elements of population"

Cooper & Schindler-



A sample, in research terms, is a group of people, objects, or items selected for measurement from a larger population. To ensure that the findings from the research sample can be applied to the entire population, the sample should be representative of the population.

What is the purpose of sampling?

Inferential statistics enable us to determine a population's characteristics by directly observing only a portion (or sample) of the population, allowing us to draw conclusions about populations from samples.

We obtain a sample of the population for a variety of reasons, including the fact that it is rarely practical and almost never cost-effective.

Many populations are quite large

- **Inaccessibility of some populations:** Access to some populations is so difficult that only a sample can be used. Prisoners, people with severe mental illness, and disaster survivors are just a few examples. And so on. The inaccessibility could be due to a lack of funds, time, or simply access.
- **Destructiveness of observation:** Sometimes just observing a product's desired characteristic destroys it for its intended use. Quality control is a good example of this. For example, a fuse must be destroyed to determine its quality and whether it is defective.
- As a result, if you tested all of the fuses, they'd all blow up.
- **Accuracy and sampling:** A sample of the study population may be more accurate than the entire population. A population that has been incorrectly identified can provide less reliable data than a sample that has been carefully selected.

2.5.2 Sampling Plan and Sampling Frame

Definition of Sampling Plan

A sampling plan is a term widely used in research studies that provide an outline on the basis of which research is conducted. It tells which category is to be surveyed, what should be the sample size and how the respondents should be chosen out of the population. Sampling plan is the base from which the research starts and includes the following major decisions:

i. Choose the population

Choosing the category of the population to be surveyed is the first and the foremost decision in a sampling plan that initiates the research.

ii. Determine the Sample Size

The second decision in sampling plan is determining the size of the sample i.e., how many objects in the sample is to be surveyed. Generally, "the larger the sample size, the more is the reliability" and therefore, researchers try to cover as many samples as possible.

iii. Decide the Sampling Procedure

The final decision that completes the sampling plan is selecting the sampling procedure i.e., which method can be used such that every object in the population has an equal chance of being selected. Generally, the researchers use the probability sampling to determine the objects to be chosen as these represents the sample more accurately.

Definition of Sampling Frame

A sampling frame is a list or database from which a sample can be used. In market research terms, a sampling frame is a database of potential respondents that can be drawn from, to invite to take part in a given research project.

Notes

2.5.3 Steps involved in Sampling Process

The following are the series of steps that are involved in the sampling process:

- Define the population
- Determine the sampling frame
- Select the sampling techniques / method
- Determine the sample size
- Execute the sampling process

2.5.3.1 Sample Selection Methods

The sample selection methods can be broadly classified into:

- Probability Sampling
- Non-Probability Sampling

2.5.3.2 Probability Sampling

The various Probability sampling methods are given below:

- Simple random sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling
- Multistage Sampling

2.5.4 Simple Random Sampling

A probability sample in which each element of population has a known & equal chance of selection

E.g. Population = Total students in AMITY (591)

Sample size = 25 students

$$\text{Probability of selection} = \frac{\text{Sample size}}{\text{Population size}}$$

$$= \frac{25}{591}$$

$$= 0.042 \text{ or } 4.2\%$$



Systematic Sampling

A probability sample drawn by applying a calculated skip interval to a sample frame.

Formula to calculate skip interval-

$$\text{Skip interval (k)} = \frac{\text{Population sample frame (N)}}{\text{Sample Size (n)}}$$

E.g. Population = Total students in AMITY (591)

Sample size = 25 students

$$\text{Skip interval (k)} = \frac{\text{Population sample frame (N)}}{\text{Sample Size (n)}}$$

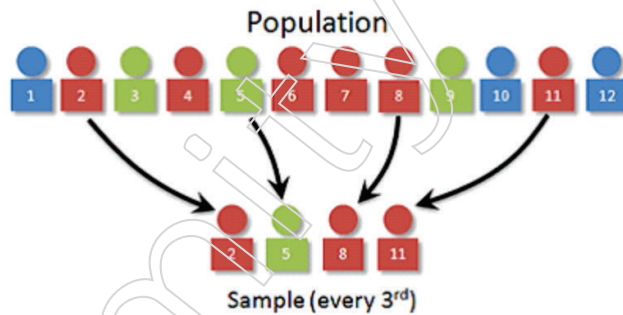
$$= 591/25$$

$$= 23.64 = 23$$

Select any number randomly between 1 – 23, and then select rest 24 numbers with a gap of 23 numbers.

$$k = 23$$

Select any number randomly between 1 – 23, and then select rest 24 numbers with a gap of 23 numbers.



Stratified Sampling

A probability sampling technique in which the population is divided into different sub-homogeneous groups or strata and samples are randomly selected from such sub-groups or strata.

E.g. Population = Total students in JSPM (591)

Sample size = 25 students

Sub groups-

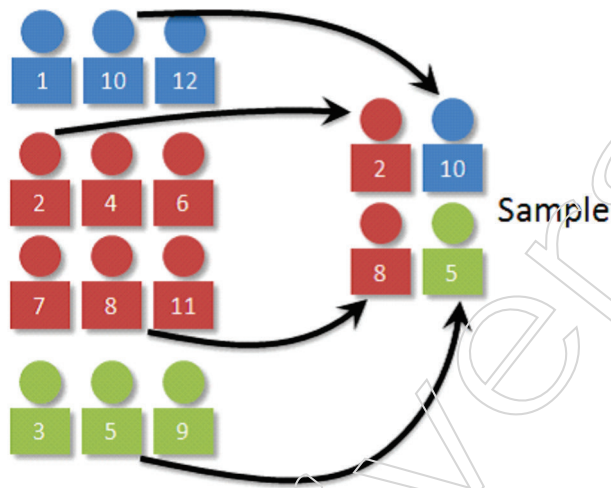
MBA = 272	MMM = 42
MPM = 59	DBM = 92

Notes

DIEM = 37

MCA = 59

MCM = 30



Cluster Sampling

A probability sampling technique in which the population is divided into several small sub groups and some groups selected randomly for study.

e.g. MBA = 272

MMM = 42

MPM = 59

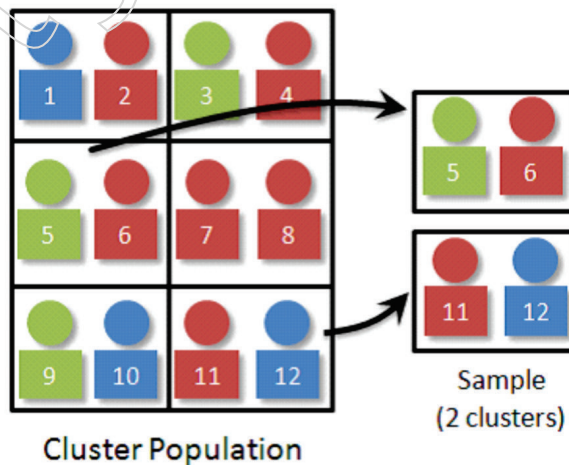
DBM = 92

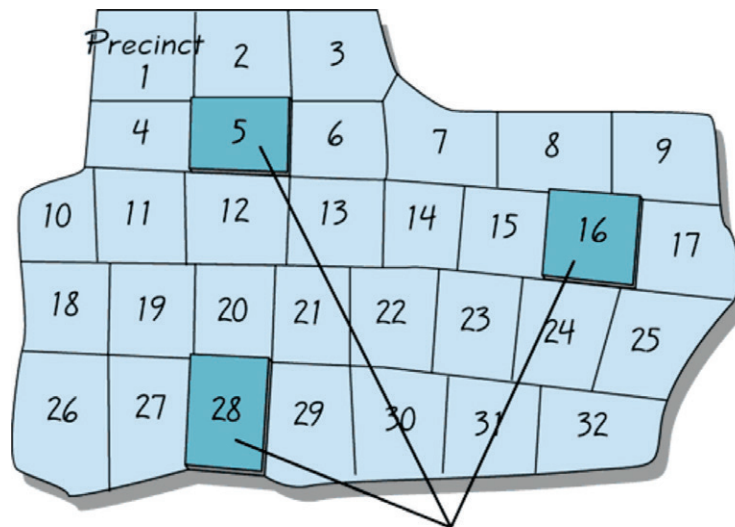
DIEM = 37

MCA = 59

MCM = 30

MCM = 30 and MMM = 42 randomly selected from above groups.





Interview all voters in shaded precincts.

Notes

Non - Probability Sampling

Also known as 'deliberate,' 'purposeful,' or 'judgement' sampling. The items have been chosen with care.

The researcher's choice of elements to include in the sample. There is no guarantee that each element will be given an equal chance to be chosen as a sample.

Convenience Sampling

A non-probability sampling technique where researcher use any readily available individuals as participants

Most cheapest & easiest to conduct

Least reliable

Researcher is free to select anybody as sample

Purposive or Judgmental Sampling

A non-probability sample that conforms to certain criteria. The units or elements are purposively selected.

Purposive sampling is a sampling method in which elements are chosen based on purpose of the study. Purposive sampling may involve studying the entire population of some limited group or a subset of a population). As with other non-probability sampling methods, purposive sampling does not produce a sample that is representative of a larger population, but it can be exactly what is needed in some cases - study of organization, community, or some other clearly defined and relatively limited group

Quota Sampling

A purposive sampling where relevant characteristics are used to stratify the samples.

Notes

This is useful to increase the representativeness of samples

MBA = 272	46%	12
MMM = 42	7%	2
MPM = 59	10%	2
DBM = 92	16%	4
DIEM = 37	6%	2
MCA = 59	10%	2
MCM = 30	5%	1
Total = 591	100%	25

Snowball Sampling

A non- probability sampling in which subsequent participants are referred by current sample elements.

Initial individuals are discovered/identified

These initial individuals refer others who are similar to them and so on. Like a snowball gathers subjects as its rolls along.

2.5.6 Sampling and Non-Sampling Errors

Definition of Sampling Error

Sampling error refers to differences between the sample and the population that, exist only because of the observations that happened to be selected for the sample.

Increasing the sample size will reduce this type of error.

Definition of Non-Sampling Error

Non-sampling errors are more serious, and they occur as a result of data acquisition errors or incorrect sample observation selection.

Non-Sample Errors

Non-sample errors can be classified into:

- Non-response Error
- Response Error

Non-Response Error

A non-response error occurs when units selected as part of the sampling procedure do not respond in whole or in part.

Response Error

A response or data error is any systematic bias that occurs during data collection, analysis or interpretation.

Response errors occur due to the following factors:

- Respondent Error (example: lying, forgetting, etc.,)
- Interviewer Errors
- Recording Errors
- Poorly designed questionnaires.
- Measurement errors.

Summary

At the end of this module the participant have covered:

- Analyzing Research Modelling
- Defining Data Collection and its Methods
- Explaining Questionnaire Designing
- Describing Measurement and Scaling
- Analyzing Sampling

Exercise:

1. Sampling is divided into two types, viz. and.....
 - a) Data collection and Analyzing
 - b) Analyzing and Interpreting
 - c) Both a and b
 - d) None of the above
2. A _____, in research terms, is a group of people, objects, or items selected for measurement from a larger population
 - a) Sample
 - b) Item
 - c) Object
 - d) None of the above
3. A _____ is a term widely used in research studies that provide an outline on the basis of which research is conducted
 - a) sampling plan
 - b) sorting plan
 - c) separating plan
 - d) none of the above
4. _____ is also known as 'deliberate,' 'purposeful,' or 'judgement' sampling
 - a) Non - Probability Sampling
 - b) Probability Sampling
 - c) Snowball Sampling
 - d) Quota sampling

Notes

5. Purposive sampling is a sampling method in which elements are chosen based on _____

- a) purpose of the study
- b) subject of the study
- c) nature of the study
- d) none of the above

Answers:

1. a) Data collection and Analyzing
2. a) Sample
3. a) sampling plan
4. a) Non - Probability Sampling
5. a) purpose of the study

Module-3: Data Analysis Techniques

Notes

Key learning Outcomes

At the end of this module, participants will be able to:

1. Identify and analyse descriptive statistics
2. Analyse and utilize the hypothesis testing
3. Identify the importance of parametric and non-parametric tests
4. Analyse and perform principle component factor analysis
5. Identify the importance of data analysis

Structure

Unit 3.1: Descriptive Statistics

- 3.1.1 Introduction to Descriptive Research Design
- 3.1.2 Applications of Descriptive Research
- 3.1.3 Descriptive Research Methods

Unit 3.2: Hypothesis Testing

- 3.2.1 Introduction to Hypothesis
- 3.2.2 Types of Hypothesis
- 3.2.3 Testing a Hypothesis and its Significance Levels
- 3.2.4 Type I and Type II Errors
- 3.2.5 One - tailed and Two - tailed Tests
- 3.2.6 Confidence Interval
- 3.2.7 Bayesian Statistics

Unit 3.3: Parametric and Non-Parametric Tests

- 3.3.1 Introduction to Parametric and Non-Parametric Tests
- 3.3.2 Z-test
- 3.3.3 t-test
- 3.3.4 Correlation & Regression
- 3.3.5 Chi-square test
- 3.3.6 Factor Analysis

Unit 3.4: Data Analysis

- 3.4.1 Principle Component Factor Analysis
- 3.4.2 Introduction to SPSS
- 3.4.3 Data Creation in SPSS
- 3.4.4 Run example on parametric test
- 3.4.5 Run example on non-parametric test

Notes

Unit 3.1: Descriptive Statistics

Unit Objectives

At the end of this unit, participants will be able to:

- An Introduction to Descriptive Research, its Definition and Characteristics
- Application of Descriptive Research with examples
- Descriptive Research Methods

3.1.1 Introduction to Descriptive Research Design

Introduction to Descriptive Research

Descriptive research used to describe a population, situation or phenomenon in accurately and systematically. It can answer questions such as what, where, when and how, but it cannot why questions.

A descriptive research design should use a large range of research methods to investigate one or more variables. Not like in experimental research, the scholar doesn't control or manipulate any of the variables, but only observes and calculate them.

Definition

A Descriptive Research Design is concerned with describing the characteristics of a particular individual or a group.

3.1.2 Applications of Descriptive Research

To understand the top objective of research goals, organizations currently use descriptive research within the following ways:

- **Define Respondent Characteristics:** The aim of using close-ended questions is to draw concrete conclusions about the respondents. This might be the necessity of derive patterns, traits, and behaviours of the respondents. It could even be to know from a respondent, their attitude, or opinion about the phenomenon. For instance, understanding from millennial the hours per week they spend on browsing the web. All this information helps the organization researching to make informed business decisions.
- **Measure Data Trends:** Researchers measure data trends over time with a descriptive research design's statistical capability. Consider if an apparel company researches different demographics like age groups from 24-35 and 36-45 on a replacement range launch of summer wear. If one in every of those groups doesn't take too well to the new launch, it provides insight into what clothes are like and what is not. The brand drops the garments and apparel that customers don't like.
- **Conduct Comparison:** Organizations also use a descriptive research design to know how different groups answer to a particular product or service. For instance, an apparel brand creates a survey asking general questions that measure the brand's image. The identical study also asks demographic

questions like age, income, gender, geographical location, etc. This marketing research helps the organization understand what aspects of the brand appeal to the population and what aspects don't. It also helps make product or marketing fixes or even create a new product line to cater to high growth potential groups.

- **Validate Existing Conditions:** Researchers widely use descriptive research to assist ascertain the research object's prevailing conditions and underlying patterns. Because of the non-invasive research method and therefore the use of quantitative observation and some aspects of qualitative observation, researchers observe each variable and conduct an in-depth analysis. Researchers also use it to validate any existing conditions that may be prevalent in an exceedingly population.
- **Conduct Research at different times:** The analysis will be conducted at different periods to establish any similarities or differences. This also allows any number of variables to be evaluated. For verification, studies on prevailing conditions may be repeated to draw trends.

3.1.3 Descriptive Research Methods

Observational Method

Animal and human behaviour are closely observed using the observational method (also known as field observation). Naturalistic observation and laboratory observation are the two main types of observational methods.

The most significant benefit of using a naturalistic approach to research is that researchers can observe participants in their natural settings. Proponents claim that this provides more ecological validity than laboratory observation.

The extent to which research can be applied in real-life situations is referred to as ecological validity.

Laboratory observation proponents frequently argue that the results obtained with laboratory observation are more meaningful than those obtained with naturalistic observation because the laboratory has more control.

Naturalistic observations are usually more time consuming and expensive than laboratory observations. Both naturalistic and laboratory observation are important in the advancement of scientific knowledge, of course.

Case Study Method

Case study research entails an in-depth examination of a single person or a group of people. Case studies frequently result in testable hypotheses and allow us to investigate unusual phenomena. Case studies aren't good for determining cause and effect, and they're useless for making accurate predictions.

Expectancy effects and atypical individuals are two serious issues with case studies. Expectancy effects are underlying biases held by the experimenter that may influence the actions taken during research. These biases can cause participants' descriptions to be misrepresented. Defining atypical people can lead to faulty generalisations and a loss of external validity.

Notes

Survey Method

Participants in survey method research respond to questions via interviews or questionnaires. Researchers describe the responses given by participants after they have answered the questions. The questions must be properly constructed in order for the survey to be both reliable and valid. Questions should be written in a clear and understandable manner.

Unit-3.2: Hypothesis Testing

Notes

Unit Outcomes

At the end of this unit, participants will be able to:

- Define Hypothesis
- Describe different types of Hypothesis
- Explain Testing Hypothesis and its Significance Levels
- Identify Type - I and Type - II Errors with Examples
- Demonstrate One-tailed and Two-tailed Test
- Analyze Confidence Intervals
- Analyze Bayesian Statistics

3.2.1 Introduction to Hypothesis

Definition

“Hypothesis may be defined as a proposition or a set of propositions set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some investigation in the light of established facts” (Kothari, 1988).

Characteristics of Hypothesis

A hypothesis ought to have the subsequent characteristic features:

- A hypothesis must be precise and clear. If it's not precise and clear, then the inferences drawn on its basis wouldn't be reliable.
- A hypothesis should be capable of being placed to check. Very often, the analysis programmes fail owing to its incapability of being subject to testing for validity. Therefore, some prior study may be conducted by the research worker in order to make a hypothesis testable. A hypothesis “is tested if different deductions are made to from it, which in turn can be confirmed or disproved by observation”.
- A hypothesis must state relationship between two variables, in the case of relational hypotheses.
- To be considered reliable, the hypothesis must be clear and precise.
- If the hypothesis is a relational hypothesis, the relationship between variables should be stated.
- The hypothesis should be specific and leave room for further testing.

3.2.2 Types of Hypothesis

There are several types of hypothesis likes:

- Simple Hypothesis

Notes

- Complex Hypothesis
- Empirical Hypothesis
- Null Hypothesis
- Alternative Hypothesis
- Logical Hypothesis
- Statistical Hypothesis

Simple Hypothesis: In simple hypothesis there exists relationship between two variables one is called independent variable or cause and other is dependent variable or effect. For example

- Smoking leads to Cancer
- High rate of unemployment leads to crimes.

Complex Hypothesis: In complex hypothesis there exist a relationship among more variables (more than two dependent and independent). For example

- Smoking and other drugs lead to cancer, chest infections etc.
- The high rate of unemployment, poverty, illiteracy leads to crimes like robbery, rape, prostitution & killing etc.

Empirical / Working Hypothesis: When a theory is put to the test through observation and experiment, it becomes an empirical hypothesis, or working hypothesis. It's no longer just a thought or a hypothesis. It's a matter of trial and error, and possibly rearranging those independent variables.

Null Hypothesis: The null hypothesis, H_0 , denotes a theory that has been proposed but not proven, either because it is believed to be true or because it is intended to be used as a basis for argument. The null hypothesis in a clinical trial of a new drug, for example, might be that the new drug is no better than the current drug on average. We'd write H_0 : on average, there's no difference between the two drugs.

The null hypothesis is given special attention. This is because the null hypothesis is concerned with the statement being tested, whereas the alternative hypothesis is concerned with the statement that will be accepted if/when the null hypothesis is rejected.

After the test has been completed, the final conclusion is always expressed in terms of the null hypothesis. "Reject H_0 in favour of H_1 " or "Do not reject H_0 " are the only options; "Reject H_1 " or "Accept H_1 " are never options.

"Do not reject H_0 " does not imply that the null hypothesis is correct; rather, it implies that there is insufficient evidence to reject H_0 in favour of H_1 . When the null hypothesis is rejected, it implies that the alternative hypothesis is true.

Alternative Hypothesis: The alternative hypothesis, H_1 , is a statement of the purpose of a statistical hypothesis test. In a clinical trial of a new drug, for example, the alternative hypothesis could be that the new drug has a different effect than the current drug on average. We'd write H_1 : on average, the two drugs have different effects.

An alternative hypothesis is that the new drug is, on average, superior to the current drug. In this case, we'd write.

H1: On average, the new drug is better than the current drug.

After the test has been completed, the final conclusion is always expressed in terms of the null hypothesis. "Reject H₀ in favour of H₁" or "Do not reject H₀" are the two options. We never say "Reject H₁" or even "Accept H₁" as a conclusion.

"Do not reject H₀" does not imply that the null hypothesis is correct; rather, it implies that there is insufficient evidence to reject H₀ in favour of H₁. When the null hypothesis is rejected, it implies that the alternative hypothesis is true.

Logical Hypothesis: A logical hypothesis is a proposed explanation for which there is only a small amount of evidence. In general, you want to turn a logical hypothesis into an empirical hypothesis by testing your theories or postulations.

Statistical Hypothesis: A statistical hypothesis is a claim about the parameters or form of a probability distribution for a specific population or populations, or, more broadly, about a probabilistic mechanism that is supposed to generate the observations.

3.2.3 Testing a Hypothesis and its Significance Levels

Testing a Hypothesis

Testing a hypothesis refers to verifying whether, the hypothesis is valid or not. Hypothesis testing attempts to check whether, to accept or not to accept the null hypothesis. The procedure of hypothesis testing includes all the steps that a researcher undertakes for making a choice between the two alternative actions of rejecting or accepting a null hypothesis. The various steps involved in hypothesis testing are as follows:

Step 1: Specify the Null Hypothesis

The null hypothesis (H₀) states that two or more groups or factors have no effect, relationship, or difference. A researcher's primary goal in a research study is to disprove the null hypothesis.

For example, there is no difference in intubation rates between children aged 0 and 5.

The survival rates of the intervention and control groups are identical (or, the intervention does not improve survival rate).

There is no link between the type of injury and whether the patient was given an IV in the prehospital setting.

Step 2: Specify the Alternative Hypothesis

The alternative hypothesis (H₁) asserts that a difference or effect exists. This is typically the hypothesis that the researcher is attempting to prove. The alternative hypothesis can be one-sided (only provides one direction, for example, lower) or two-sided (provides both directions). Even when our true hypothesis is one-sided, we frequently use two-sided tests because accepting the alternative hypothesis requires more evidence against the null hypothesis.

The success rate of intubation varies depending on the age of the patient being treated (two-sided).

Notes

Notes

The intervention group's time to resuscitation from cardiac arrest is shorter than the control group's (one-sided).

Step 3: Set the Significance Level (α)

The significance level is usually set at 0.05 (represented by the Greek letter alpha— α). This means that if your null hypothesis is true, there is a 5% chance that you will accept your alternative hypothesis. The greater the burden of proof required to reject the null hypothesis, or to support the alternative hypothesis, the smaller the significance level.

Step 4: Determine the Test Statistic and the P-Value that corresponds.

We present some basic test statistics to evaluate a hypothesis in another section. In most cases, hypothesis testing employs a test statistic that compares groups or investigates relationships between variables. A confidence interval is commonly used to describe a single sample without establishing relationships between variables.

Step 5: Drawing a Conclusion

P-value \leq significance level (α) \Rightarrow Reject your null hypothesis in favour of your alternative hypothesis. Your result is statistically significant.

P-value $>$ significance level (α) \Rightarrow Fail to reject your null hypothesis. Your result is not statistically significant.

If your null hypothesis is true, the p-value describes the likelihood of obtaining a sample statistic as or more extreme by chance alone. The result of your test statistic is used to calculate this p-value. Your p-value and significance level are used to draw conclusions about the hypothesis.

Significance Level

The probability of rejecting the null hypothesis when it is true is known as the significance level, also known as alpha or α . A significance level of 0.05, for example, indicates a 5% chance of concluding that a difference exists when there is none.

Because of their technical nature, these definitions can be difficult to comprehend. The concepts are much easier to understand with a picture!

The significance level determines how far the line on the graph will be drawn from the null hypothesis value. We need to shade the 5% of the distribution that is furthest away from the null hypothesis to graph a significance level of 0.05.

3.2.4 Type I and Type II Errors

An investigator's conclusion, like a judge's, could be incorrect. A sample may not be representative of the population simply by chance. As a result, the sample results do not reflect reality in the population, and random error leads to an incorrect conclusion. A type I error (false-positive) occurs when an investigator rejects a null hypothesis that is true in the population, while a type II error (false-negative) occurs when an investigator fails to reject a null hypothesis that is true in the population.

Although type I and type II errors are impossible to completely avoid, the investigator can reduce the likelihood of them by increasing the sample size (the

larger the sample, the lesser is the likelihood that it will differ substantially from the population).

Bias can also lead to false-positive and false-negative results (observer, instrument, recall, etc.). (Bias errors, on the other hand, are not classified as type I or type II errors.) Such errors are inconvenient because they are often difficult to detect and cannot be quantified.

3.2.5 One - tailed and Two - tailed Tests:

One - tailed Test

A test of H_0 which assumes, that the difference between sample parameter and population statistics is in only one direction.

e.g. Maximum 15kgs of chemical waste is produced per batch of 60kgs. However a random sample of 100 batches gives an average of 16kgs of chemical waste per patch. Test at 10% level of significance, whether average quantity of waste per batch has increased?

When the hypothesis about the population parameter is rejected for the value of sample statistic falling into outside tail of the distribution, then it is known as one-tailed test.

Two - tailed Test

A test of H_0 which assumes, that the difference between sample parameter and population statistics is in both directions. i.e. sample parameter is either greater or less than population statistics.

e.g. Average height of 20 students = 168 cms. Can this be considered as a sample from large population of average height of 169cms?

3.2.6 Confidence Interval

A confidence interval is a set of numbers that contains an unknown population parameter. If you draw a random sample many times, the population mean will appear in a certain percentage of the confidence intervals. The confidence level is expressed as a percentage.

Confidence intervals are most commonly used to bound the mean or standard deviation, but they can also be used to bound regression coefficients, proportions, rates of occurrence (Poisson), and population differences.

There is a common misunderstanding of how to interpret confidence intervals, just as there is a misunderstanding of how to interpret P values. In this case, the confidence level isn't the most important factor.

If you can assess many intervals and know the value of the population parameter, the confidence level represents the theoretical ability of the analysis to produce accurate intervals. There's no room for probabilities other than 0 or 1 in a specific confidence interval from one study—it either contains the population value or it doesn't. You can't choose between these two options because you don't know the population parameter's value.

Notes

“Because the parameter is an unknown constant, no probability statement can be made about its value.”

—Jerzy Neyman, the inventor of the confidence interval.

This will be clearer after we talk about the grain.

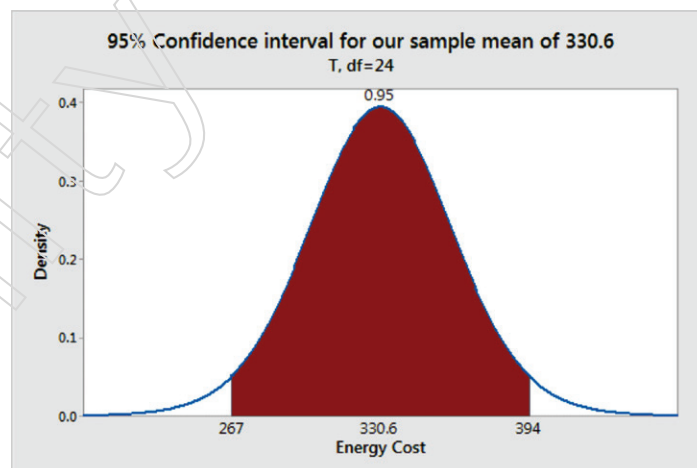
Because the procedure tends to produce intervals that contain the parameter, confidence intervals serve as good estimates of the population parameter. The point estimate (the most likely value) and a margin of error around that point estimate make up confidence intervals. The margin of error describes the level of uncertainty surrounding a sample estimate of a population parameter.

In this vein, confidence intervals can be used to evaluate the precision of a sample estimate. A narrower confidence interval [90 110] for a specific variable indicates a more precise estimate of the population parameter than a wider confidence interval [50 150].

Margin of Error

Let's look at how confidence intervals are used to account for that margin of error. We'll use the same tools we used to understand hypothesis tests to accomplish this. Using probability distribution plots, the t-distribution, and the variability in our data, I'll create a sampling distribution. Our confidence interval will be based on the energy cost data set we've been using.

When we looked at the significance levels, we saw a sampling distribution centred on the null hypothesis value, with the outer 5% of the distribution shaded. We need to shift the sampling distribution so that it is centred on the sample mean and shade the middle 95% for confidence intervals.



The shaded area depicts the range of sample means that you'd get 95% of the time if you used our sample mean as the population mean point estimate. Our 95 percent confidence interval is this range [267 394].

It's easier to understand how a confidence interval represents the margin of error, or the amount of uncertainty, around a point estimate when you look at the graph. Given the information available, the sample mean is the most likely value for the population mean. However, the graph shows that other random samples drawn from the same population could have different sample means within the shaded area, which is not unusual. These other possible sample means all point to a different conclusion.

These graphs can be used to calculate probabilities for specific values. However, because the population mean is unknown, you won't be able to plot it on the graph. As a result, as Neyman pointed out, you can't calculate probabilities for the population mean!

3.2.7 Bayesian Statistics

Bayesian statistics are a mathematical procedure that applies probabilities to statistical problems. In Bayesian statistics, the interpretation of probability is a description of how certain some statement, or proposition, is true.

- If the probability is 1, then we are sure that the statement is true
- If the probability is 0, then we are sure that the proposition is false.
- If the probability is 0.5, then we are as uncertain state, as we would be about a fair coin toss.
- If the probability is 0.95, then we're quite sure the statement is true, but it wouldn't be too surprising to us if we found out the statement was false.

0.....probability.....1

The above figure can say that the probability can be used to describe degrees of certainty, or how plausible some statement is. 0 and 1 are the two extremes of the scale and correspond to complete certainty. However, probabilities are not static quantities. When we get more information, our probabilities can change.

It might sound like there is nothing more to Bayesian statistics than just thinking about a question and then blurting out a probability that feels appropriate. For example, we may be on "Who Wants to be a Millionaire?" and not know the answer to a question, so we might think the probability that it is A is 25%. But if we call our friend using "phone a friend", and our friend says, "It's definitely A", then we would be much more confident that it is A! our probability probably wouldn't go all the way to 100%.

We will now look at a simple example to demonstrate the basics of how Bayesian statistics works.

- We start with some probabilities at the beginning of the problem are called prior probabilities.
- And how exactly these get updated when we get more information, these updated probabilities are called posterior probabilities.
- To make all these clearer, we will use a table that we will call a Bayes' Box to help us calculate the posterior probabilities easily.

Assume there are two balls in a bag, where:

- at least one of them is black,
- but we're not sure whether they're both black,

Notes

- or whether one is black and one is white.

For the above we consider two possibilities either both are black in colour or one is white and another is black so that we can label our two competing hypotheses BB and BW. So, at the beginning of the problem, we know that one and only one of the following statements/hypotheses is true:

Suppose an experiment is performed to help us determine which of these two hypotheses is true. The experimenter reaches into the bag, pulls out one of the balls, and observes its colour. The result of this experiment is

The Baye's Box

A Bayesian analysis starts by choosing some values for the prior probabilities with our two competing hypotheses BB and BW, and we need to choose some probability values to describe how sure we're that each of these is true. Since we are taking two hypotheses then there will be two prior probabilities, one for BB and one for BW. For simplicity, we will assume that we don't have much of an idea which is true, and so we will use the following prior probabilities:

$$P(BB) = 0.5$$

$$P(BW) = 0.5.$$

The above two hypotheses are mutually exclusive (they can't both be true) and exhaustive (one of these is true; it can't be some undefined third option). The choice of 0.5 for the two prior probabilities describes the fact that, before we did the experiment, we were very uncertain about which of the two hypotheses was true. Now present a Bayes' Box, which lists all the hypotheses that might be true, and the prior probabilities. There are some extra columns which we haven't discussed yet, and will be needed in order to figure out the posterior probabilities in the final column. The first column of a Bayes' Box is that the list of hypotheses we're considering. In this case there are just two. If need to construct a Bayes' box for a new problem then just think about what the possible answers to the problem are, and list them in the first column. The 2nd column lists the prior probabilities for every hypothesis. Above, before we did the experiment, we decided to say that there was a 50% probability that BB is true and a 50% probability that BW is true, hence the 0.5 values in this column. The prior column should always sum to 1. Remember, the prior probabilities only describe our initial uncertainty, before taking the data into account.

Hypotheses	prior	likelihood	prior × likelihood	posterior
BB	0.5			
BW	0.5			
Totals:	1			

Likelihood

The third column is called likelihood by which we can calculating the posterior probabilities. It is synonymous with probability. In statistics, likelihood is a very unique style of probability. To fill in the third column of the Bayes' Box, we need to calculate two

likelihoods, so you can tell from this that the likelihood is something different for each hypothesis. But what is it exactly?

Here is the Bayes' Box with the likelihood column filled in

Hypotheses	prior	likelihood	prior × likelihood	posterior
BB	0.5			
BW	0.5			
Totals:	1			

First calculate the value of the likelihood for the BB hypothesis. Remember, the data we are analysing here is that we chose one of the balls in the bag "at random", and it was black. The likelihood for the BB hypothesis is therefore the probability that we would get a black ball if BB is true.

Imagine that BB is true. That means both balls are black. What is the probability that the experiment would result in a black ball? That's easy – it's 100%! So, we put the number 1 in the Bayes Box as the likelihood for the BB hypothesis.

Now imagine instead that BW is true. That would mean one ball is black and the other is white. If this were the case and we did the experiment, what would be the probability of getting the black ball in the experiment? Since one of the two balls is black, the chance of choosing this one is 50%. Therefore, the likelihood for the BW hypothesis is 0.5, and that's why we put 0.5 in the Bayes' Box for the likelihood for BW.

In general, the likelihood is the probability of the data that you actually got, assuming a particular hypothesis is true. In this example it was fairly easy to get the likelihoods directly by asking "if this hypothesis is true, what is the probability of getting the black ball when we do the experiment?" Sometimes this is not so easy, and it can be helpful to think about ALL possible experimental outcomes/data you might have seen – even though ultimately, we just need to select the one that actually occurred.

Hypotheses	Possible Data	Probability
BB	Black Ball	1
	White Ball	0
BW	Black Ball	0.5
	White Ball	0.5

This table demonstrates a method for calculating the likelihood values, by considering not just the data that actually occurred, but all data that might have occurred. Ultimately, it is only the probability of the data which actually occurred that matters, so this is highlighted in blue.

The Mechanical Part

The third column of the Bayes' Box is the product of the prior probabilities and the likelihoods, calculated by simple multiplication. The result will be called "prior times likelihood", but occasionally we will use the letter h for these quantities. This is the un-

Notes

normalized posterior. It does not sum to 1 as the posterior probabilities should, but it is at least proportional to the actual posterior probabilities.

To find the posterior probabilities, we take the prior likelihood column and divide it by its sum, producing numbers that do sum to 1. This gives us the final posterior probabilities, which were the goal all along. The completed Bayes' Box is shown below:

Hypotheses	prior	likelihood	prior × likelihood	posterior
BB	0.5	1	0.5	0.667
BW	0.5	0.5	0.25	0.333
Totals:	1		0.75	1

We can see that the posterior probabilities are not the same as the prior probabilities, because we have more information now! The experimental result made BB a little bit more plausible than it was before. Its probability has increased from 1/2 to 2/3.

Interpretation

The posterior probabilities of the hypotheses are proportional to the prior probabilistic and the likelihoods. A high prior probability will help a hypothesis have a high posterior probability. To understand what this means about reasoning, consider the meanings of the prior and the likelihood. There are two things that can contribute to a hypothesis being plausible:

If the prior probability is high. That is, the hypothesis was already plausible, before we got the data.

If the hypothesis predicted the data well. That is, the data was what we would have expected to occur if the hypothesis had been true.

Bayes Rule:

Bayes' rule is an equation from probability theory and conditional probabilities.

For example, the left-hand side of the equation is $P(A|B)$ and that means the probability of A given B. That is, it's the probability of 'A' after taking into account the information 'B'. In other words, $P(A|B)$ is a posterior probability, and Bayes' rule tells us how to calculate it from other probabilities. Bayes' rule is true for any statements A and B.

$$P(A|B) = P(B|A) P(A) / P(B)$$

In Bayesian Statistics A has been replaced by H, and B has been replaced by D. The reason for these letters is that we should interpret H as hypothesis and D as data. Then we can interpret Bayes' rule as telling you the probability of a hypothesis given some data, in other words, a posterior probability.

$$P(H|D) = P(H) P(D|H) / P(D)$$

In Bayesian statistics, most of the terms in Bayes' rule have special names. Some

of them even have more than one name, with different scientific communities preferring different terminology. Here is a list of the various terms and the names we will use for them:

$P(H|D)$ is the posterior probability. It describes how certain or confident we are that hypothesis H is true, given that we have observed data D . Calculating posterior probabilities is the main goal of Bayesian statistics!

$P(H)$ is the prior probability, which describes how sure we were that H was true, before we observed the data D .

$P(D|H)$ is the likelihood. If you were to assume that H is true, this is the probability that we would have observed data D .

$P(D)$ is the marginal likelihood. This is the probability that we would have observed data D , whether H is true or not

Notes

Unit-3.3: Parametric and Non-Parametric Tests

Unit Objectives:

At the end of this unit the participant will be able to learn:

- Introduction to Parametric and Non-Parametric Tests
- Z-test
- t-test
- F-test
- Correlation and Regression
- Chi-square Test
- Factor Analysis

3.3.1 Introduction to Parametric and Non-Parametric Tests

Parametric Test

Parametric test is a testing procedure that requires assumption about the type of population or parameters.

Parametric tests have following advantages:

1. Parametric tests are more powerful here data is derived from interval and ratio measurement.
2. In parametric tests, it's assumed that the data follows normal distributions. Examples of parametric tests are (a) Z-Test, (b) T-Test and (c) F-Test.
3. Observations must be independent i.e., selection of any one item should not affect the chances of selecting any others be included in the sample.

The following tests are based on the assumption that the samples were drawn from normally distributed populations:

- F - test
- t - test
- Z - test

Non-Parametric Test

A group of alternative techniques known as non-parametric tests were developed since, it was not always possible to make a rigid assumption about the population distribution from which, and the samples were being drawn. The prominent examples of non-parametric test are:

- Chi Square test of Independence
- Goodness of fit

3.3.2 Z-test

1. One Sample Test One sample tests can be categorized into 2 categories.

z Test 1. When sample size is > 30

P_1 = Proportion in sample 1

P_2 = Proportion in sample 2

Example: You are working as a purchase manager for a company. The following information has been supplied by two scooter tire manufacturers.

	Company A	Company B
Mean life (in km)	13000	12000
S.D (in km)	340	388
Sample size	100	100

In the above, the sample size is 100; hence a Z-test may be used.

2. Testing the hypothesis about difference between two means: This can be used when two population means are given and null hypothesis is $H_0: P_1 = P_2$

Example: In a city during the year 2000, 20% of households indicated that they read Femina magazine. Three years later, the publisher had reasons to believe that circulation has gone up. A survey was conducted to confirm this. A sample of 1,000 respondents were contacted and it was found 210 respondents confirmed that they subscribe to the periodical 'Femina'.

From the above, can we conclude that there is a significant increase in the circulation of 'Femina'? Solution: We will set up null hypothesis and alternate hypothesis as follows:

Null Hypothesis is $H_0: \mu = 15\%$

Alternate Hypothesis is $H_A: \mu > 15\%$

This is a one-tailed (right) test

$$= 8.33$$

As the value of Z at 0.05 = 1.64 and calculated value of Z falls in the rejection region, we reject null hypothesis, and therefore we conclude that the sale of 'Femina' has increased significantly

3.3.3 T-test

T-test is used in the following circumstances: When the sample size $n < 30$. Discuss with following example:

There are two nourishment programmes: 'A' and 'B'. Two groups of children are subjected to this. Their weight is measured after six months. The first group of children subjected to the program 'A' weighed 44, 37, 48, 60, 41 kgs. at the end of programme. The second group of children were subjected to nourishment program 'B' and their weight was 42, 42, 58, 64, 64, 67, 62 kgs. at the end of the programme. From the above, can we conclude that nourishment programme 'B' increased the weight of the

Notes

children significantly, given a 5% level of confidence.

Null Hypothesis: There is no significant difference between Nourishment programme 'A' and 'B'. Alternative Hypothesis: Nourishment programme B is better than 'A' or Nourishment programme 'B' increase the children's weight significantly. Solution:

	Nourishment programme A			A Nourishment programme B	
X	$(X-X/\) = (X-46)$	$(X-X/)^2$	Y	$(Y-Y/) = (y-57)$	
44	-2	4	42	-15	
37	-9	81	42	-15	
48	2	4	58	1	
60	14	196	64	7	
41	-5	25	64	7	
			67	10	
			62	5	
Total	0	310	399	0	

Here

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left\{ \sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 \right\}$$

$$D.F. = (n_1 + n_2 - 2) = (5 + 7 - 2) = 10$$

$$s^2 = \frac{1}{10} \{310 + 674\} = 98.4$$

$$t = \frac{46 - 57}{\sqrt{98.4 \times \left(\frac{1}{5} + \frac{1}{7} \right)}}$$

$$= \frac{-11}{\sqrt{98.4 \times \left(\frac{12}{35} \right)}}$$

$$= \frac{-11}{\sqrt{33.73}} = -\frac{11}{5.8}$$

$$= -1.89$$

- F-Test:

Let there be two independent random samples of sizes n_1 and n_2 from two normal populations:

with variances σ_1^2 and σ_2^2 respectively. Further, let $s_1^2 = \frac{1}{n_1 - 1} \sum (X_{1i} - \bar{X}_1)^2$ and $s_2^2 = \frac{1}{n_2 - 1} \sum (X_{2i} - \bar{X}_2)^2$ be the variances of the first sample and the second samples respectively.

Then F - statistic is defined as the ratio of two χ^2 - variates. Thus, we can write

$$F = \frac{\frac{\chi_{n_1-1}^2}{n_1-1} = \frac{(n_1-1)s_1^2 / (n_1-1)}{\frac{\chi_{n_2-1}^2}{n_2-1} = \frac{(n_2-1)s_2^2 / (n_2-1)}{\frac{\sigma_1^2}{\sigma_2^2}}$$

Features of F- distribution

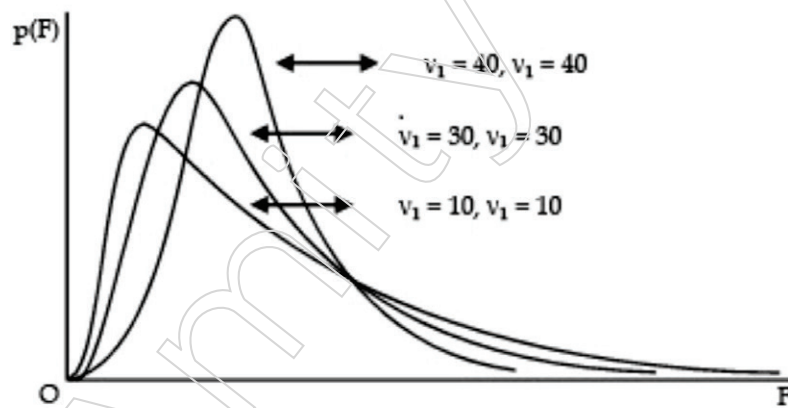
1. This distribution has two parameters $v_1 (= n_1 - 1)$ and $v_2 (= n_2 - 1)$.
2. The mean of F - variant with v_1 and v_2 degrees of freedom is $v_2 / (v_2 - 2)$

$$(v_2 / v_2 - 2)$$

We note that the mean will exist if $v_2 > 2$ and standard error will exist if $v_2 > 4$. Further, the mean > 1 .

3. The random variate F can take only positive values from 0 to ∞ . The curve is positively skewed.
4. For large values of v_1 and v_2 , the distribution approaches normal distribution.
5. If a random variate follows t-distribution with v degrees of freedom, then its square follows F-distribution with 1 and v d.f. i.e. $t_v^2 = F_{1,v}$

F and χ^2 are also related as $F_{v_1/v_2} = \chi_{v_1/v_1}^2$



3.3.4 Correlation & Regression

Various experts have defined correlation in their own words and their definitions, broadly speaking, imply that correlation is the degree of association between two or more variables.

Some important definitions of correlation are given below:

“Correlation is an analysis of covariation between two or more variables.”

A.M. Tuttle –

Notes

"If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in other(s) then they are said to be correlated."

L.R. Connor–

"When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."

Croxton and Cowden–

"Correlation analysis attempts to determine the 'degree of relationship' between variables".

YaLun Chou–

Correlation Coefficient: It is a numerical measure of the degree of association between two or more variables.

The Scope of Correlation Analysis

Existing of correlation between more than two variables can implies that these variables

(i) either tend to increase or decrease together or (ii) an increase (or decrease) in one is accompanied by the corresponding decrease (or increase) in the other. The questions of the type, whether changes in a variable are due to changes in the other, i.e., whether a cause and effect type relationship exists between them, are not answered by the study of correlation analysis. If there is a correlation between two variables, it may be due to any of the following situations:

1. One of the variable may be affecting the other: A correlation coefficient calculated from the data on quantity and corresponding price of cashew would only reveal that the degree of association between them is very high. It will not give us any idea about whether price is affecting demand of cashew or vice-versa. In order to know this, we need to have some additional information apart from the study of correlation. For example if, on the basis of some additional information, we say that the price of tea affects its demand, then price will be the cause and quantity will be the effect. The causal variable is also termed as independent variable while the other variable is termed as dependent variable.
2. The two variables may act upon each other: Cause and effect relation exists in this case also but it may be very difficult to find out which of the two variables is independent.

Example: If we have data on price of wheat and its cost of production, the correlation among them may be high because higher price of wheat may attract farmers to produce more wheat and more production of wheat may mean higher cost of production, assuming that it is an increasing cost industry. Further, the higher cost of production may in turn raise the price of wheat.

For the purpose of determining a relationship between the two variables in such situations, we can take any one of them as independent variable.

3. The two variables may be acted upon by the outside influences: In this case we might get a high rate of correlation between the two variables, however, apparently no cause and effect type relation seems to exist between them.

Example: The demands of the two commodities, say X and Y, may be positively correlated because the incomes of the consumers are rising. Coefficient of correlation obtained in such a situation is called a spurious or nonsense correlation.

4. A high value of the correlation coefficient may be obtained due to sheer coincidence (or pure chance): This is another situation of spurious correlation. Given the data on any two variables, one may obtain a high value of correlation coefficient when in fact they do not have any relationship

Example: A high value of correlation coefficient may be obtained between the size of shoe and the income of persons of a locality Merits and Limitations of Coefficient of Correlation.

The only merit of Karl Pearson's coefficient of correlation is that it is the most popular method for expressing the degree and direction of linear association between the two variables in terms of a pure number, independent of units of the variables. This measure, however, suffers from certain limitations, given below:

1. Coefficient of correlation r does not give any idea about the existence of cause and effect relationship between the variables. It is possible that a high value of r is obtained although none of them seem to be directly affecting the other. Hence, any interpretation of r should be done very carefully.
2. It is only a measure of the degree of linear relationship between two variables. If the relationship is not linear, the calculation of r does not have any meaning.
3. Its value is unduly affected by extreme items. 4. If the data are not uniformly spread in the relevant quadrants the value of r may give a misleading interpretation of the degree of relationship between the two variables. Just Like if there have some values concentrating around a point at first quadrant and there are similar type of concentration in third quadrant, the value of r will be very high although there may be no linear relation between the variables.
5. As compared with other methods, to be discussed later in this unit, the computations of r are cumbersome and time consuming.

Regression Analysis

If the coefficient of correlation calculated for bivariate data (X_i, Y_i) , $i = 1, 2, \dots, n$, in all fairness high and a cause-and-effect kind of relation is additionally believed to be existing between them, the subsequent logical step is to get a functional relation between these variables. This functional relation is understood as regression equation of Y on X. Since the coefficient of correlation is measure of the degree of linear association of the variables, we shall discuss only simple regression equation.

This doesn't, however, imply the non-existence of non-linear regression equations.

The regression equations are useful for predicting the worth of dependent variable for given value of the independent variable. As pointed out earlier, the nature of a

Notes

regression equation is different from the nature of a mathematical equation, e.g., if $Y = 10 + 2X$ is a mathematical equation then it implies that Y is exactly equal to 20 when $X = 5$. However, if $Y = 10 + 2X$ is a regression equation, then $Y = 20$ is an average value of Y when $X = 5$.

The term regression was first introduced by Sir Francis Galton in 1877. In his study of the relationship between heights of fathers and sons, he found that tall fathers were likely to have tall sons and vice-versa. However, the mean height of sons of tall fathers was lower than the mean height of their fathers and the mean height of sons of short fathers was higher than the mean height of their fathers. During this way, a bent of the civilisation race to regress or to return to a traditional height was observed. Sir Francis Galton referred this tendency of returning to the mean height of all men as regression in his research paper, "Regression towards mediocrity in hereditary stature". The term 'Regression', originated in this particular context, is now utilized in various fields of study, even though there may be no existence of any regressive tendency.

Simple Regression

For a bivariate data (X_i, Y_i) , $i = 1, 2, \dots, n$, we are able to have either X or Y as independent variable. If X is independent variable then we can estimate the average values of Y for a given value of X . The relation used for such estimation is called regression of Y on X . If on the other hand Y is employed for estimating the average values of X , the relation termed as regression of X on Y . For a bivariate data, there will always be two lines of regression. It will be shown later that these two lines are different, i.e., one can't be derived from the other by mere transfer of terms, because the derivation of every line is dependent on a different set of assumptions.

Line of Regression of Y on X

The general form of the line of regression of Y on X is $Y_{Ci} = a + bX_i$, where Y_{Ci} denotes the average or predicted or calculated value of Y for a given value of $X = X_i$. This line has two constants, a and b . The constant a is defined because the average value of Y when $X = 0$. Geometrically, it is the intercept of the line on Y -axis. Further, the constant b , gives the average rate of change of Y per unit change in X , is known as the regression coefficient.

The above line is known if the values of a and b are known. These values are estimated from the observed data (X_i, Y_i) , $i = 1, 2, \dots, n$.

3.3.5 Chi-square Test

A chi-square test (χ^2 test) is a statistical hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true, or asymptotically true, meaning that the sampling distribution can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

It is used in the following circumstances:

1. Sample observations should be independent i.e., two individual items should be included twice in a sample.

2. The sample should contain at least 50 observations
Or total frequency should be greater than 50.
3. There should be a minimum of five observations in any cell. This is called cell frequency constraint.

For instance: Chi-square

Persons	Age Group				Total
	Under 20-40	20-40	41-51	51 & Over	
Liked the car	146	78	48	28	300
Disliked the car	54	52	32	62	200
Total	200	130	80	90	500

Is there any significant difference between the age group and preference for the car?

Example: A company marketing tea claims that 70% of population in a metro drinks a particular brand (Wood Smoke) of tea. A competing brand challenged this claim. They took a random sample of 200 families to gather data. During the study period, it was found that 130 families were using this brand of tea. Will it be correct on the part of competitor to conclude that the claim made by the company does not holds good at 5% level of significance?

Solution:

Hypothesis H0 – People who drink Wood Smoke brand is 70%.

H0 – People who drink Wood Smoke brand is not 70%.

If the hypothesis is true then number of consumers who drink this particular brand is 200×0.7

= 140.

Those who do not drink that brand are $200 \times 0.3 = 60$

Degree of freedom = $D = 2 - 1 = 1$, since there are two groups

Group	Observed (O)	Expected (E)	O-E	(O-E) ²	(O-E) ² E
Those who drink branded tea	130	140	-10	100	0.714
Those who did not drink branded tea	70	60	+10	100	1.667
	200	200	0		

$$\chi^2 = \frac{(O-E)^2}{E} = 2.381$$

Notes

A 0.5 level of significance of for 1 d.f. is equal to 3.841 (From tables). The calculated value is 2.381 is lower. Therefore, we accept the hypothesis that 70% of the people in that metro drink Wood Smoke branded tea.

3.3.6 Factor Analysis

The main purpose of Factor Analysis is to group large set of variable factors into fewer factors.

Each factor will account for one or more component. Each factor a combination of many variables.

There are two most commonly employed factor analysis procedures or methods. They are:

1. Principle component analysis
2. Common factor analysis.

When the objective is to summarise information from a large set of variables into fewer factors, principle component factor analysis is used. On the other hand, if the researcher wants to analyse the components of the main factor, common factor analysis is used.

Example: Common factor – Inconvenience inside a car. The components may be:

1. Leg room
2. Seat arrangement
3. Entering the rare seat
4. Inadequate dickey space
5. Door locking mechanism

Summary:

- Define Correlation & Regression
- Factor Analysis
- Chi-square test
- Different kinds of parametric/ nonparametric test

Questions:

1. Discuss about different kind of parametric test with example.
2. What is Correlation?
3. What is the basic use of correlation?
4. How we can use regression and when?

Unit-3.4: Data Analysis

Notes

Unit Objectives:

At the end of this unit the participant will be able to learn:

1. Analyse the concept of SPSS
2. Identify the importance of data creation in SPSS
3. Analyse and run example on parametric test
4. Analyse and run example on non-parametric test

Purposes: Customer feedback about a two-wheeler manufactured by a company.

Method: The MR manager prepares a questionnaire to study the customer feedback. The researcher has identified six variables or factors for this purpose. They are as follows:

1. Fuel efficiency (A)
2. Durability (Life) (B)
3. Comfort (C)
4. Spare parts availability (D)
5. Breakdown frequency (E)
6. Price (F)

The questionnaire may be administered to 5,000 respondents. The opinion of the customer is gathered. Let us allot points 1 to 10 for the variables factors A to F. 1 is the minimum and 10 is the maximum. Let us assume that application of factor analysis has led to grouping the variables as A, B, D, E into factor-1

F into Factor -2

C into Factor - 3

Factor - 1 can be named as Technical factor;

Factor - 2 can be named as Price factor;

Factor - 3 can be termed as Personal factor.

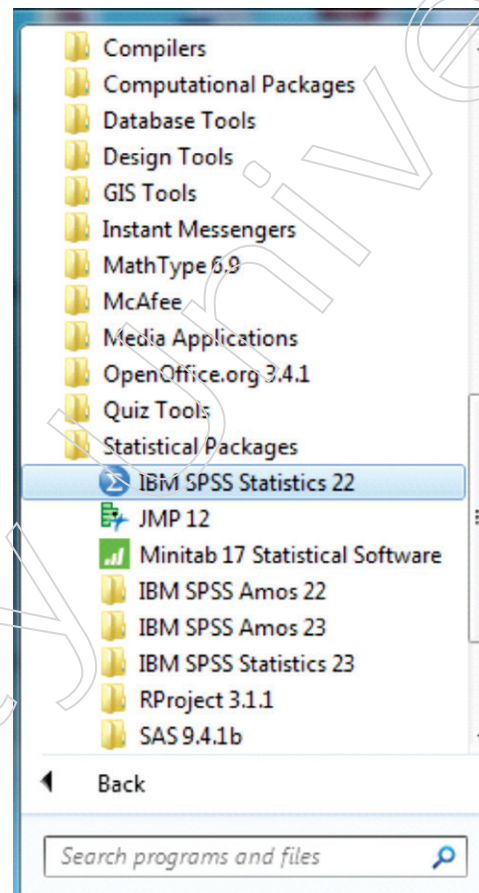
For future analysis, while conducting a study to obtain customers' opinion, three factors mentioned above would be sufficient. One main purpose of using factor analysis is to minimize the number of independent variables in the study. By having too many independent variables, the M.R study will suffer from following disadvantages:

1. Time for data collection is very high due to several independent variables.
2. Expenditure increases due to the time factor.
3. Computation time is more, resulting in delay.
4. There may be redundant independent variables.

Notes

3.4.1 Introduction to SPSS

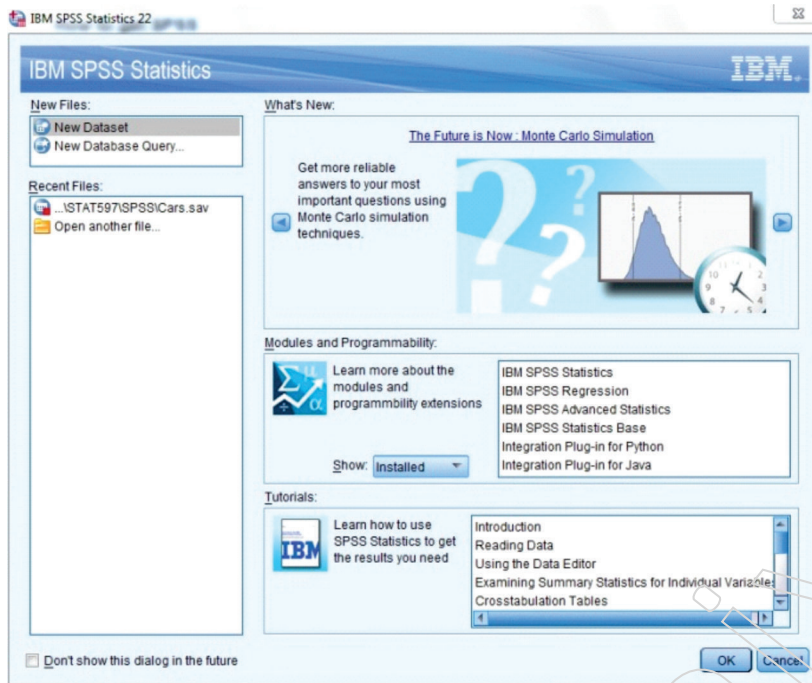
SPSS is a powerful statistical software program with a graphical interface designed for ease of use. Almost all commands and options can be accessed using pull down menus at the top of the main SPSS window. This design means that once you learn a few basic steps to access programs, it is very easy to expand your knowledge in using SPSS through the help files. To access the online SPSS help, you click on Help in the menu and then click on Topics if you want help by topic or on Tutorials for step-by-step hands-on guide. How to get SPSS is installed on all ITap (Information Technology at Purdue) machines in all ITaP labs around Statistical Packages Standard Software Programs campus. To get into the program: click



Start IBM SPSS Statistics 22

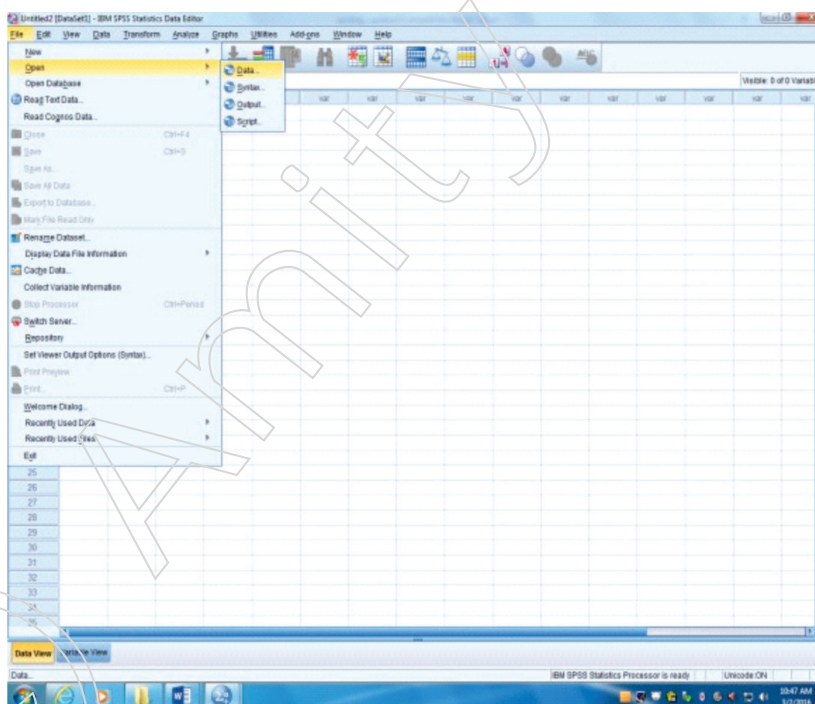
Opening SPSS Data

When SPSS is launched, a pop-up window (Error! Reference source not found.) with a few options will appear. Assume the goal is to analyze a data set, one can select New Dataset or open a file recently used or another file under Recent Files and then click OK. The other windows shows, What's New, Modules and Programmability and Tutorials, which help one to navigate SPSS.



Opening data from external files

Sometimes you have already entered the SPSS session as described above, worked on a data set for a while, and then want to open and work on another data set. You do not have to quit the current SPSS session to perform this. Simply click on the File menu, follow Open then Data... and find your file.



Opening CSV or Excel files

Once in SPSS, in the SPSS Data Editor click on File, then Open and then choose

Notes

data as shown in Figure 3, and Enter and the screen as shown in Figure 4 is given. In Look in: specify the location of the data file, under File name: specify its name; and under Files of type: specify the file type. The dataset we are working with is called Cars.csv as shown in Figure 4. (Download a copy of this data set [click here](#))

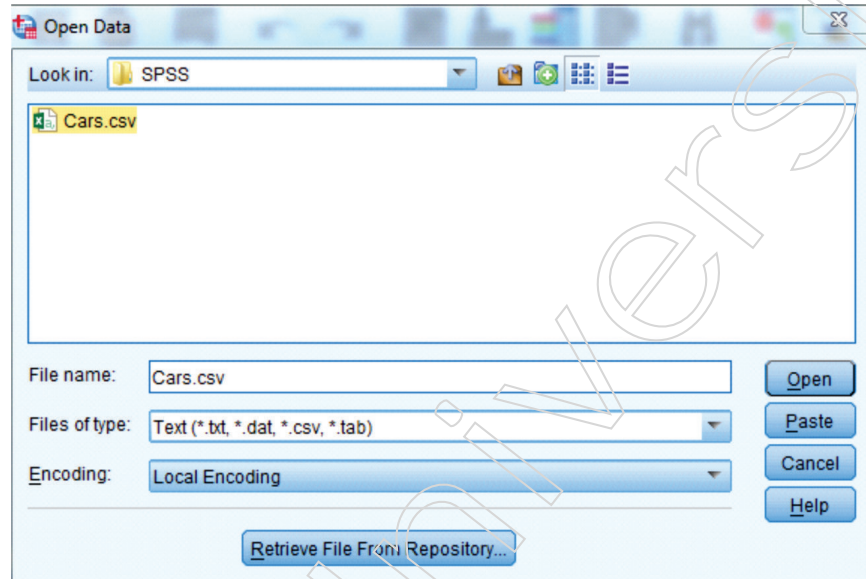
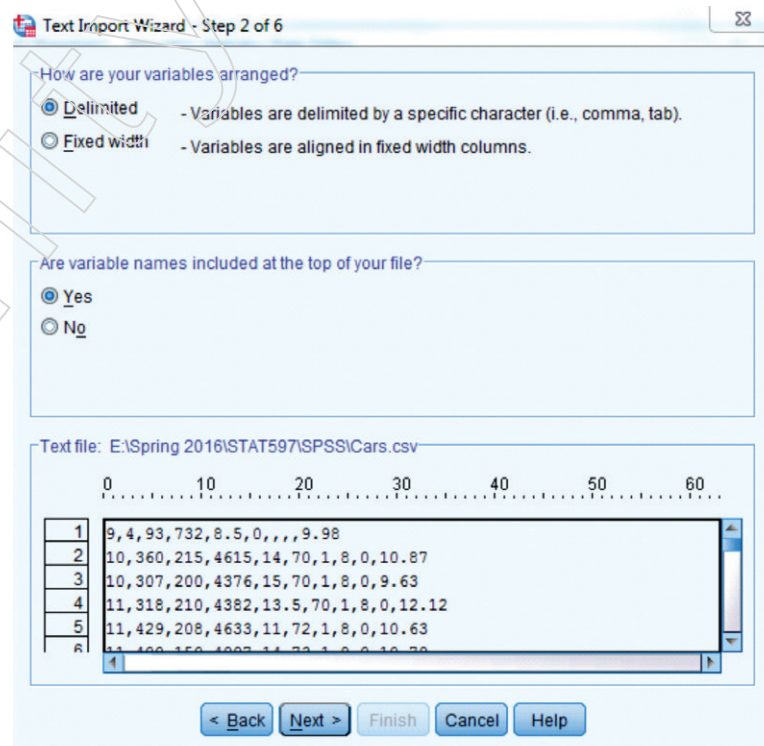


Figure below shows how one imports the data from Car.csv. Since this file has variable names at the top of the file, then click the Yes button under “Are variable names included at the top of your file?” Click on Next until the data pops in the Data View. Similar steps can be taken to import data from Excel and many spreadsheets and text files.

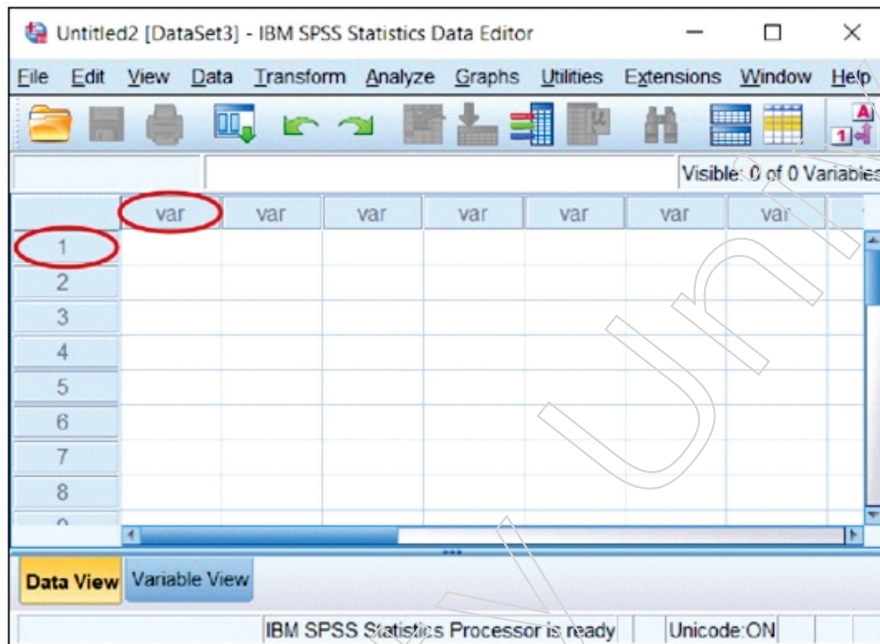


3.4.2 Data Creation in SPSS

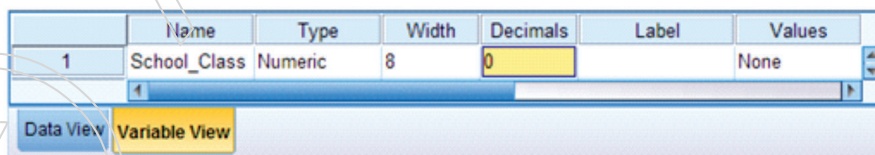
When open the SPSS program then open a blank spreadsheet in Data View. If already a data set open but want to create a new one, then click File

- New>Data to open a blank spreadsheet.

Here each of the columns is labelled “var.” The column names will represent the variables that we want to enter in our data set. Here each row is labelled with a number (“1”, “2”, “3”, and so on). The rows will represent cases that will be a part of our data set. When we enter values for our data in the spreadsheet cells, each value will correspond to a specific variable (column) and a specific case (row)



- Click the Variable View tab. Type the name for our 1st variable under the Name column. We can also enter other information about the variable, such as type (the default is “numeric”), width, decimals, label, etc. Type the name for each variable that we plan to include in your dataset. In this example, I will type “School_Class” since I plan to include a variable for the class level of each student (i.e., 1 = first year, 2 = second year, 3 = third year, and 4 = fourth year). I will also specify 0 decimals since my variable values will only include whole numbers. (The default is two decimals.)



Click the Data View tab. Any variable names that we entered in Variable View will now be included in the columns (one variable name per column). We can see that School_Class appears in the first column in this example

Notes

	School_Class	var	var
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Now we can enter values for each case. In this example, cases represent students. For each student, enter a value for class level in the cell that corresponds to the appropriate row and column. For example, the 1st person's information should appear in the 1st row, under the variable column School_Class. In this example, the 1st person's class level is "2", the second person's is "1", the third person's is "1", the fourth person's is "3", and so on

ID	School_Class
1	2
2	1
3	1
4	3
5	1
6	1
7	4
8	4
9	1
10	1
11	4
12	1
13	3

ID Variables versus Row Numbers

Now discuss a special type of variable called an ID variable, when data are collected, each piece of information is tied to a particular case. For an instance, you

distributed a survey as part of your data collection, and each survey was labelled with a number (“I,” “II,” etc.). In this example, the survey numbers essentially represent ID numbers: numbers that help us to identify which pieces of information go with which respondents in our sample. Without these ID numbers, we have no way of tracking which information goes with which respondent, and it would be impossible to enter the data accurately into SPSS. At the time of entering data into SPSS, we need to enter values for each variable that correspond to the correct person or object in our sample. It might seem like a simple solution to use the conveniently labelled rows in SPSS as ID numbers; we can enter our first respondent’s information in the row that is already labelled “I,” the second respondent’s information in the row labelled “II,” and so on.

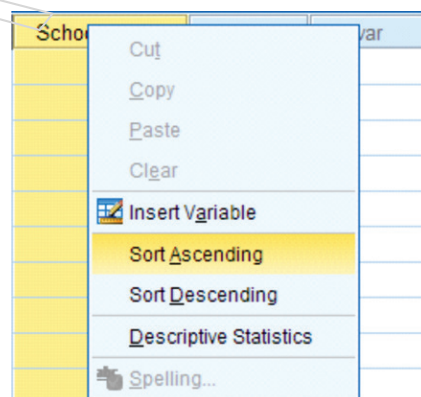
However, we should never rely on these pre-numbered rows for keeping track of the specific respondents in our sample. This is because the numbers for every row are visual guides only—they’re not attached to specific lines of data, and thus cannot be used to identify specific cases in our data. If our data become rearranged (e.g., after sorting data), the row numbers will no longer be associated with the same case as when we first entered the data. Again, the row numbers in SPSS aren’t attached to specific lines of info and might not be used to identify certain cases. Instead, you should create a variable in your dataset that is used to identify each case—for example, a variable called StudentID.

For an example that illustrates why using the row numbers in SPSS as case identifiers is flawed:

Let’s say that we have entered values for each person for the School_Class variable. We relied on the row numbers in SPSS to correspond to our survey ID numbers. Thus, for survey #1, we entered the first respondent’s information in row 1, for survey #2 we entered the second person’s information in row 2, and so on.

But suppose the data get reorganized in the spreadsheet view. A common way of reorganised data is by sorting. Sorting will rearrange the rows of data so that the values appear in ascending or descending order. If we right-click on any variable name, we can select “Sort Ascending” or “Sort Descending.” in the example below, the data are sorted in ascending order on the values for the variable School_Class.


But what happens if we need to view a specific respondent’s information? Or perhaps we need to double-check our entry of the data by comparing the original survey to the values you entered in SPSS. Now that the data have been rearranged, there is no way to identify which row corresponds to which participant/survey number.



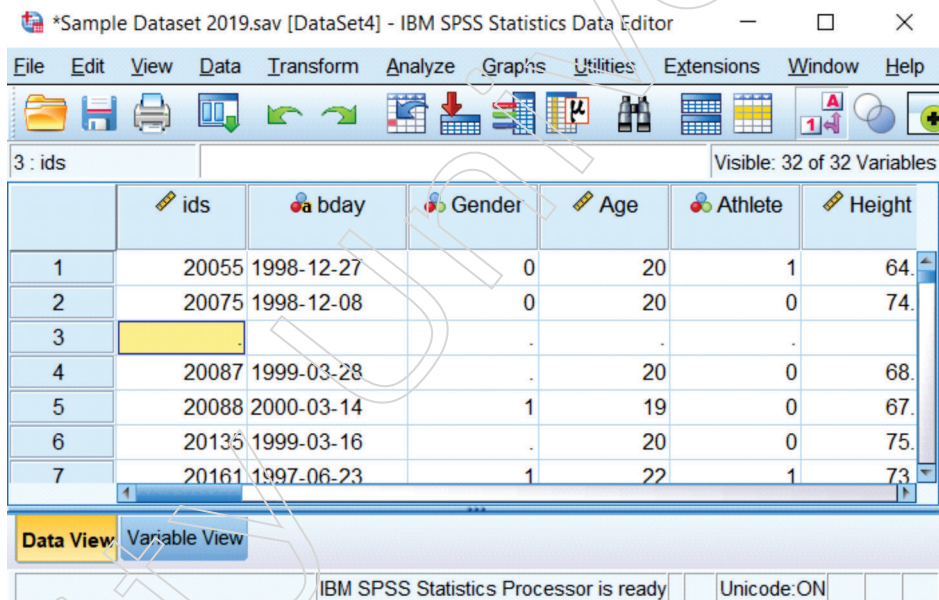
Notes

Inserting A Case

To insert a new case into a dataset:

- In Data View, click a row number or individual cell below where we want our new row to be inserted.
- We can insert a case in several ways: Click Edit > Insert Cases; Right-click on a row and select Insert Cases from the menu; or Click the Insert Cases icon ().

A new, blank row will appear above the row or cell we selected. Values for each existing variable in our dataset will be missing (indicated by either a "." or a blank cell) for our newly created case since we have not yet entered this information



*Sample Dataset 2019.sav [DataSet4] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

3 : ids Visible: 32 of 32 Variables

	ids	bday	Gender	Age	Athlete	Height
1	20055	1998-12-27	0	20	1	64.
2	20075	1998-12-08	0	20	0	74.
3		
4	20087	1999-03-28	.	20	0	68.
5	20088	2000-03-14	1	19	0	67.
6	20135	1999-03-16	.	20	0	75.
7	20161	1997-06-23	1	22	1	73.

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode: ON

Deleting A Case

To delete an existing case from a dataset:

- In the Data View tab, click the case number (row) that we wish to delete. This will highlight the row for the case we selected.
- Press Delete on our keyboard, or right-click on the case number and select "Clear". This will remove the entire row from the dataset.

Deleting A Variable

To delete an existing variable from a dataset:

- In the Data View tab, click the column name (variable) that we wish to delete. This will highlight the variable column.
- Press Delete on our keyboard, or right-click on the selected variable and click "Clear." The variable and associated values will be removed.

Alternatively, we can delete a variable through the Variable View window:

- Click on the row number corresponding to the variable we wish to delete. This will highlight the row.
- Press Delete on our keyboard, or right-click on the row number corresponding to the variable we wish to delete and click “Clear”.

3.4.3 Run Example on Parametric Test

	gender	age	ethnicity	gpa	p_learning	c_community	csoc_com	clm_com	s_community	ssoc_com
1	2	2	2	1.30	9	36	16	20	12	12
2	2	2	2	1.40	5	21	7	14	25	12
3	1	2	2	1.58	7	23	9	14	30	15
4	2	2	2	1.79	9	25	7	18	25	16
5	1	2	2	1.87	7	22	5	17	28	12
6	2	3	2	2.00	6	34	18	16	29	15
7	1	2	2	2.00	5	23	10	13	20	15
8	1	2	2	2.00	7	23	8	15	26	11
9	1	3	2	2.10	5	22	11	11	17	4
10	2	3	2	2.30	7	25	14	11	19	7
11	1	3	2	2.40	8	32	16	16	19	11
12	2	2	2	2.48	7	20	1	19	18	11
13	1	4	2	2.50	7	24	10	14	16	10
14	1	2	2	2.50	6	22	5	17	40	20
15	1	3	4	2.50	5	28	15	13	25	12
16	1	2	2	2.50	5	25	14	11	30	15
17	1	3	2	2.55	7	22	11	11	15	5
18	1	2	2	2.60	6	23	10	13	31	15
19	1	3	2	2.60	9	33	14	19	32	13
20	2	3	2	2.60	7	34	18	16	25	15
21	1	3	2	2.62	2	19	15	4	39	19
22	1	3	4	2.65	5	38	19	19	40	20
23	1	2	2	2.70	6	27	9	18	30	16
24	1	2	2	2.70	7	19	5	11	28	15

One sample Test of Hypothesis

The screenshot shows the SPSS 'Analyze' menu with the following path highlighted: **Analyze** > **Compare Means** > **One-Sample T Test...**. A red box with the text "Follow the menu as indicated." is overlaid on the 'One-Sample T Test...' option. The background shows a portion of the data table with columns 'lrm_com' and 's_community' visible.

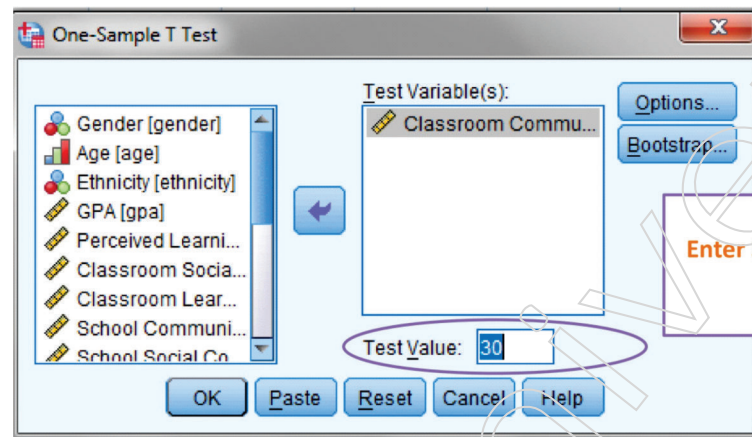
Notes

One sample Test of Hypothesis

In this example, we will test the following null hypothesis:

H_0 : There is no difference between the sample mean for the variable *Classroom Community* and $\mu = 30$.

Select and move the *Classroom Community* variable to the Test Variable(s) box.



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Classroom Community	169	28.84	6.242	0.480

3.4.4 Run Example on Non-parametric Test

The chi-square test could be used to determine if a basket of fruit contains equal proportions of apples, bananas, oranges, and peaches

Fruits	Count
orange	1
orange	1
mango	2
banana	3
lemon	4
banana	3
orange	1
lemon	4
lemon	4
orange	1
mango	2
banana	3
lemon	4
banana	3

orange	1
lemon	4
lemon	4

SPSS Steps:

Get the data

chisquare_fruit.sav [DataSet2] - SPSS Data Editor

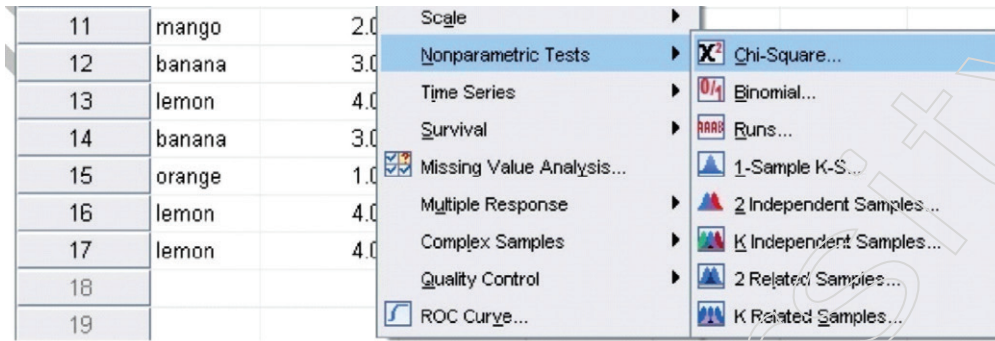
	fruits	count	var
1	orange	2.00	
2	orange	1.00	
3	mango	2.00	
4	banana	3.00	
5	lemon	2.00	
6	banana	3.00	
7	orange	1.00	
8	lemon	4.00	
9	lemon	4.00	
10	orange	1.00	

chisquare_fruit.sav [DataSet2] - SPSS Data Editor

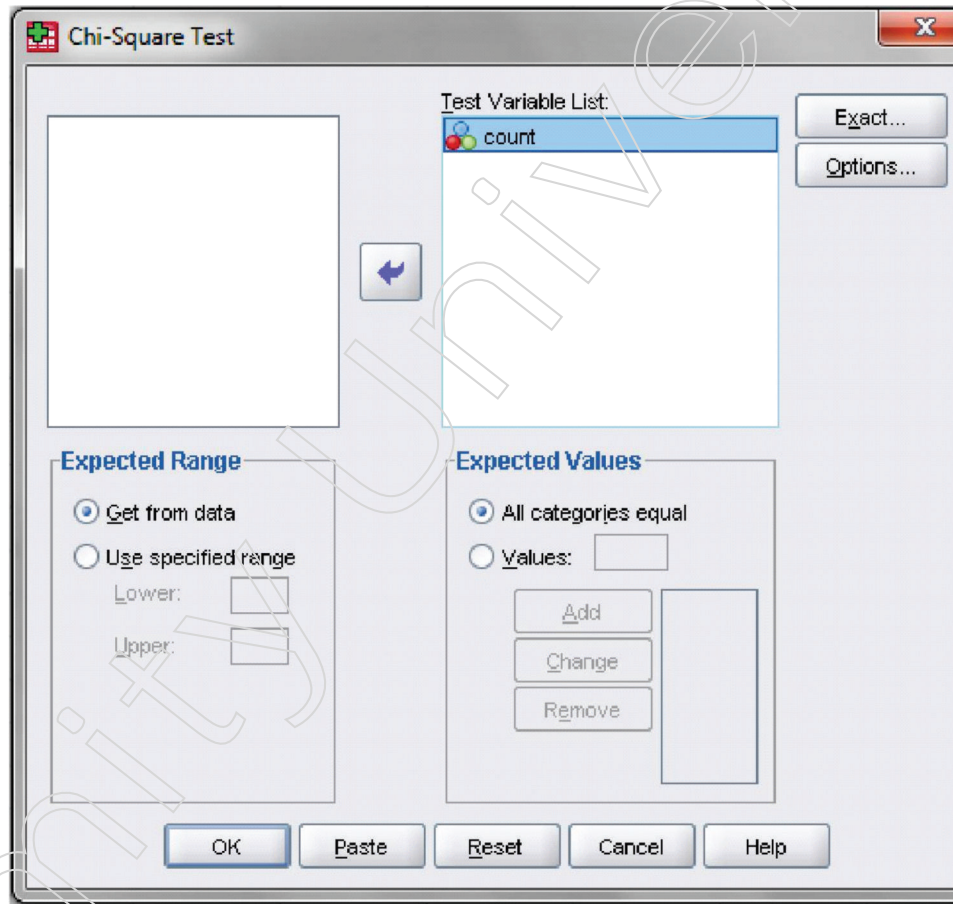
	fruits	count	var	var	var
1	orange	2.00			
2	orange	1.00			
3	mango	2.00			
4	banana	3.00			
5	lemon	2.00			
6	banana	3.00			
7	orange	1.00			
8	lemon	4.00			
9	lemon	4.00			
10	orange	1.00			

- Reports
- Descriptive Statistics
- Tables
- Compare Means
- General Linear Model
- Generalized Linear Models
- Mixed Models
- Correlate
- Regression
- Loglinear
- Neural Networks
- Classify
- Data Reduction

Notes



Get the count in the test variable list



Click ok and get out put following bellow.

	Observed N	Expected N	Residual
1	4	4.2	-.2
2.	4	4.2	-.2
3.	4	4.2	-.2
4	5	4.2	.8
Total	17		

	Count
Chi-Square	.176
df	3
Asymp. Sig.	.981

Interpretation

Here p value is 0.981 which is more than 0.05. Hence it is not significant and we fail to reject the null hypothesis and conclude that there is no significant difference in the proportions of apples, bananas, oranges, and peaches.

We could also test to see if a basket of fruit contains 10% apples, 20% bananas, 50% oranges, and 20% peaches. For this we have to define the proportions by checking the button "Values" and keep on adding

Summary

- Introduction to SPSS
- The way of Creating Data file
- Run example on parametric test

Questions

1. What is SPSS what is the usefulness of it.
2. Give the pictorial representation of the run an example on parametric / nonparametric test
3. Define how to create a data in useful manner.

Exercises:

1. _____ research is aimed at expanding knowledge and does not involve inventing or creating anything.
 - a. Basic
 - b. Exploratory
 - c. Action
 - d. Descriptive
2. Conducting an experiment on Newton's 3rd law of motion is an example of _____ research.
 - a. Action
 - b. Exploratory
 - c. Basic
 - d. Descriptive
3. Which of the following is an example of Applied research?

Notes

- a. Conducting an experiment on Einstein's Special Theory of Relativity
 - b. Devising solutions for arresting employee attrition
 - c. Conducting an archaeological study on a few historical artifacts
 - d. Conducting a survey related to preference of face wash products
4. Which of the following options is FALSE about empirical research?
- a. The findings are subject to verification by experiment or observation
 - b. You need to collect data to prove or disprove the hypotheses
 - c. Involves guidelines and techniques by which you can utilise historical sources, artifacts, and other evidence for researching and establishing facts
 - d. Is a data-based research technique
5. Comparing the carpentry tools used during the Gupta and Maurya dynasties is an example of _____ research.
- a. Explanatory
 - b. Exploratory
 - c. Descriptive
 - d. Historical
6. Which of the following research techniques are conducted by companies during later phases of decision-making?
- a. Descriptive
 - b. Exploratory
 - c. Explanatory
 - d. Both a and c
7. Which of the following is/are example/examples of Causal research?
- a. How incentives affect employee performance
 - b. How employee attrition affects profitability
 - c. How pricing strategies affect customer loyalty
 - d. All of these
8. Which of the following is NOT RELATED to Explanatory research?
- a. Formulating and testing research hypotheses
 - b. Unstructured
 - c. Highly structured
 - d. Conducted in the later phases of decision-making
9. Which of the following is NOT RELATED to Descriptive research?
- a. Unstructured

- b. Framing and asking research questions
 - c. Formulating and testing research hypotheses
 - d. Conducted in the later phases of decision-making
10. Read the below statements and identify the correct option.
- a. The findings of 1- _____ research offer a final conclusion or conclusive evidence to the research problem.
 - b. _____ research involves studying only one variable
 - c. _____ research is aimed at devising solutions for an immediate problem in a company, the society, or a person
 - d. _____ research is conducted to analyse how certain changes affect existing standard procedures
11. Which of the following is FALSE about hypothesis?
- a. Hypothesis is a tentative statement, which is subject to verification
 - b. Hypothesis is a tested, well-substantiated, complete explanation for a set of proven factors
 - c. Hypothesis is conceptually different from theory
 - d. Hypothesis is a testable relationship between at least two variables
12. A research firm conducts a study to establish the minimum purchasing power required for the medium and large retail stores as Rs. 150 million and Rs. 300 million, respectively. Identify the INCORRECT statement.
- a. The null hypothesis is - total purchasing power is less than Rs. 150 million
 - b. One of the alternative hypotheses is - total purchasing power is between Rs. 150 million and Rs. 300 million
 - c. The null hypothesis is - total purchasing power is more than Rs. 300 million
 - d. One of the alternative hypotheses is - total purchasing power is more than Rs. 300 million
13. A renowned automobile company is under the process of launching a new luxury car. The car is aimed at catering to the HNI (High Net Worth Individual) population. The company wants to conduct a detailed market research on the popular choices of luxury cars in the country. It recruits a research team, which comes up with the research problem - "Is luxury car a popular among HNI clients?" Which of the following is FALSE about the case?
- a. "The HNI population does not prefer luxury cars" can possibly be the null hypothesis
 - b. "At least 50% of the HNI population prefers luxury cars" can possibly be one of the alternative hypotheses
 - c. Only one alternative hypothesis exists in this case
 - d. "Less than 50% of the HNI population prefers luxury cars" can possibly be one of the alternative hypotheses

Notes

14. Read the below statements and identify the wrong one(s).
1. Complex hypothesis establishes a causal relationship between two variables
 2. Null hypothesis establishes a relationship among more than two variables
 3. An alternative hypothesis is used for a reverse strategy
 4. "Performance at work is not related to salary alone" is an example of alternative hypothesis
- a. Only 1 and 3 are wrong
 - b. Only 4 is wrong
 - c. Only 2 and 4 are wrong
 - d. All 4 statements are wrong
15. A renowned Fashion magazine conducts a market research to find out the need for advertisement. The research team constructs the below hypotheses: H₀: At least 30 % of the readers consists of women H₁: Less than 30 % of the readers consists of women What decision would the management take if the research team commits a type II error?
- a. The management invests unnecessarily in advertisements
 - b. The management does not invest in advertisements
 - c. The management prepares a budget for promotional cost
 - d. The management hires more salespersons to carry out extensive sales across the country
- a. 1 - Descriptive
2 - Exploratory
3 - Applied
4 - Historical
 - b. 1 - Exploratory
2 - Descriptive
3 - Applied
4 - Explanatory
 - c. 1 - Action
2 - Causal
3 - Applied
4 - Explanatory
 - d. None of these

Answers

- 1- a

- 2- c
- 3- b
- 4- c
- 5- d
- 6- d
- 7- d
- 8- b
- 9 -c
- 10 -b
- 11- b
- 12- c
- 13- c
- 14- d
- 15- b

Notes

© Amity University

Module-4: Inferential Statistics and Prescriptive Analytics

Key Learning Outcomes:

At the end of this module the participant will be able to:

1. Identify the importance of machine learning
2. Analyse the concept of regression
3. Identify the importance of data analysis

Structure:

Unit -4.1: Machine Learning

- 4.1.1 Challenges for Big Data Analytics
- 4.1.2 Introduction to Machine Learning
- 4.1.3 Concepts of Machine Learning
- 4.1.4 Use cases of Machine Learning in Research

Unit-4.2: Regression

- 4.2.1 Introduction to Regression
- 4.2.2 Ordinary Least Squares
- 4.2.3 Ridge Regression
- 4.2.4 Polynomial Regression
- 4.2.5 Bayesian Linear Regression
- 4.2.6 Lasso Regression
- 4.2.7 K Nearest Neighbours Regression
- 4.2.8 Logistic Regression & Classification tree
- 4.2.9 Clustering

Unit-4.3: Data Analysis

- 4.3.1 Unsupervised Learning
- 4.3.2 Creating Data through Designed Experiments
- 4.3.3 Creating Data through Active Learning
- 4.3.4 Creating Data through Reinforcement Learning

Unit-4.1: Machine Learning

Notes

Unit Objectives:

At the end of this unit, participants will be able to learn:

- What is big data
- The challenges for big data analysis.
- What is machine Learning
- Several Use case of Machine Learning.

4.1.1 Challenges for Big Data Analytics

In this digitalized world, we are producing a large number of knowledge (data) in every minute. The number of knowledge (data) produced in every minute makes it challenging to store, manage, utilize, and analyze it. Even large business enterprises are struggling to seek out the ways to form this huge amount of useful data. Today, the number of knowledge (data) produced by large business enterprises is growing, as mentioned before, at a rate of 40 to 60% per year. Simply storing this huge amount of knowledge (data) is not going to be all that useful and this can be the reason why organizations are looking at options like data lakes and big data analysis tools that can help them in handling big data to a great extent. Now, let's take a quick look at some challenges faced in Big Data analysis:

1. Lack of proper understanding of massive Big Data

Companies fail in their Big Data initiatives because of insufficient understanding. Employees might not know what data is, its storage, processing, importance, and sources. Data professionals may know what's occurring, but others mightn't have a transparent picture.

For example, if employees do not understand the importance of data storage, they may not keep the backup of sensitive data(knowledge). They may not use databases properly for storage. As a result, when this important data is required, it cannot be retrieved easily.

Solution

Big Data workshops and seminars must be held at companies for everybody. Basic training programs must be arranged for all the workers who are handling data regularly and are a part of the Big Data projects. A basic understanding of data concepts must be inculcated by all levels of the organization.

2. Data growth issues

One of the foremost promising challenges of Big Data is storing all these huge sets of data(knowledge) properly. The number of data(knowledge) being stored in data centers and databases of companies is increasing rapidly. As these data sets grow exponentially with time, it gets extremely difficult to handle.

Most of the info is unstructured and comes from documents, videos, audios, text

Notes

files and other sources. This implies that we cannot find them in databases.

Solution

In order to handle these large data sets, companies are choosing modern techniques, like compression, tiring, and reduplication.

- Compression is employed for reducing the amount of bits in the data that means reducing its overall size.
- Reduplication is that the process of reducing duplicate and unwanted data from a data set.
- Data tiring want to store data in different storage tiers. It ensures that the info is residing within the most appropriate storage space. Data tiers can be public cloud, private cloud, and flash storage, depending on the data size and importance.

Companies are also opting for Big Data tools, such as Hadoop, NoSQL and other technologies.

3. Confusion while Big Data tool selection

Companies often get confused while choosing the simplest tool for Big Data analysis and storage. Is HBase or Cassandra the best technology for data storage? Is Hadoop Map Reduce good enough or will Spark be a better option for data analytics and storage?

These questions bother companies and sometimes they're unable to find the answers. They end up making poor decisions and selecting an inappropriate technology. As a result, money, time, efforts and work hours are wasted.

Solution

The best procedure to go to seek professional help or hire experienced professionals who have rather more knowledge about these tools. Otherwise go for Big Data consulting. Here, consultants will give a recommendation of the best tools, based on our company's scenario. Based on their advice, we can work out a strategy and then select the best tool for us.

4. Lack of data professionals

To handle these modern technologies and Big Data tools, companies need skilled data professionals may be data scientists, data analysts and data engineers who are much more experienced in working with the tools and making sense out of big data sets. Data handling tools are changes rapidly, for that reason companies face a problem of lack of Big Data professionals.

Solution

Companies are investing more cash within the recruitment of skilled professionals. They also can arranged training programs to the existing staff to get the most out of them.

5. Securing data

Securing these vast sets of knowledge is one of the daunting challenges of massive information. Often companies are so busy in understanding, storing and analyzing their data sets that they push data security for later stages. But this is not a smart move as unprotected data repositories can become breeding grounds for malicious hackers.

Solution

Companies are recruiting more cyber security professionals to protect their data. Other steps taken for securing data like:

- Data encryption
- Data segregation
- Identity and access control
- Implementation of endpoint security
- Real-time security monitoring
- Use Big Data security tools, such as IBM Guardian

6. Integrating data from a variety of sources

Data in a company comes from a range of sources, such as social media pages, ERP applications, customer logs, monetary reports, e-mails, presentations and reports created by employees. Combining all this data to prepare reports is a challenging task.

Data integration is crucial for analysis, reporting and business intelligence, so it has to be perfect.

Solution

Companies need to solve their information integration problems by purchasing the right tools. Some of the best data integration tools are mentioned below:

- Talent Data Integration
- Centerprise Data Integrator
- ArcESB
- IBM InfoSphere
- Xplenty
- Informatica Power Center
- CloverDX
- Microsoft SQL
- QlikView
- Oracle Data Service Integrator

In order to place massive Data to the best use, corporate need to begin doing things differently. This means hiring better staff, changing the management, reviewing existing business policies and the technologies being used. To boost deciding they can hire a Chief Data Officer – a step that is taken by many of the fortune 500 companies.

Notes

7. Need for Synchronization Across Disparate Data Sources

As information sets are getting larger and more diverse, there is a big challenge to incorporate them into an analytical platform. If this is unmarked, it'll make a gap and result to wrong messages and insights.

4.1.2 Introduction to Machine Learning

Introduction to Machine Learning for Beginners

We have seen Machine Learning as a buzzword for the past few years, the rationale for this could be the high amount of information production by applications, the rise of computation power within the past few years and therefore the development of higher quality of algorithms. Machine Learning is used anywhere from automating mundane tasks to offering intelligent insights, industries in every sector try to benefit from it. We may already be using a device that utilizes it. For example, a wearable fitness tracker like Fitbit, or an intelligent home assistant like Google Home. But there are much more examples of ML in use.

- Prediction — Machine learning may also be utilized in the prediction systems. Considering the loan example, to compute the probability of a fault, the system will need to classify the available data in groups.
- Image recognition — Machine learning can be used for face detection in an image as well. There is a separate category for each person in a database of several people.
- Speech Recognition — It's the translation of spoken words into the text. It's employed in voice searches and more. Voice user interfaces include voice dialing, call routing, and appliance control. It can also be employed a simple data entry and the preparation of structured documents.
- Medical diagnoses — ML is trained to recognize cancerous tissues.
- Financial industry and trading — companies use ML in fraud investigations and credit checks.

4.1.3 Concept of Machine Learning

What's Machine Learning?

Why will we have to care about machine learning?

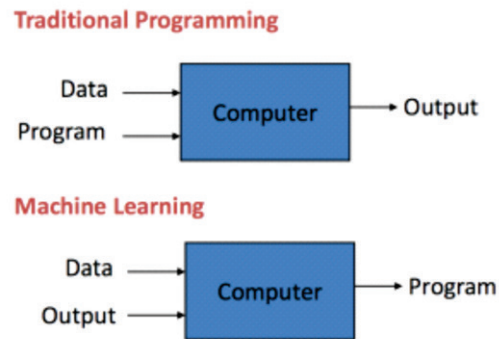
“A breakthrough in machine learning would be worth ten Microsoft's Entrepreneur”
Bill Gates, Former Chairman, Microsoft

Machine Learning is getting computers to program themselves. If programming is automation, then machine learning is automating the method of automation.

Writing software is that the bottleneck, we don't have enough good developers. Let the information do the work rather than people. Machine learning is that the way to make programming scalable.

In Traditional Programming: Data and program is run on the PC to provide the output.

But in Machine Learning: Data and output is run on the PC to make a program. This program can be used in traditional programming.



Machine Learning used

Machine learning used in several fields such as

- Web search: ranking page by clicking for promote the site as a 1 site.
- Computational biology: rational design drugs in the computer based on past experiments.
- Finance: decide who to send what credit card offers to. Evaluation of risk on credit offers. How to decide where to invest money.
- E-commerce: Predicting client churn. Whether or not or not a group action fallacious.
- Space exploration: space probes and radio astronomy.
- Robotics: how to handle uncertainty in new environments. Self-driving car.
- Information extraction: Ask questions over databases across the web.
- Social networks: Knowledge on relationships and preferences. Machine learning to extract value from knowledge.
- Debugging: Use in technology issues like debugging. Labour intensive method. May recommended suggest where the bug could be.

Key Elements of Machine Learning

There are tens of thousands of machine learning algorithms and hundreds of new algorithms are developed every year. But each and every machine learning algorithm has three basic components:

- **Representation:** how to represent knowledge.
Examples likes decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensembles and others.
- **Evaluation:** the way to evaluate candidate programs (hypotheses).
Examples likes as accuracy, prediction and recall, squared error, likelihood, posterior probability, cost, margin, entropy k-L divergence and others.
- **Optimization:** the approach candidate programs are generated known as the search method. Like as combinatorial optimization, convex optimization, constrained optimization.

Notes

All machine learning algorithms are combinations of these three components. A framework for understanding all algorithms.

Types of Learning

There are four types of machine learning:

- **Supervised learning:** (also called inductive learning) Training data includes desired outputs. This is spam this is not, learning is supervised.
- **Unsupervised learning:** Training data does not include desired outputs. Example is clustering. It's difficult to explain what's good learning and what's not.
- **Semi-supervised learning:** Training data includes a few desired outputs.
- **Reinforcement learning:** Rewards from a sequence of actions. All varieties like it, it's the foremost ambitious type of learning. Supervised learning is that the most mature, foremost studied and the type of learning utilized by most machine learning algorithms. Learning with supervision is much easier than learning without supervision. Inductive Learning is where we are given examples of a function in the form of data (x) and the output of the function (f(x)). The goal of inductive learning is to learn the function for new data (x).
- **Classification:** when the function being learned is discrete.
- **Regression:** when the function being learned is continuous.
- **Probability Estimation:** when the output of the function is a probability.

Machine Learning in Practice

Machine learning algorithms are only a very small part of using machine learning in practice as a data analyst or data scientist. the process often looks like:

- Start Loop
- Understand the domain, prior knowledge and goals. Talk to domain experts. Often the goals are very unclear. We frequently have more things to try than we can possibly implement.
- Data integration, selection, cleaning and pre-processing. This is often the most time-consuming part. It's important to have high quality info. The more data we have, the more it sucks because the data is dirty. Garbage in, garbage out.
- Learning models. The fun part. This part is very mature. The tools are general.
- Interpreting results. Sometimes it does not matter how the model works as long it delivers results. Several domains require that the model is understandable.
- Consolidating and deploying discovered knowledge. The majority of projects that are successful in the lab are not used in practice. It is very hard to get something used. 6. End Loop
- It's not a one-shot process; it's a cycle. We need to run the loop until we get a result that we can use in practice. Also, the info can change, requiring a new loop.

4.1.4 Use Case of Machine Learning:

The different kind of use case about machine Learning are discussed below such as,

1. Customer Service Automation

Managing the increasing number of online customer interactions features a big challenge for several organizations because they don't have the sufficient customer support staff to accommodate the quantity of inquiries they're receiving and therefore the old solution of outsourcing issues to a call centre is solely unacceptable for many of today's customers. Advances in machine learning algorithms have made it possible for chat bots and other automated systems to fill these needs with automating routine and low priority tasks, companies can free up employees to handle more high-level customer service. When implemented properly, machine learning in business can streamline issue resolution and make sure that customers can get the kind of helpful assistance that turns them into loyal brand advocates.

2. Cyber Security

As networks become increasingly complex, cyber security experts have worked hard to retort to the ever-expanding scope of security threats. Since there introduced rapid change in malware and hacking techniques which make harder to counter, but the proliferation of Internet of Things (IoT) devices has fundamentally altered the cyber security landscape. Attacks can come from anywhere, at any time, and in any form. Fortunately, machine learning algorithms have allowed cyber security efforts to stay walking with these rapid changes. Predictive analysis makes it possible to spot and attenuate threats faster than ever, and machine learning can track user behaviour within a network to identify irregularities and gaps in existing security measures.

3. Visual Perception

By using machine learning applications, more and more devices now have feature object visualising capabilities. An autonomous vehicle, for example knows another car when it sees one, even if programmers didn't provide it with an actual example of that car to use as a reference. Retail stores are even using this technology to assist speed up the checkout process. Cameras detect the things customers place in their cart and may automatically charge their accounts at the time of leaving the shop.

4. Fraud Detection

Now the quantity of monetary transactions happening online has raised consumer awareness about various forms of fraud. While the purchaser enjoys the convenience of having the ability to form purchases and payments online, they need to understand that their financial data is being protected in the process. MasterCard companies and banks have responded by turning to machine learning algorithms that may review vast amounts of transactional data to spot suspicious activity. While these sorts of checks are nothing new, machine learning in business has drastically expanded and accelerated the scope of these reviews. Consistent with industry research, machine learning solutions can detect up to 95 percent of fraud and minimize investigation time by 70 percent.

Notes

5. Communication

Avoiding mistakes and misunderstandings is very important in any quite communication, but especially so for today's businesses. Whether it's electronic mail correspondence, customer reviews, video conferencing, or text-based documents altogether their varied forms, simple grammatical errors, inappropriate tone, or inaccurate translations can cause a spread of problems. Machine learning programs have taken communication far beyond the heady days of Microsoft's Clippy. Thanks to natural language processing, real-time language translation, and speech recognition, these machine learning examples are able to help people communicate clearly and accurately. While many of us wish to complain about autocorrect features, they also appreciate being saved from embarrassing mistakes and inappropriate tone.

6. Digital Marketing

Much of today's marketing initiatives are dispensed online through a spread of digital platforms and software applications. As companies gather data about customers and their purchasing habits, marketing teams can use that information to form a posh picture of their audience and identify which individuals are more likely to seek out their products and services. Machine learning algorithms help marketers to form sense of all that data, identifying key trends and features that allow them to segment opportunities more narrowly. The identical technology enables digital marketing automation on an enormous scale. Ad platforms can be setup to dynamically identify new potential customers and direct the acceptable marketing material to them in the right place at the correct time.

As machine learning continues to advance, the range of applications and use cases will definitely expand within the 2020s. With the new decade just getting underway, it's worth keeping an eye fixed how machine learning use cases will be deployed to boost efficiency, reduce costs, and deliver better user experiences.

7. Process Automation

Intelligent Process Automation (IPA) is the combination of artificial intelligence and automation by the utilisation of machine learning. From automating manual data entry, to more complex use cases like automating insurance risk assessments. The cognitive technology like natural language processing, machine vision and deep learning, machines can augment traditional rule-based automation and overtime learn to try and do them better. Most IPA solutions already done by utilizing ML-powered capabilities beyond simple rule-based automation. The business benefits are much more extensive than cost saving and include better use of costly equipment or highly skilled employees, faster decisions and actions, service and merchandise innovations, and overall better outcomes. By using ML in over rate, within the enterprise the human worker to focus on product innovation and service improvement; allowing the corporate to transcend conventional performance trade-offs and achieve unparalleled levels of quality and efficiency.

8. Sales Optimization

The enterprises are saving consumer data for years, because it's also the place with the foremost potential for immediate financial impact from implementing machine learning. That's why every enterprise needing to gain a competitive edge are applying

ML to both marketing and sales challenges in order so as accomplish strategic goals. Some popular marketing techniques that depend on machine learning models include intelligent content and ad placement or predictive lead scoring. By adopting machine learning within the enterprise, companies can rapidly evolve and personalize content to meet the ever-changing needs of prospective customers. ML models are also being used for customer sentiment analysis, sales forecasting analysis, and customer churn predictions. With these solutions, sales managers are alerted before to specific deals or customers that are risk.

9. Collaboration

The key to getting the foremost out of machine learning in the enterprise lies within the enterprise tapping into the capabilities of both machine learning and human intelligence. ML-enhanced collaboration tools have the potential to spice up efficiency, quicken the innovation of latest ideas and lead to improved outcomes for teams that collaborate from disparate locations. Nemertes' 2018 UC and collaboration concluded that about 41 percent of enterprises plan to use AI in their unified communications and collaboration applications. Some uses cases in the collaboration space include:

- Video intelligence, audio intelligence and image intelligence can add context to content being shared, making it simpler for customers to find the files they require. Image intelligence coupled with object detection, text and handwriting recognition helps improve meta data indexing for enhance search.
- Real time language translation, facilitates communication and collaboration between global workgroups in their native languages.
- Integrating chatbots into team applications enables linguistic communication like alerting team members or polling them for status updates.

That is just the tip of the iceberg, machine learning offers significant potential benefits for companies adopting it as part of their communications strategy to reinforce data access, collaboration and control of communication endpoints.

Summary

- Discuss about Big data challenges.
- Concept of machine Learning.
- Different Use Case of Machine Learning

Questions

1. Discuss about the Different Challenges of big data and how to overcome from that.
2. Define several use care of machine learning.
3. Discuss about the several use of Machine Learning.

Notes

Unit-4.2: Regression

Unit Objectives:

At the end of this unit, participants will be able to learn:

- Learn about regression
- Various types of Regression
- Learn about Clustering.

4.2.1 Introduction to Regression

If the coefficient of correlation calculated for bivariate data (X_i, Y_i) , $i = 1, 2, \dots, n$, is logically high and a cause-and-effect type of relation is also believed to be existing between them, the next logical step is to obtain a functional relation between these variables. This functional relation is known as regression equation. The coefficient of correlation is measure of the degree of linear association of the variables.

The regression equations are useful for predicting the value of dependent variable with respect to the given value of the independent variable. The characteristic of a regression equation is different from the characteristic of a mathematical equation, e.g., if $Y = 10 + 2X$ is a mathematical equation then it implies that Y is exactly equal to 20 when $X = 5$.

However, if $Y = 10 + 2X$ is a regression equation, then $Y = 20$ is an average value of Y when $X = 5$.

The term regression was first introduced by Sir Francis Galton in 1877.

In his study of the relationship between heights of fathers and sons, he found that tall fathers were likely to have tall sons and vice-versa. The average height of sons of tall fathers was lower than the average height of their fathers and the average height of sons of short fathers was higher than the average height of their fathers. In this way, a tendency of the human race to regress or to return to a normal height was observed. Sir Francis Galton referred this tendency of returning to the average height of all men as regression in his research paper, "Regression towards mediocrity in hereditary stature". The term 'Regression', originated in this particular context, is now used in various fields of study, even though there may be no existence of any regressive tendency.

4.2.2 Ordinary Least Squares Regression (OLS)

It is identify with another name linear regression (simple or multiple counting on the amount of explanatory variables).

Here a model with p explanatory variables, the OLS regression model writes:

$$Y = \beta_0 + \sum_{j=1..p} \beta_j X_j + \varepsilon$$

where Y is that the dependent variable,

β_0 , is that the intercept of the model, X_j corresponds to the j th explanatory variable of the model ($j = 1$ to p), and e is the random error with expectation 0 and variance σ^2 .

Here we consider n observations, the estimation of the anticipated value of the dependent variable Y for the i th observation is given by:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij}$$

The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values. This minimization results in the subsequent estimators of the parameters of the model:

$$[\beta = (X'DX)^{-1} X'Dy \quad \sigma^2 = 1/(W - p^*) \sum_{i=1}^n w_i(y_i - \hat{y}_i)] \text{ where}$$

β is the that vector of the estimators of the β_i parameters,

X is that the matrix of the explanatory variables preceded by a vector of 1s,

y is that the vector of the n observed values of the dependent variable

p^* is that the number of explanatory variables to which we add 1 if the intercept isn't fixed,

w_i is that the weight of the i th observation,

W is that the sum of the w_i weights,

D may be a matrix with the w_i weights on its diagonal.

The vector of the predicted values can be written as follows:

$$\hat{y} = X (X' DX)^{-1} X'Dy$$

Limitation of OLS regression

The limitations of the OLS regression come from the constraint of the inversion of the $X'X$ matrix: it's required that the rank of the matrix is $p+1$, and a few numerical problems may arise if the matrix isn't well behaved. If the matrix rank equals q where q is strictly lower than $p+1$, few variables are far away from the model, either because they're constant or because they belong to a block of collinear variables.

Variable Selection within the OLS Regression

An automatic selection of the variables is performed if the user selects a too high number of variables compared to the amount of observations. Theoretically the limit is $n-1$, as with greater values the $X'X$ matrix becomes non-invertible.

The deleting of a number of the variables may however not be optimal: in some cases we'd not add a variable to the model because it's almost collinear to some other variables or to a block of variables, but it'd be that it would be more relevant to truncate a variable that is already within the model and to the new variable.

For the above reason and also, in order to handle the cases where there a lot of explanatory variables, other methods have been developed.

Prediction

Linear regression is usually use to predict outputs' values for new samples

Notes

4.2.3 Ridge Regression

Ridge regression is a method of model tuning, used to analyse any data that suffers from multi collinearity. When the issue of multi collinearity occurs, least-squares are unbiased, and variances are large, thus the results in predicted values to be far away from the actual values.

The cost function for ridge regression:

$$\text{Min}(\|Y - X(\theta)\|^2 + \lambda\|\theta\|^2)$$

Lambda is that the penalty term. λ also denoted by an alpha parameter in the ridge function. By changing the values of alpha, we're controlling the penalty term. Higher the values of alpha, larger is that the penalty and therefore the magnitude of coefficients is minimised.

- It shrinks the parameters to prevent multi collinearity
- It reduces the model complexity by coefficient shrinkage

For any variety of regression machine learning models, the same regression equation forms the base which is written as:

$$Y = XB + e$$

Where Y is that the dependent variable, X represents the independent variables, B is that the regression coefficients to be calculable, and e represents the errors are residuals. Once we tend the lambda function to this equation, the variance that's not evaluated by the overall model is considered. Once the data is ready and identified to be part of L2 regularization, there are steps that one can undertake.

Standardization

In ridge regression, at first standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. Which make a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. For standardization all ridge regression calculations are based on standardized variables. Once the ultimate regression coefficients are displayed, they're adjusted into their original scale. However, the ridge trace is on a consistent scale.

Bias and Variance Trade-off

Bias associated in variance trade-off is mostly difficult once it involves to building ridge regression models on an actual dataset. However, following the general trend which one needs to remember is:

- The bias increases as λ increases.
- The variance decreases as λ increases.

Assumptions of Ridge Regressions

The assumptions of ridge regression are the similar as that of linear regression: linearity, constant variance, and independence. However, as ridge regression doesn't provide confidence limits, the distribution of errors to be traditional needn't to be assumed.

4.2.4 Polynomial Regression

Here we can do a polynomial regression on the data to fit a polynomial equation to it. It is very difficult to match a linear regression line low value of error. Hence, we are only able to use the polynomial regression to match a polynomial line so that we are able to achieve a minimum error or minimum cost function. The equation of the polynomial regression be:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2$$

Now we can say a general equation of a polynomial regression is:

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \dots + \theta_n X^n + \text{residual error}$$

Pros of using Polynomial Regression

- Polynomial provides the best approximate relationship between the dependent and independent variable.
- A Broad range of function can be fit under it.
- Polynomial basically fits a wide range of curvature.

Cons of Polynomial Regression

- If there present one or two outliers in the data then they can seriously affect the results of the nonlinear analysis.
- These are too sensitive to the outliers.
- In addition, there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression.

4.2.5 Bayesian Linear Regression

In the Bayesian analysis, we have got to formulate linear regression using probability distributions with respect to point estimates. Here the response, y , isn't estimated as one value, but is assumed to be drawn from a probability distribution. The model for Bayesian Regression with the response sampled from a normal distribution is:

$$y \sim N(\beta^T X, \sigma^2 I)$$

Output, y is generated from a Gaussian Distribution characterized by a mean and variance.

The linear regression towards the mean = the transpose of the weight matrix X the predictor matrix.

Variance is that the square of the standard deviation σ (multiplied by the scalar matrix because this is often a multi-dimensional formulation of the model). Main focus of Bayesian Regression isn't to search out the single "best" value of the model parameters, but also to work out the posterior distribution for the model parameters. This is because the model parameters are assumed to come back from a distribution. The posterior probability of the model parameters is conditional upon the training inputs and outputs:

Notes

$$P(\beta|y, X) = \frac{P(y|\beta, X) * P(\beta|X)}{P(y|X)}$$

Here, $P(\beta|y, X)$ is the posterior probability distribution of the model parameters given the inputs and outputs. Which is adequate to the likelihood $P(y|\beta, X)$ of the info, multiplied by the prior probability of the parameters and divided by a normalization constant. It is a simple expression of Bayes Theorem, the basic underpinning of Bayesian Inference:

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Normalization}}$$

In contrast to Ordinary least square method, here have a posterior distribution for the model parameters that's proportional to the likelihood of the info multiplied by the prior probability of the parameters. In the following main two primary benefits of Bayesian Regression are

- **Priors:** If we've got domain knowledge about the model parameters then we are able to include them in our model, unlike within the frequentist (is a sort of statistical inference that pulls conclusions from sample data by emphasizing the frequency or proportion of the information) approach which assumes everything there's to understand about the parameters comes from the data. If we've got no estimates ahead of time, then we can use non-informative priors for the parameters such as a traditional distribution.
- **Posterior:** The results of performing Bayesian Regression may be a distribution of possible model parameters based on the information and also the before quantify our uncertainty about the model: if we have fewer datum, the posterior distribution are more displayed.

As the amount of datum increases, the likelihood washes out the prior, and within the case of infinite data, the outputs for the parameters converge to the values obtained from OLS.

The formulation of model parameters as distributions, we start with an initial estimate, our prior, and as we gather more evidence, our model becomes less wrong. Bayesian reasoning may be a natural extension of our intuition. Often, we have an initial hypothesis, and as we collect data that either supports or disproves our ideas, we alter our model of the world (ideally this is how we would reason)!

Implementing Bayesian Linear Regression towards the Mean

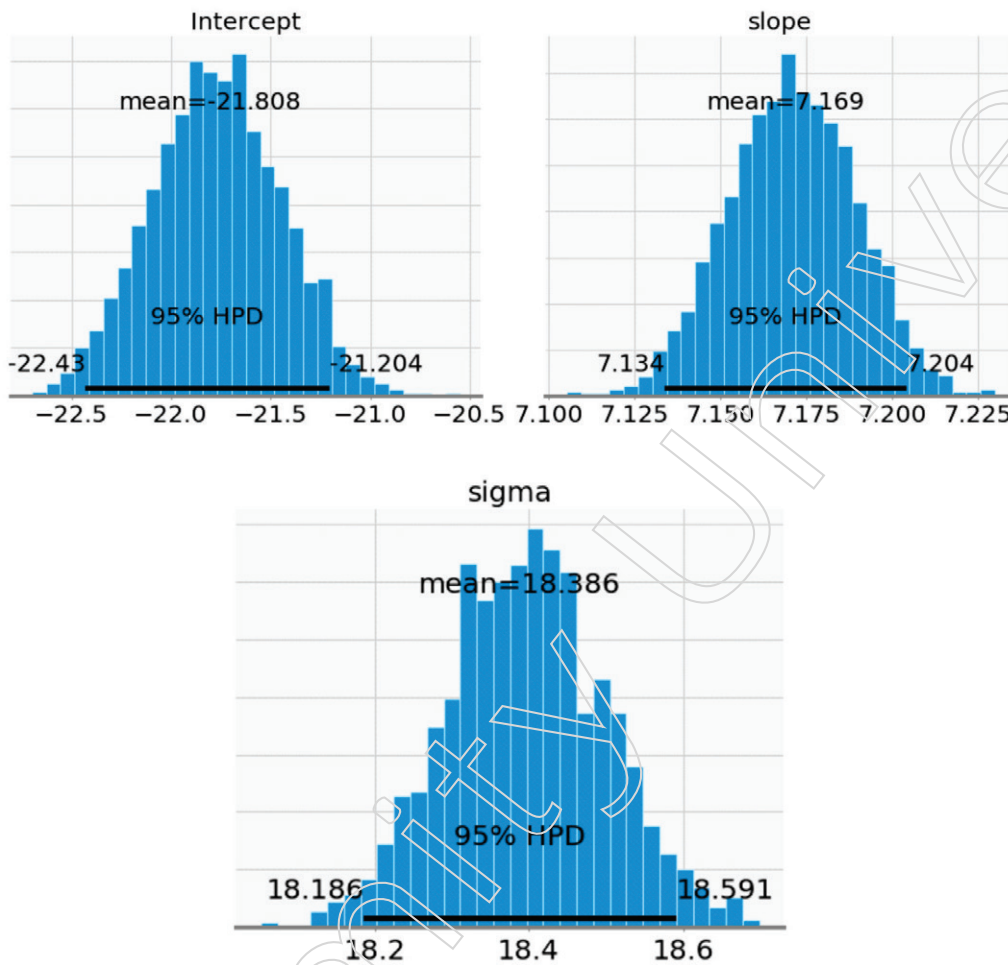
For practice, evaluating the posterior distribution for the model parameters is intractable for continuous variables, so we use sampling methods to draw samples from the posterior in order to approximate the posterior. The technique of drawing random samples from a distribution to approximate the distribution is one application of Monte Carlo methods.

Bayesian Linear Modelling Application

Here we'll not discuss about the code but the fundamental procedure for implementing Bayesian Regression, i.e.: specify priors for the model parameters (may be normal distributions), creating a model mapping the training inputs to the training

outputs, and then have a Markov Chain Monte Carlo (MCMC) algorithm draw samples from the posterior distribution for the model parameters. The end result will be posterior distributions for the parameters. We can inspect these distributions to introduce a way of what is occurring.

The first pic show the approximations of the posterior distributions of model parameters. These are the result of 1000 steps of MCMC, meaning the algorithm drew 1000 steps from the posterior distribution.

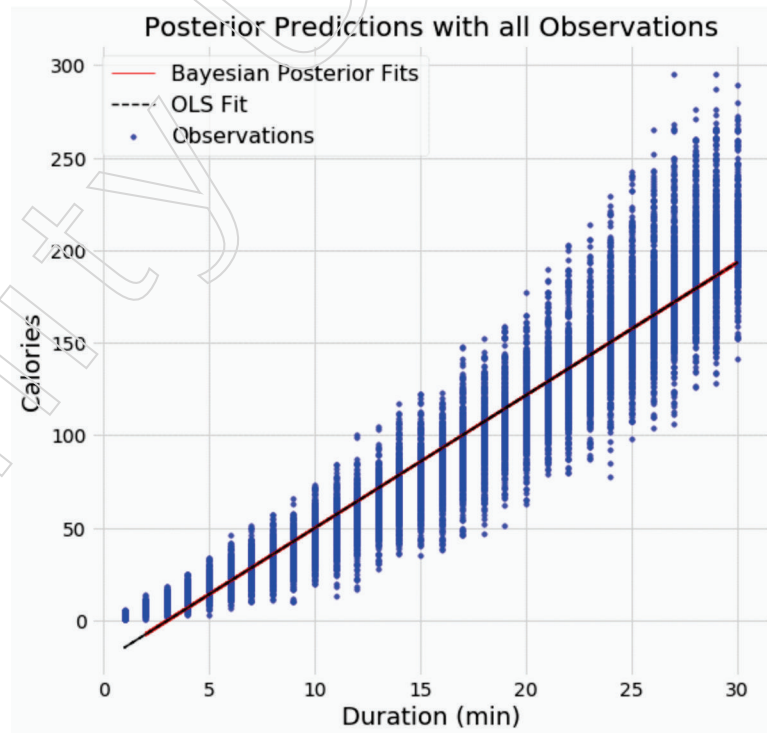
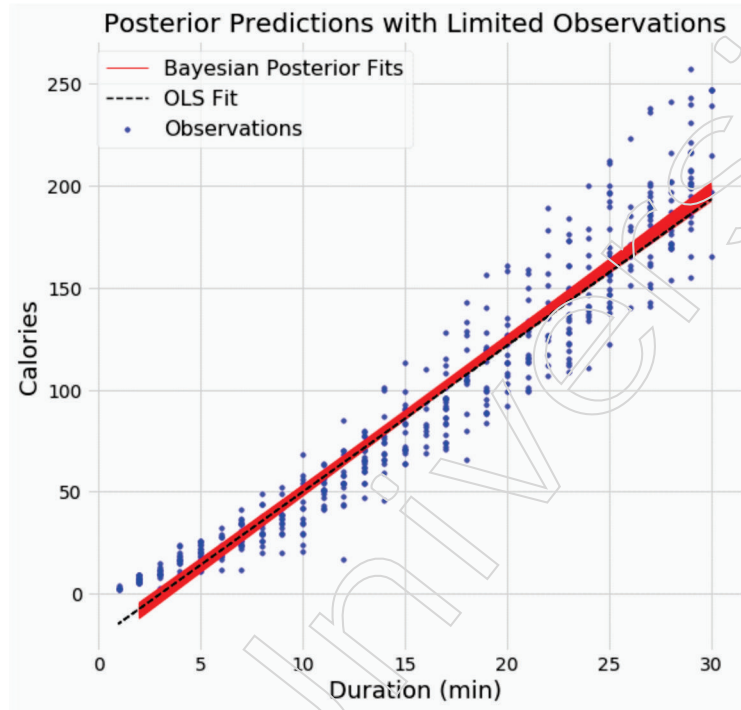


If we tend to compare the mean values for the slope and intercept to those obtained from OLS (the intercept from OLS was -21.83 and the slope was 7.17), we tend to see that they are very similar. However, whereas we are able to use the mean as one purpose estimate, we also have a range of possible values for the model parameters. As the number of data points increases, this range will shrink and converge one a single value representing greater confidence in the model parameters.

When we wish to show the linear match from a Bayesian model, rather than showing of solely estimate, we are able to draw a variety of lines, with each one representing a different estimate of the model parameters. As the number of data points increases, the lines begin to overlap because there is less uncertainty in the model parameters. In order to demonstrate the effect of the number of data points in the model, I used two models, the first, with the resulting fits shown on the left, used 500

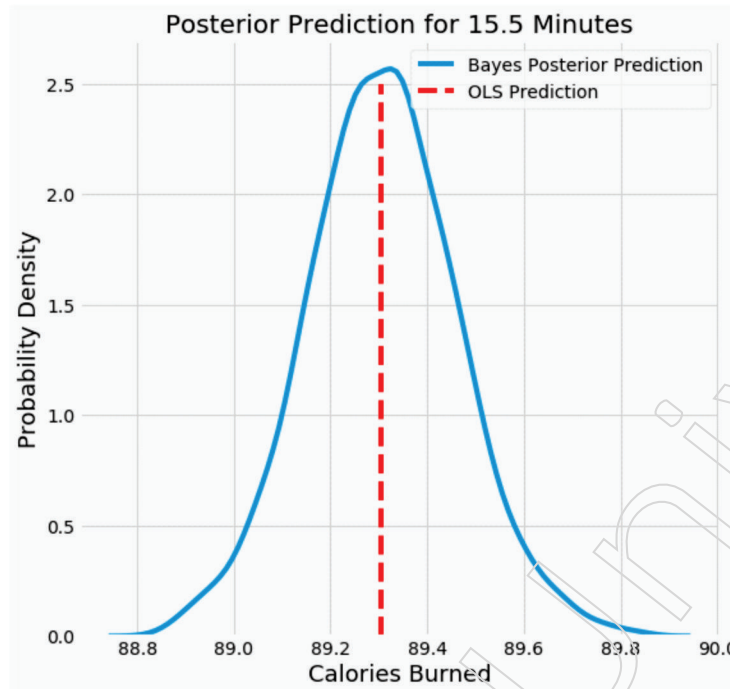
Notes

data points and the one on the right used 15000 data points. Each graph shows 100 possible models drawn from the model parameter posteriors.



There is much more variation in the fits when using fewer data points, which represents a greater uncertainty in the model. With all of the data points, the OLS and Bayesian Fits are nearly identical because the priors are washed out by the likelihoods from the data.

When predicting the output for a single data point using our Bayesian Linear Model, we also do not get a single value but a distribution. Following is the probability density plot for the number of calories burned exercising for 15.5 minutes. The red vertical line indicates the point estimate from OLS.



We see that the probability of the number of calories burned peaks around 89.3, but the full estimate is a range of possible values.

4.2.6 Lasso Regression

What is lasso regression? Lasso regression formula and example.

Lasso regression is one of the regression models that are available to analyze the data. LASSO stands for Least Absolute Shrinkage and Selection Operator. It was developed in 1989. It's basically an alternate to the classic method of least squares estimate to avoid many of the problems with over fitting once we have an outsized number of independent variables

Lasso regression is one of the regularization methods that creates frugal models with large number of features, where large means either of the below two things:

1. Massive enough to reinforce the tendency of the model to over-fit.
2. Large enough to cause computational challenges. This matter can arise in case of millions or billions of features.

Lasso regression performs L1 regularization that's it adds the penalty such as absolutely the worth of the magnitude of the coefficients. Here the minimization objective is as followed.

Minimization objective = LS Obj + λ (sum of absolute value of coefficients)

Notes

Where LS Obj stands for Least Squares Objective (linear regression objective) without regularization and λ is the turning factor to control the amount of regularization and the bias will increase with the increasing value of λ and the variance will decrease as the amount of shrinkage (λ) increases.

What large coefficient signifies?

Using large coefficient, we're putting a large emphasis on the particular characteristics that it can be a good predictor of the outcome. And when it is too large, the algorithm starts modelling complex relations to calculate the output & ends the over fitting for the particular data. Lasso regression merges a factor with the sum of the absolute value of the coefficients.

Now let us discuss lasso regression formula with an example:

The lasso regression estimate is defined as

Here the turning factor λ controls the strength of penalty, that is

- When $\lambda = 0$: We get same coefficients as simple linear regression
- When $\lambda = \infty$: All coefficients are zero
- When $0 < \lambda < \infty$: We get coefficients between 0 and that of simple linear regression

So, when λ is in between the two extremes, we are balancing the below two ideas.

- Fitting a linear model of y on X
- Shrinking the coefficients

But the character of L1 regularization penalty causes some coefficients to be shrunken to zero. Hence, not like ridge regression, lasso regression is in a position to perform variable choice within the linear model. So as the value of λ increases, more coefficients will be set to value zero (provided fewer variables are selected) and so among the nonzero coefficients, more shrinkage will be employed. The below working example will explain it well.

Working example:

For analyzing the prostate-specific substance and therefore the clinical measures among the patients united nation agency where close to have their prostates removed, ridge regression will provide smart results provided there are a good number of true coefficients. But if there are solely a couple of coefficients to predict the results lasso regression is the higher option to have accurate results since lasso can perform better than ridge when the coefficients are few.

Why Use Lasso Regression?

The advantage of lasso regression compared to least squares regression lies in the bias-variance trade-off.

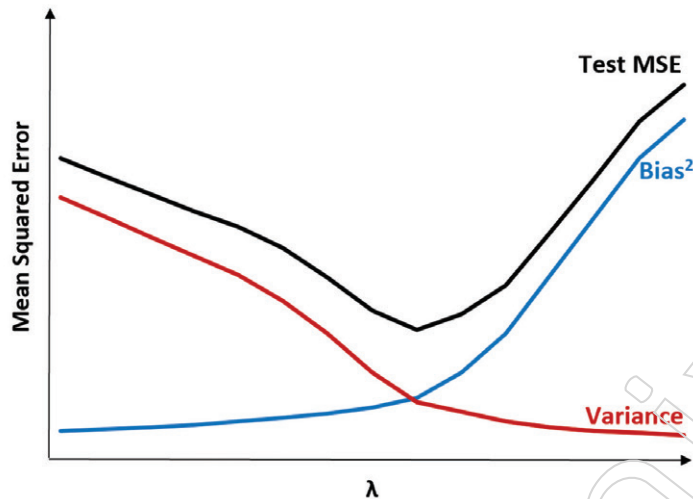
Recall that mean squared error (MSE) is a metric we can use to measure the accuracy of a given model and it is calculated as:

$$\text{MSE} = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

$$MSE = \text{Variance} + \text{Bias}^2 + \text{Irreducible error}$$

The basic idea of lasso regression is to introduce a little bias so that the variance can be substantially reduced, which leads to a lower overall MSE.

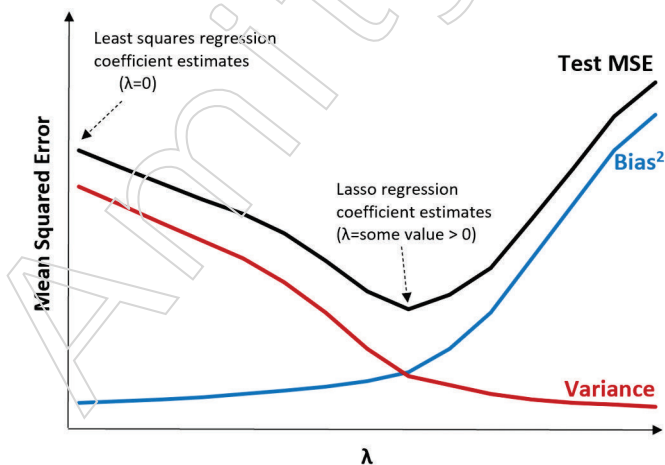
To illustrate this, consider the following chart:



Notice that as λ increases, variance drops substantially with very little increase in bias. Beyond a certain point, though, variance decreases less rapidly and the shrinkage in the coefficients causes them to be significantly underestimated which results in a large increase in bias.

We can see from the chart that the test MSE is lowest when we choose a value for λ that produces an optimal trade-off between bias and variance.

When $\lambda = 0$, the penalty term in lasso regression has no effect and thus it produces the same coefficient estimates as least squares. However, by increasing λ to a certain point we can reduce the overall test MSE.



This means the model fit by lasso regression will produce smaller test errors than the model fit by least squares regression.

Steps to Perform Lasso Regression in Practice

The following steps may be accustomed to perform lasso regression:

Notes

Step 1: Calculate the correlation matrix and VIF values for the predictor variables.

First, we must always produce a correlation matrix and calculate the variance inflation factor values for every variable quantity.

If we detect high correlation between predictor variables and high VIF values (some texts define a “high” VIF value as 5 while others use 10) then lasso regression is likely appropriate to use.

However, if there’s no multi collinearity present within the data then there could also be no need to perform lasso regression within the first place. Instead, we are able to perform ordinary least squares regression.

Step 2: Fit the lasso regression model and choose a worth for λ .

Once we determine that lasso regression is acceptable to use, we can fit the model (using popular programming languages like R or Python) using the optimal value for λ .

To determine the optimal value for λ , we are able to fit several models using different values for λ and choose λ to be the value that produces very cheap test MSE.

Step 3: Compare lasso regression to ridge regression and ordinary method of least squares regression.

Lastly, we are able to compare our lasso regression model to a ridge regression model to work out least squares regression model to determine which model produces all time low test MSE by using k-fold cross-validation.

Depending on the link between the predictor variables and therefore the response variable, it’s entirely possible for one of these three models to outperform the others in different scenarios.

4.2.7 K Nearest Neighbours Regression

k-nearest neighbours algorithm (k-NN) may be a non-parametric classification method initially developed by Evelyn Fix and Joseph Hodges in 1951, again later expanded by Thomas Cover. It’s used for classification and regression and also the input consists of the k closest training examples in data set. Whereas the output depends on whether k-NN is employed for classification or regression

- In k-NN classification, the output could be a class membership. An object is classed by a plurality vote of its neighbours, with the article being assigned to the category commonest among its k nearest neighbours (k may be a small positive integer). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbour.
- In k-NN regression, the output is that the property value for the item. This value is that the average of the values of k nearest neighbours.

It is a useful technique to assign weights to the contributions of the neighbours, so the nearer neighbours contribute more to the common than the more distant ones. For instance a typical weighting scheme consists in giving each neighbour a weight of $1/d$, where d is that the distance to the neighbour.

For k-NN regression the neighbours are taken from a group of objects the object property value is known. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the info.

KNN Algorithm

- Load the info
- Initialize K to our chosen number of neighbours
- For each example within the data
- Calculate the distance between the query example and also the current example from the info.
- Add the space and therefore the index of the instance to an ordered collection
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- Pick the primary K entries from the sorted collection
- Get the labels of the chosen K entries
- If regression, return the mean of the K labels
- If classification, return the mode of the K labels

4.2.8 Logistic Regression & Classification Tree

Logistic Regression and Decision Tree classification are two of the most popular and basic classification algorithms none of the algorithms is better than the other and one's superior performance is often credited to the nature of the data being worked upon.

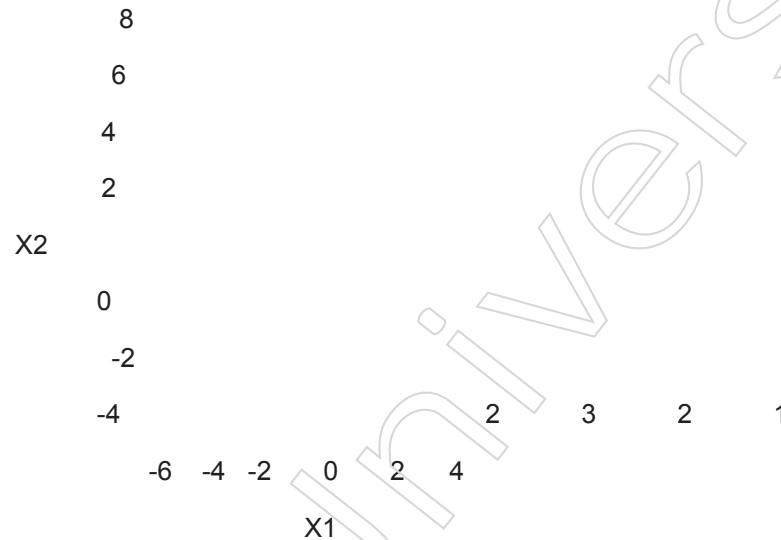
We can compare the two algorithms on different categories –

Criteria	Logistic Regression	Decision Tree Classification
Interpretability	Less interpretable	More interpretable
Decision Boundaries	Linear and single decision boundary	Bisects the space into smaller spaces
Ease of Decision Making	A decision threshold has to be set	Automatically handles decision making
Over fitting	Not prone to over fitting	Prone to over fitting
Robustness to noise	Robust to noise	Majorly affected by noise
Scalability	Requires a large enough training set	Can be trained on a small training set

In a Classification problem, we have a training sample of n observations on a class variable Y that takes values $1, 2, \dots, k$, and p predictor variables, X_1, \dots, X_p . We are supposed to find a model for predicting the value so f_Y from new X values. In theory, the solution is simply a partition of the X space in to k disjoint sets, A_1, A_2, \dots, A_k , such that the predicted value of Y is j if X belongs to A_j , for $j = 1, 2, \dots, k$. If the X variables take

Notes

ordered values, two classical solutions are linear discriminate analysis and another one is nearest neighbour classification. These methods output sets A_j with pie Classification tree methods output rectangular sets A_j by recursively partitioning the data set one X variable at a time for easy interpret. For example, Figure 1 gives an example wherein there are three classes and two X variables to show the decision tree structure. A key advantage of the tree structure is its applicability to any number of variables, whereas the plot on its left is limited to at most two.



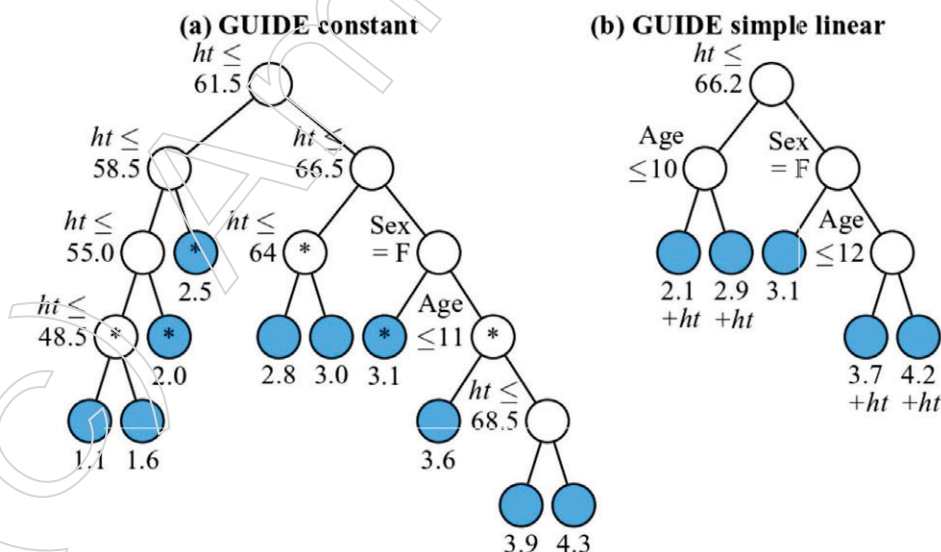
The first published classification tree algorithm is THAID. Employing a measure of node impurity based on the distribution of the observed Y values in the node, THAID splits a node by exhaustively searching over all X and S for the split X^*S^* that minimizes the total impurity of its two child nodes. If X takes ordered values, the set S is an interval of the form $(-\infty, c]$. Otherwise, S is a subset of the values taken by X . The process is applied recursively on the data in each child node. Splitting stops if the relative decrease in impurity is below a pre specified threshold. Algorithm 1 gives the pseudo code for the basic steps.

Algorithm 1 Pseudo code for tree construction by exhaustive search

- Start at the root node.
- For each X , find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split X^*S^* that gives the minimum overall X and S .
- If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.
- Choose the variable X^* associated with the X_r that has the smallest significance probability.
- Find the split set X^* belongs to S^* that minimizes the sum of Gini indexes and use it to split the node into two child nodes.
- If a stopping criterion is reached, exit. Otherwise, apply steps 2–5 to each child node.
- Prune the tree with the CART method

A regression tree is similar to a classification tree, except that the Y variable takes ordered values and a regression model is fitted to each node to give the predicted values of Y. Historically, the first regression tree algorithm is AID [36], which appeared several years before THAID. The AID and CART regression tree methods follow Algorithm 1, with the node impurity being the sum of squared deviations about the mean and the node prediction the sample mean of Y. This yields piece-wise constant models. Although they are simple to interpret, the prediction accuracy of these models often lags behind that of models with more smoothness. It can be computationally impracticable, however, to extend this approach to piece-wise linear models, because two linear models (one for each child node) must be fitted for every candidate split a regression tree algorithm by Quinlan,22 uses a more computationally efficient strategy to construct piecewise linear models. It first constructs a piece-wise constant tree and then fits a line a regression model to the data in each leaf node. Because the tree structure is the same as that of a piece-wise constant model, the resulting trees tend to be larger than those from other piece-wise linear tree.

Because the total model complexity is shared between the tree structure and the set of node models, the complexity of a tree structure often decreases as the complexity of the node models increases. Therefore, the user can choose a model by trading off tree structure complexity against node model complexity. Piece-wise constant models are mainly used for the insights their tree structures provide. But they tend to have low prediction accuracy, unless the data are sufficiently informative and plentiful to yield a tree with many nodes. The trouble is that the larger the tree, the harder it is to derive insight from it. Trees (a) and (e) are quite large, but because they split almost exclusively on ht, we can infer from the predicted values in the leaf nodes that FEV increases monotonically with ht. The piece-wise simple linear (b) and quadratic (c) models reduce tree complexity without much loss (if any) of interpretability. Instead of splitting the nodes, ht now serves exclusively as the predictor variable in each node. This suggests that ht has strong linear and possibly quadratic effects. On the other hand, the splits on age and sex point to interactions between them and ht. These interactions can be interpreted with the help of Figures 6 and 7, which plot the data values of FEV and ht and the fitted regression functions, with a different symbol and colour for each node.



Notes

- **Clustering:**

Clustering is that the task of dividing the population or data points into a variety of groups such data points within the same groups are more likely to other data points within the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Let's understand this with an example. Suppose, you're the top most person of a rental store and wish to understand preferences of your costumers to rescale your business. Is it possible for you to look at details of every costumer and devise a unique business strategy for each and every one of them? Definitely not but, what you'll do is to cluster all of your costumers into ten groups based on their purchasing habits and use a separate strategy for costumers in each of these ten groups. And this can be what we call clustering.

Now, that we understand what is clustering. Let's take a look at a glance of the categories of clustering

Types of Clustering

Broadly speaking, clustering will be divided into two subgroups:

Hard Clustering: In hard clustering, each datum either belongs to a cluster completely or not. For instance, within the above example each customer is put into one group out of the ten groups.

Soft Clustering: In soft clustering, rather than putting each datum into a separate cluster, a probability or likelihood of that datum to be in those clusters is assigned. As an example, from the above scenario each costumer is assigned a probability to be in either of ten clusters of the mercantile establishment.

Types of clustering algorithms:

Since the task of clustering is subjective, the means that is used for achieving this goal are plenty. Every methodology follows a special set of rules for outlining the 'similarity' among datum. In fact, there are more than hundreds of clustering algorithms. But few of the algorithms are used popularly, let's take a look at them in detail:

Connectivity models: As the name suggests, these models are supported the notion that the datum closer in data space exhibit more similarity to each other than the datum lying farther away. These models can follow two approaches within the first approach, they begin with classifying all data points into separate clusters & then aggregating them because the distance decreases.

In the second approach, all data points are classified as one cluster and then partitioned as per the distance increases. Also, the selection of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big data sets. Examples of these models are hierarchical clustering algorithm and its variants.

- **Centroid models:** These are iterative clustering algorithms during which the notion of similarity is springs by the closeness of a datum to the centroid of the clusters. K-Means clustering algorithm could be a popular algorithm that falls into

this category. In these models, the no. of clusters required at the end must have to be mentioned beforehand, which makes it important to process prior knowledge of the data set. These models run iteratively to seek out the local optima.

- **Distribution models:** These clustering models are supported the notion of how probable is it that all datum within the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from over fitting. A preferred example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.
- **Density models:** These models supported the data space for areas of varied density of datum in the data space. It isolates various different density regions and assign the datum within these regions within the same cluster. Popular examples of density models are DBSCAN and OPTICS.

Summary

- What is regression
- Discuss about different kind of regression.
- Concept of Bayesian analysis
- Concept of clustering

Questions

1. Discuss about different kind of regression
2. Define the Bayesian analysis
3. What is Lasso regression
4. Difference between Logistic Regression & Classification tree

Notes

Unit-4.3: Data Analysis

Unit Objectives:

At the end of this unit, participants will be able to learn:

- Learn about unsupervised learning
- Design Experiments
- Active Learning.

4.3.1 Unsupervised Learning

Unsupervised learning is a machine learning algorithm to draw inferences from data sets consisting of input data without labelled responses. In different pattern recognition problems, the training data consists of a set of input vectors x without any corresponding target values. The goal of unsupervised learning problems to discover groups of similar examples within the data (clustering), or to determine how the data is distributed in the space (density estimation).

Issues with Unsupervised Learning

There are several issues such as:

- It is harder than Supervised Learning
- How do we know if results are meaningful since no answer labels are available?
- The results need external evaluation
- Define an objective function on clustering (internal evaluation)

What is the need of Unsupervised Learning despite of these issues?

- Interpret large datasets is very costly and hence we can label only a few examples manually. Example: Speech Recognition
- There may be cases where we don't know how many/what classes is the data divided into. Example: Data Mining
- We may want to use clustering to gain some insight into the structure of the data before designing a classifier.

Unsupervised Learning classified into two categories:

- Parametric Unsupervised Learning

Here sample data comes from a population which follows a probability distribution based on a fixed set of parameters. In a normal family of distributions, all members have the same shape and are parameterized by mean and standard deviation. That means if we have idea about the mean and standard deviation, and that the distribution is normal, we know the probability of any future observation. Parametric UL construct the Gaussian Mixture Models and utilize Expectation-Maximization algorithm for predicting the class of the sample in question. It is harder than the standard supervised

learning because there are no answer labels available and hence there is no correct measure of accuracy available to check the result.

- **Non-parametric Unsupervised Learning**

In non-parameterized unsupervised learning, the data is grouped into clusters, where each cluster indicate something about categories and classes present in the data. This method is basically used to model and analyze data for small sample sizes. In nonparametric models do not require the modeler to make any assumptions about the distribution of the population, and that's why sometimes it's called distribution-free method.

4.3.2 Making information through Designed Experiments

Planning, Design, & Analysis are the main three components to creating data through designed experiments.

Planning should always begin with a well-formed hypothesis.

Some of the main components we wish to form in this process are as follows:

- What is the question you want answered?
- What is the population in question?
- What are dependent and independent variables?

Analysis

When conducting an experiment, there are 3 main characteristic to consider. These three aspects of an experiment allow us to assess our population's variability.

- Randomization
- Replication
- Blocking

Randomization

The purpose of randomization is to form positive that if there's variation in outcomes that's associated with outside factors, then it is distributed across treatment teams.

Replication

When conducting an experiment, we have a tendency to ask the variability of outcomes. For instance; if I were to run a given experiment however once and that I was looking on an outcome that may have occurred due to random likelihood. The purpose here is to grasp the broad spectrum of prospects or outcomes, it's vital that we have a tendency to replicate the experiment consequently.

Statistical Power

The concept of applied mathematical power means that if our experiment concludes such that we reject the NULL hypothesis and settle for the choice hypothesis,

Notes

it's the likelihood that it would not be due to random chance. Best practice is 80% applied mathematical power.

So, to change this even further; If our hypothesis seems to be correct, what's the likelihood that we didn't get that outcome just due to random likelihood.

Blocking

Blocking is employed to assist management variability by creating treatment teams additional alike. Inside of a given cluster, we would possibly see that differences are minimal, however across alternative teams that would be much larger. One example of this might be blocking an experiment by gender.

- Blocking by variables — use above for the sake of obstruction.
- Randomized Complete Block style (RCBD) experiment

T-test

After accumulating information from our experiment, one fast and east take a look at statistical significance we would possibly run is termed as a t-test.

- Consider your hypothesis or central research question:
- NULL hypothesis — let's keep this simple, the null hypothesis is pretty much when you're wrong. For the mtcars dataset, the null hypothesis might be something like a vehicle horse power has no effect on miles per gallon.
- Alternative hypothesis — conversely the alternative hypothesis means that there was a difference. If we are able to confirm with statistical significance the impact of the independent variable on the dependent variable, we'd say that we reject the null hypothesis and choose the alternative hypothesis.
- Is this a one- or two-sided test?
- One sided test — when you are testing whether a given variable is greater than another then it's a one-sided test; if you're testing whether it's less than another... still one-sided.
- Two-sided test — when you are testing that a given variable is not equal to another, then that is two sided '>' or '<' than in a single test.
- Were your results statistically significant?
- People use the term statistically significant left and right with little consideration of what it actually means. What's that speech communication is that if we run our test and our data is suggesting a that our hypothesis is correct, statistical significance is effectively knowing that it's not likely due to random chance.
- The standard here is 95% confidence, or a less than or equal to likelihood of 5%.
- What is statistical power?
- Similar to statistical significance; given that the alternative hypothesis is true, power represents the likelihood that the null hypothesis will be rejected.
- The standard for power is 80%.

Sample Size

For a given experiment one issue to contemplate is that of the sample size. So as to hit at a needed range for this requires a handful of alternative variables including targeted statistical power & significance.

Another measure is that of impact size. Impact size represents the distinction between the average of 2 groups divided by the standard deviation of both groups combined.

The > the distance between groups < of a sample to validate it. The smaller the difference the greater the likelihood that the observed distance is only due to chance.

In order to calculate any of these values including effect size, statistical power, p-value, etc. Load up the package power and use the `pwr.anova.test` to identify the odd variable out here.

k — number of groups

n — sample size per group

f = effect size

sig.level = significance level

power = statistical power

- Making Information through Active Learning:

Discuss with a case study Algorithmically Choosing Training information to enhance Alexa's Natural-Language Understanding

Alexa's ability to retort to client requests is essentially the result of machine learning models trained on annotated information. The models are fed sample texts like "Play the blue blood song 1999" or "Play River by Joni Mitchell". In every text, labels are connected to particular words — Song Name for "1999" and "River", as an example, and Artist Name for Prince and Joni Mitchell. By analyzing annotated information, the system learns to classify unannotated information on its own.

Regularly grooming Alexa's models on new information improves their performance. However annotation is dear, so we'd prefer to annotate solely the foremost informative training examples — those which will yield the best reduction in Alexa's error rate. Selecting those examples mechanically is known as active learning.

At the annual meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), we presented a new approach to active learning that, in experiments, improved the accuracy of machine learning models by 7% to 9%, relative to training on randomly selected examples.

We compared our technique to four other active-learning strategies and showed gains across the board. Our new approach is 1% to 3.5% better than the best-performing approach previously reported. In addition to extensive testing with previously annotated data (in which the labels were suppressed to simulate unannotated data), we conducted a smaller trial with unlabeled data and human annotators and found that our results held, with improvements of 4% to 9% relative to the baseline machine learning models.

Notes

The goal of active learning is to canvass as many candidate examples as possible to find those with the most informational value. Consequently, the selection mechanism must be efficient. The classical way to select examples is to use a simple linear classifier, which assigns every word in a sentence a weight. The sum of the weights yields a score, and a score greater than zero indicates that the sentence belongs to a particular category.

For instance, if the classifier is trying to determine whether a sentence belongs to the category music, it would probably assign the word “play” a positive weight, because music requests frequently begin with the word “play”. But it might assign the word “video” a negative weight, because that’s a word that frequently denotes the customer’s desire to play a video, and the video category is distinct from the music category.

Such weights are learned from training examples. During training, the linear classifier is optimized using a loss function, which measures the distance between its performance and perfect classification of the training data.

Typically, in active learning, examples are selected for annotation if they receive scores close to zero — whether positive or negative — which means that they are near the decision boundary of the linear classifier. The hypothesis is that hard-to-classify examples are the ones that a model will profit from most.

Researchers have also investigated committee-based methods, in which linear models are learned using a number of different loss functions. Some loss functions emphasize getting the aggregate statistics right across training examples; others emphasize getting the right binary classification for any given example; still others impose particularly harsh penalties for giving the wrong answer with high confidence; and so on.

A graph showing how different loss functions (black lines) divide training data in different ways. Easily classified examples (red and green X’s) are less informative than examples that fall closer to classification boundaries (grey X’s).

Traditional committee-based methods also select low-scoring examples, but they add another criterion: at least one of the models must disagree with the others in its classification. Again, the assumption is that hard-to-classify examples will be the most informative.

In our experiments, we explored a variant on the committee-based approach. First, we tried selecting low-scoring examples on which the majority of linear models have scores greater than zero. Because this majority positive filter includes examples with all-positive scores, it yields a larger pool of candidates than the filter that enforces dissent. To select the most informative examples from that pool, we experimented with several different re-ranking strategies.

Most importantly, we used a conditional-random-field (CRF) model to do the re-ranking. Where the linear models classify requests only according to domain — such as music, weather, smart home, and so on — the CRF models classify the individual words of the request as belonging to categories such as Artist Name or Song Name.

If the CRF easily classifies the words of a request, the score increases; if the CRF struggles, the score decreases. (Again, low-scoring requests are preferentially selected

for annotation.) Adding the CRF classifier does not significantly reduce the efficiency of the algorithm because we execute the re-ranking only on examples where the majority of models agreed.

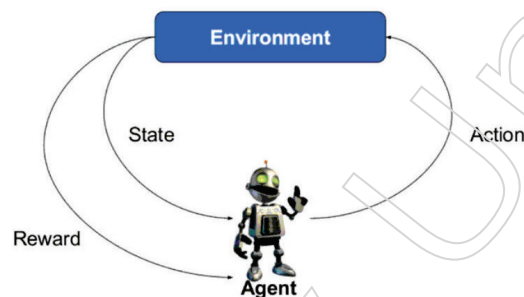
For re-ranking, we add the committee scores and then take the absolute value of the sum. This permits individual models on the committee to provide high-confidence classifications, so long as strong positive scores are offset by strong negative scores.

The committee approaches reported in the literature enforced dissent among the models; interestingly, using the criterion of majority scores greater than zero yielded better results, even without the CRF. With the CRF, however, the error rate shrank by an additional 1% to 2%.

Ref: <https://www.amazon.science/blog/active-learning-algorithmically-selecting-training-data-to-improve-alexas-natural-language-understanding>.

4.3.4 Creating Data through Reinforcement Learning

Typical RL Scenario



Here are some important terms used in Reinforcement AI:

- Agent: It is an assumed entity which performs actions in an environment to gain some reward.
- Environment (e): A scenario that an agent has to face.
- Reward (R): An immediate return given to an agent when he or she performs specific action or task.
- State (s): State refers to the current situation returned by the environment.
- Policy (π): It is a strategy which applies by the agent to decide the next action based on the current state.
- Value (V): It is expected long-term return with discount, as compared to the short-term reward.
- Value Function: It specifies the value of a state that is the total amount of reward. It is an agent which should be expected beginning from that state.
- Model of the environment: This mimics the behaviour of the environment. It helps us to make inferences to be made and also determine how the environment will behave.
- Model based methods: It is a method for solving reinforcement learning problems which use model-based methods.

Notes

- Q value or action value (Q): Q value is quite similar to value. The only difference between the two is that it takes an additional parameter as a current action.

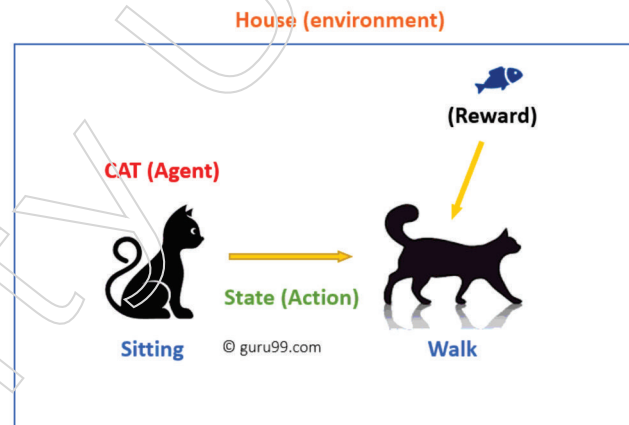
How Reinforcement Learning Works?

Let's see some simple example which helps us to illustrate the reinforcement learning mechanism.

Consider the scenario of teaching new tricks to a cat

- As cat doesn't understand English or any other human language, we can't tell her directly what to do. Instead, we follow a different strategy.
- We emulate a situation, and the cat tries to respond in many different ways. If the cat's response is the desired way, we will give her fish.
- Now whenever the cat is exposed to the same situation, the cat executes a similar action with even more enthusiastically in expectation of getting more reward(food).
- That's like learning that cat gets from "what to do" from positive experiences.
- At the same time, the cat also learns what not to do when faced with negative experiences.

Explanation about the example:



In this case,

- Our cat is an agent that's exposed to the environment. During this case, it is our house. An example of a state might be our cat sitting, and that we employed a selected word in certain cat to walk.
- Our agent reacts by performing an action transition from one "state" to a different "state."
- For example, our cat goes from sitting to walking.
- The reaction of an agent is an action, and therefore the policy could be method of selecting an action given a state in expectation of higher outcomes.
- After the transition, they may get a reward or penalty in return.

Reinforcement Learning Algorithms

There are three approaches to implement a Reinforcement Learning algorithm.

Value-Based

In a value-based Reinforcement Learning method, we ought to aim to maximize a worth function $V(s)$. During this method, the agent is expecting a long-term return of the current states under policy π .

Policy-based

In a policy-based RL method, we try and come up with such a policy that the action performed in every state helps us to achieve maximum gift within the future.

Two styles of policy-based methods are:

- Deterministic: For any state, the identical action is produced by the policy π .
- Stochastic: Every action contains a certain probability, which is decided by the subsequent equation. Stochastic Policy :

$$P\{a|s\} = P\{A, = a|S, =S\}$$

Model-Based

In this Reinforcement Learning method, we wish to create a virtual model for each and every environment. The agent learns to perform therein specific environment.

Characteristics of Reinforcement Learning

Here are important characteristics of reinforcement learning:

- There isn't any supervisor, only a true number or gift signal
- Sequential higher cognitive process
- Time plays a vital role in Reinforcement problems
- Feedback is often delayed, not instantaneous
- Agent's actions determine the subsequent data it receives

Types of Reinforcement Learning

Two varieties of reinforcement learning methods are:

Positive:

It is defined as an occurrence that happens thanks to specific behaviour. It increases the strength and also the frequency of the behaviour and impacts positively on the action taken by the agent.

This type of Reinforcement helps us to maximise performance and sustain change for a more extended period. However, an excessive amount of Reinforcement may cause over-optimization of state, which can affect the results.

Negative:

Negative Reinforcement is defined as strengthening of behaviour that occurs because of a negative condition which should have stopped or avoided. It helps us to define the minimum stand of performance. However, the downside of this method is that it provides enough to fulfil the minimum behaviour.

Notes

Learning Models of Reinforcement

There are two important learning models in reinforcement learning:

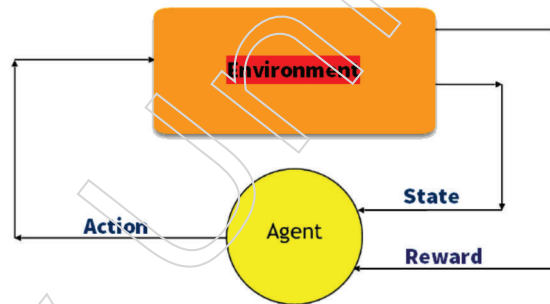
- Markov Decision Process
- Q learning

Markov Decision Process

The following parameters are used to get a solution:

- Set of actions- A
- Set of states -S
- Reward- R
- Policy- π
- Value- V

The mathematical approach for mapping a solution in reinforcement Learning is known as a Markov Decision Process or (MDP).

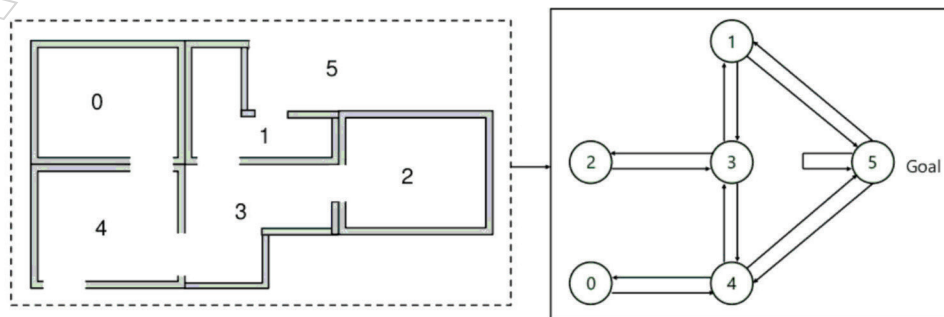


Q-Learning

Q learning may be a value-based method of supplying information to tell which action an agent should take.

Let's understand this method by subsequent example:

- There are five rooms inside a building which are connected by doors.
- Each room is numbered 0 to 4
- The outside of the building are often be one big outside area (5)
- Doors no 1 and 4 lead into the building from room 5



Next, we might wish to associate a gift value to every door:

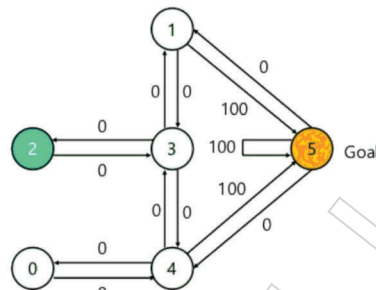
- Doors which lead on to the goal have a gift of 100
- Particular doors which is not directly connected to the target room gives zero reward
- As doors are two-way, and two arrows are assigned for every room
- Every arrow within the above image contains an intermediate gift value

Explanation

In this image, we'll view that room represents a state

Agent's movement from one room to a distinct represents an action

In the following image, a state is described as a node, while the arrows show the action.



For example, an agent traverse from room number 2 to 5

- Initial state = state 2
- State 2-> state 3
- State 3 -> state (2,1,4)
- State 4-> state (0,5,3)
- State 1-> state (5,3)
- State 0-> state 4

Summary

- Define Learning
- Discuss about unsupervised Learning.
- Creating Data through several experiments.
- Reinforcement Learning

Exercises:

1. Regression coefficient is independent of change of
 - a) . origin
 - b) Subject
 - c) Data
 - d) None of the above

Notes

2. In the case ofregression, one variable is affected by a linear combination of another variable.
 - a) simple linear
 - b) complex linear
 - c) non linear
 - d) none of the above
3.analysis is based on the statistical principle of multivariate statistics, which involves observation and analysis of more than one statistical variable at a time
 - a) Multivariate
 - b) Simple
 - c) Both a and b
 - d) None of the above
4. In discriminant analysis,groups are compared.
 - a) Three
 - b) two or more
 - c) one
 - d) none of the above
5. If the discriminant analysis involves two groups, there arecentroids
 - a) Four
 - b) Three
 - c) Two
 - d) one
6.analysis is concerned with the measurement of the joint effect of two or more attributes.
 - a) Simple
 - b) Conjoint
 - c) Complex
 - d) All of the above
7. Forselection, the market researcher can conduct interview with the customers directly.
 - a) Attributes
 - b) Sample
 - c) Item
 - d) None of the above
8. Theis a part-worth or utility for each level of each attribute

- a) Output
 - b) Input
 - c) Both a and b
 - d) None of the above
9. When the objective is to summarise information from a large set of variables into fewer factors, analysis is used.
- a) principle component factor
 - b) sub factors
 - c) data
 - d) none of the above
10. Correspondence analysis is atechnique.
- a) Descriptive/Exploratory
 - b) Precise/gist
 - c) both a or b
 - d) None of the above
11. In a typical correspondence analysis, a cross-tabulation table of frequencies is first
- a) standardized
 - b) rationalized
 - c) subdued
 - d) none of the above
12. Analysis is a technique used for classifying objects into groups.
- a) Cluster
 - b) Group
 - c) Collective
 - d) None of the above
13. Theapplication of cluster analysis is in customer segmentation and estimation of segment sizes
- a) marketing
 - b) selling
 - c) buying
 - d) none of the above
14. An advantage of the non-metric models is that they permit the researcher to andpreference data.
- a) categorize, examine

Notes

- b) identify, scrutinize
- c) collect, analyze
- d) none of the above

15. The spatial display of data provided by MDS is also sometimes referred to as

- a) perceptual mapping
- b) conceptual mapping
- c) geographical mapping
- d) none of the above

Answers:

1. origin
2. simple linear
3. Multivariate
4. two or more
5. two
6. Conjoint
7. attributes
8. output
9. principle component factor
10. descriptive/exploratory
11. standardized
12. Cluster
13. marketing
14. categorize, examine
15. perceptual mapping

Module-5: Field Project & Report Writing

Notes

Key Learning outcomes:

At the end of this module the participants will be able to:

1. Analyse prewriting considerations
2. Analyse literature review writing
3. Analyse report writing

Structure

Unit 5.1 : Prewriting Considerations

- 5.1.1 Topic
- 5.1.2 Audience
- 5.1.3 Purpose

Unit 5.2: Literature Review Writing.

- 5.2.1 What is Literature Review?
- 5.2.2 Purpose of Literature Review
- 5.2.3 Types of Literature Reviews
- 5.2.4 Structure & Writing Style
- 5.2.5 Stages of Literature Review Development
- 5.2.6 Ways to organize Literature Review
- 5.2.7 Writing Literature Review

Unit 5.3: Report Writing

- 5.3.1 Meaning of Research Report
- 5.3.2 Types of Report
- 5.3.3 Components of a Research Report
- 5.3.5 APA style essentials
- 5.3.6 Citing & Referencing Sources
- 5.3.7 Footnotes
- 5.3.8 Key Considerations/factors

Notes**Unit 5.1 : Prewriting Considerations****Unit Outcomes**

At the end of this unit, you will learn:

- How to Choose a Topic
- How to get attention of audience.
- Why Choose the topics.

5.1.1 : Topic

Prewriting is the first stage of the writing process, and it requires the writer to think about three main factors: topic, audience, and purpose.

A student may be faced with one of two types of topics: assigned topics or topics chosen by the student. If the topic has been assigned, the assignment instructions will limit and determine the approach to take. Instructions must be carefully read and instructions must be followed to the letter. If the student is given free reign over the topic, it is critical that they consider the value and significance of the final product.

A writer should choose something he is passionate about and knows a lot about, but he should also think about the effect he wants to achieve and the reaction he wants from the reader. Any topic can spark a lively debate if the following options are considered: choosing an unusual topic or taking a fresh and unique approach to an old one.

5.1.2 Audience

For communication to be effective, the audience's experience and knowledge of the subject must be considered: too technical and specialised information may be beyond the reader's comprehension; a too basic or simple approach will bore the reader.

The question to consider is: What will the reader gain by reading this essay? The goal will be to educate, entertain, or persuade. These goals are frequently combined in a paper, with each goal serving as a function of the others.

The main goal of prewriting activities is to determine the paper's focus. The point of focus is where all of your energy is focused. The paper will be vague, superficial, and likely disorganised if the topic is too broad.

5.1.3 Purpose

Consider the audience to see if the topic is narrow enough. If your audience lacks specific knowledge of the subject, you may want to take a more general approach. Our own knowledge of the subject also limits you. You can't be precise about something you don't understand. Of course, research will provide you with the necessary information on a subject.

After you've decided on a strategy, you can start gathering ideas. Remember that you can always change your paper's focus if you give yourself enough time to make the necessary changes. If you're having trouble narrowing down your topic, a prewriting activity might help.

Summary:

- What is Prewriting
- Choosing a topic
- The audience of the topics.

Questions:

- Why we choosing a topics for prewriting is a vital issue?
- Who are the main audience of different kind of topics?
- What is the purpose?

Notes

Notes

Unit-5.2: Literature Review Writing

Unit Out Comes

At the end of this unit, you will learn:

- Learn about Literature
- What is the purpose
- Structure and Writing style
- Types of Literature Review

5.2.1 What is Literature Review?

Do you know that we, human beings, are the most intelligent living beings on earth? Thanks to our stellar intelligence, we can utilise the knowledge that has been preserved or accumulated over eons. Human knowledge comprises three equally crucial phases - namely preservation, transmission, and advancement. Research helps in advancement of knowledge so that an updated knowledge reservoir is created and transmitted for the benefit of mankind.

Human beings build upon the recorded and accumulated knowledge of the past and this constant endeavour of adding to the vast reservoir of knowledge in every possible field makes advancement of human race possible. You, as a researcher, need to ensure that considerable work has already been done on topics related to your field of investigation. You are required to be familiar with all previous projects, research, and theory related to the research problem you are dealing with. You need to conduct a thorough review of research and theoretical literature to ensure such familiarity.

In this unit, you are going to study the meaning and importance of reviewing literature, and identify the sources and steps of writing review of literature.

The term “review” means “to organise the knowledge of the specific research area to create a knowledge pool so that your study adds on to and enriches the field of research.” The term “literature” stands for “the knowledge of a specific area of investigation, related to a given discipline, which includes theoretical, research-oriented, and practical studies.” Thus, review of literature is the process of creating new and updating existing knowledge pools, related to specific disciplines, which add on to and enrich fields of research.

5.2.2 Purpose of Literature Review:

A literature review has several functions. We have understood the meaning of the terms “review of literature”, it is time that we learn why and how a successful review of literature would help us to drive our research. The purpose why we choose literature review, are like as

- It provides the opportunity to show what research has already been done on any given subject.
- Review of Literature provides researchers with theories, ideas, explanations

or hypothesis that may prove useful in the formulation of a new problem

- Review of Literature guides researchers on the availability of adequate evidence that solves the problem sufficiently this initiative avoids the replication of research
- Review of Literature serves as prominent sources for hypothesis - researchers can formulate research hypotheses based on available studies
- Review of Literature suggests data sources, methodology, and statistical techniques apt for the solution of the research problem
- Review of Literature helps researchers locate comparative data and findings useful in the correct interpretation of results

5.2.3 Types of Literature Reviews:

Narrative literature review

Critiques and summarises the body of a work of literature. A narrative review can also be used to draw conclusions about a topic and identify gaps or inconsistencies in a body of knowledge. To conduct a narrative literature review, you must have a sufficiently focused research question.

Systematic literature review

In comparison to most other types of literature reviews, systematic literature review necessitates a more rigorous and well-defined approach. A systematic literature review is thorough and includes information about the timeframe in which the literature was chosen. Meta-analysis and meta-synthesis are the two types of systematic literature reviews.

When you do a meta-analysis, you combine the results of multiple studies on the same topic and analyse them using standardised statistical procedures. Patterns and relationships are identified, and conclusions are drawn, in meta-analysis. Meta-analysis is linked to a deductive research strategy.

Meta-synthesis, on the other hand, is based on techniques that aren't statistical. This method combines, evaluates, and interprets the results of several qualitative research studies. When using an inductive research approach, a meta-synthesis literature review is usually performed.

Argumentative literature review

As the name implies, an argumentative literature review selects literature to support or refute an argument, deeply ingrained assumption, or philosophical problem already established in the literature. It should be noted that one of the major drawbacks of an argumentative literature review is the possibility of bias.

Integrative literature review

An integrative literature review examines, critiques, and synthesises secondary data about a research topic in order to generate new frameworks and perspectives on the subject. Integrative literature review will be your only option if your research does not involve primary data collection and analysis.

Notes

Theoretical literature review

Theoretical literature review is concerned with a body of knowledge that has accumulated in relation to a topic, concept, theory, or phenomenon. Theoretical literature reviews are useful for determining what theories already exist, their relationships, and the extent to which existing theories have been investigated, as well as for developing new hypotheses to test.

5.2.4 Structure & Writing Style

Writing up the review

Key features to address in our literature review:

- the methods we have used to find the papers
- the papers we have read
- the criteria used to analyze the papers
- what we hoped to find in our review of the literature
- what we found from reading the literature
- what if any gaps in the literature we found
- where our research question fits in
- Style of writing

The style of writing we use is important and also needs to express:

- Themes arising from papers read rather than being a summary of each paper
- Examples of where authors agree or disagree on particular points, ideas or conclusions
- Key theories being examined and how different authors are using or applying the theories
- Thoughts on the usefulness of the literature in response to your research question

Literature review template

Here we discuss about a simple pattern for writing up a systematic literature review. This is a very simple outline, that's why be sure to discuss with your supervisor to ensure that their requirements are met and that specific elements of literature review/research are covered.

Literature review outline

Literature Review Outline

- I. Introduction
- Describe the overall topic that you have been investigating, why it is important to the field, and why you are interested in the topic.
 - Identify themes and trends in research questions, methodology, and findings. Give a “big picture” of the literature.
- II. Theme A¹
- Overview of characteristics of the theme (commonalities, differences, nuances)
 - Sub-theme – narrow but grouped findings related to the theme
 - Study 1 (Research question(s), Methods/Participants, Related Findings)
 - Study 2 (Research question(s), Methods/Participants, Related Findings)
 - Study 3 (Research question(s), Methods/Participants, Related Findings)
 - Sub-theme – narrow but grouped findings related to the theme
 - Study 4 (Research question(s), Methods/Participants, Related Findings)
 - Study 5 (Research question(s), Methods/Participants, Related Findings)
 - Study 6 (Research question(s), Methods/Participants, Related Findings)
 - Etc., etc., etc. with other findings that fit Theme A; studies can be repeated if there are multiple findings that fit under more than one theme. However, no need to re-write methods/participants in detail (just enough to remind the reader about the study).
- III. Theme B – follow a, b, c, and so on from above
- IV. Keep repeating with themes
- V. Conclusion: *An evaluation/critique of the existing literature. Write several paragraphs.*
- What are the contributions of this literature to the field?
 - What are the overall strengths?
 - What are the overall weaknesses?
 - What might be missing?
 - What are some next steps for research? The next steps should explicitly address how to “correct” for strengths, weaknesses, and gaps.

Use APA level headings 1, 3, & 4. See Creswell (page 112) or APA manual for formatting.

Review of the Literature on Girl Culture (Level 1)

Resistance (Level 3)*Overview of resistance.* (Level 4)*Resistance to teachers.* (Level 4)*Resistance as strategic.* (Level 4)*Resistance as subconscious.* (Level 4)

¹ Remember: The theme is a broad word or phrase that synthesizes a more narrow group of related findings. E.g., a theme of “Resistance” would include types of resistance, resistance to whom, resisting what, etc.

Jackson

RES 5000/6000

Appalachian State University

5.2.5 Stages of Literature Review Development:

In the following discuss about the different stages of development of literature review.

1. Concise the topic and select papers accordingly

Consider the specific area of study. Think about the field of interests.

Talk to professor, brainstorm, and read lecture notes and recent issues of periodicals in the field.

Notes

Notes

Limit the scope to a smaller topic area.

2. Search for literature

Define the source selection criteria (i.e., articles published between a specific date range, focusing on a specific geographic region, or using a specific methodology).

Using keywords, search a library database.

Reference lists of recent articles and reviews can lead to other useful papers.

Include any studies contrary to your point of view.

Read the selected articles thoroughly and evaluate them

Evaluate and synthesize the studies' findings and conclusions.

Note the following:

- Assumptions by some or most researchers.
- Methodologies, testing procedures, subjects, material tested researchers use
- Experts in the field: names/labs that are frequently referenced
- Conflicting theories, results, methodologies
- Popularity of theories and how this has/has not changed over time

4. Organize the papers by looking for patterns and by developing subtopics

Note the following:

Findings that are common/contested

Important trends in the research

The most influential theories

Move them around if

(a) they fit better, under different headings,

(b) need to establish new topic headings.

Develop headings/subheadings that reflect the major themes and patterns.

6. Write the paper

Follow the organizational structure including the headings and subheadings. Make certain that each section links logically to the one before and after. Structure every section by themes or subtopics, not by individual theorists or researchers.

Note: If each paragraph begins with a researcher's name, it might indicate that, instead of evaluating and comparing the research literature from an analytical point of view, we have simply described what research has been done.

Prioritize analysis over description.:

- For example, look at the following two passages and note that Student A merely describes the literature, whereas Student B takes a more analytical and evaluative approach by comparing and contrasting. You can also see that this evaluative approach is well signalled by linguistic markers indicating

logical connections (words such as “however,” “moreover”) and phrases such as “substantiates the claim that,” which indicate supporting evidence and Student B’s ability to synthesize knowledge.

Student A: San (2000) concludes that personal privacy in their living quarters is the most important factor in nursing home residents’ perception of their autonomy. He suggests that the physical environment in the more public spaces of the building did not have much impact on their perceptions. Neither the layout of the building nor the activities available seem to make much difference. Ram and Ramen make the claim that the need to control one’s environment is a fundamental need of life (2001), and suggest that the approach of most institutions, which is to provide total care, may be as bad as no care at all. If people have no choices or think that they have none, they become depressed.

Student B: After studying residents and staff from two intermediate care facilities in Calgary, Alberta, San (2000) came to the conclusion that except for the amount of personal privacy available to residents, the physical environment of these institutions had minimal if any effect on their perceptions of control (autonomy). However, French (1998) and Haroon (2000) found that availability of private areas is not the only aspect of the physical environment that determines residents’ autonomy. Haroon interviewed 115 residents from 32 different nursing homes known to have different levels of autonomy (2000). It was found that physical structures, such as standardized furniture, heating that could not be individually regulated, and no possession of a house key for residents limited their feelings of independence. Moreover, Hope (2002), who interviewed 225 residents from various nursing homes, substantiates the claim that characteristics of the institutional environment such as the extent of resources in the facility, as well as its location, are features which residents have indicated as being of great importance to their independence.

7. Review the work

Look at the topic sentences of each paragraph. If read only these sentences, the paper presented a clear position, logically developed, from beginning to end? The topic sentences of each paragraph should indicate the main points of literature review.

Make an outline of each section of the paper and decide whether need to add information, to delete irrelevant information, or to re-structure sections.

Read work out loud by which able to identify the need of punctuation marks to signal pauses or divisions within sentences, where have made grammatical errors, or where the sentences are unclear.

Since the purpose of a literature review is to demonstrate that the writer is familiar with the important professional literature on the chosen subject,

Make certain that all of the citations and references are correct

Text should be written in a clear and concise academic style; it should not be descriptive in nature or use the language of everyday speech.

There should be no grammatical or spelling errors.

Sentences should flow smoothly and logically.

Notes

5.2.6 Ways to organize Literature Review:

The following steps are described about the way to organize the literature review.

Chronological (by date): This is a common way for the topic that have been talked about for a long time and have changed over its history. Organise it in stages of how the topic has changed: the first definitions of it, then major time periods of change as researchers talked about it, then how it is thought about today.

Broad-To-Specific: Another approach is to start with a section for reviewing the general type of issue, then narrow down to increasingly specific issues in the literature until to reach the articles that are most specifically similar to the research question, thesis statement, hypothesis, or proposal. This can be a good way to introduce a lot of background and related facets of selected topic when there is not much directly on topic but we are tying together many related, broader articles.

Major Models Or Major Theories: If there are multiple models or prominent theories, then it is a good idea to outline the theories or models that are applied the most in our articles. That way we can group the articles we read by the theoretical framework that each prefers, to get a good overview of the prominent approaches to our concept.

Prominent Authors: If a certain researcher started a field, and there are several famous people who developed it more, a good approach can be grouping the famous author/researchers and what each is known to have said about the topic and then organise other authors into groups by which famous authors' ideas they are following.

Contrasting Schools Of Thought: If there a dominant argument comes up in our research, with researchers taking two sides and talking about how the other is wrong, then we may want to group our literature review by those schools of thought and contrast the differences in their approaches and ideas.

5.2.7 Writing Literature Review:

- Once we've found credible, up-to-date, and accurate information sources, we'll need to start putting together a picture of our subject area and the research that's been done on it.
- To get a complete picture, we look at a variety of print and electronic sources.
- Begin sorting information into categories related to our research topic.
- Take notes and identify themes, but we should continue to analyse the documents critically.
- Our literature review helped us identify the various sub-topics.
- We should describe our subject, show that we understand it, and explain what research has been done and how it will affect our own research.
- Our work should be fully referenced in order to avoid plagiarism.
- We should use quotes from other authors as needed, but we should not rely on them.
- The introduction to a literature review should be written first. This is a brief

summary of our research focus. It allows us to define our research area, highlighting any areas that we have excluded and why.

- After the introduction, present the main body of our literature review, which will make up the majority of our work.
- Finally, present our conclusions, which should summarise our findings and, hopefully, justify our choice of research topic.
- A reference list and/or a bibliography are required. This is a comprehensive list of all the sources we used and/or consulted.

Summary

- Discuss about Literature review
- Different kind of Stages of Literature Review Development
- Ways to organize Literature Review
- Types of Literature Reviews
- Writing Literature Review

Questions

1. How you can write a review
2. Describe the different kind of Literature review
3. How you can organise a literature review.

Notes

Unit-5.3: Report Writing

Unit Outcomes:

At the end of this unit, you will learn:

- Learn about what is the research report.
- Types of report
- APA style
- Key factors.

5.3.1 Meaning of Research Report

A report is a very formal document that is written for various purposes, such as 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.

There are three features that, together, characterize report writing at a very basic level:

- a predefined structure,
- independent sections
- reaching unbiased conclusions.

Predefined structure: In bigger sense, these headings may indicate sections within a report, such as an introduction, discussion, and conclusion.

Independent sections: Each section in a report is separately written, because if the reader want to selectively identify the report sections they are interested in, rather than reading the whole report through in one go from start to finish.

Unbiased conclusions: A third element of report writing is that it is an unbiased and objective form of writing.

5.3.2 Types of Report

Formal or Informal Reports:

Formal reports are meticulously structured; they emphasise objectivity and organisation, contain a great deal of detail, and are written in a style that avoids personal pronouns. Informal reports are typically short messages written in a natural, casual tone. An internal memorandum can be thought of as a non-formal report.

Short or Long Reports

This is a perplexing categorization. A one-page memo is clearly brief, while a twenty-page report is clearly lengthy. But where does the line of demarcation lie? Keep in mind that as a report grows longer (or whatever length you decide), it begins to resemble formal reports more.

Informational or Analytical Reports

Annual reports, monthly financial reports, and reports on employee absenteeism are examples of informational reports that carry objective data from one part of an organisation to another. Scientific research, feasibility reports, and real-estate appraisals are examples of analytical reports that attempt to solve problems.

Proposal Report

The proposal is a problem-solving report with a twist. A proposal is a written document that explains how one company can meet the needs of another. The majority of government agencies use “requests for proposal,” or RFPs, to publicise their requirements. The RFP identifies a requirement, and potential suppliers submit proposal reports outlining how they will meet that requirement.

Vertical or Lateral Reports

The direction in which a report travels is classified in this way. Vertical reports are reports that are more upward or downward in the hierarchy, and they help with management control. Lateral reports, on the other hand, help with organisation coordination. A lateral report is one that travels between units at the same organisational level (for example, the production and finance departments).

Internal or External Reports

Internal reports circulate within the company. External reports, such as company annual reports, are written for distribution outside of the organisation.

Periodic Reports

Periodic reports are sent out on a regular basis. They are usually upwardly directed and serve to control management. The uniformity of periodic reports is aided by pre-printed forms and computer-generated data.

Functional Reports

Accounting reports, marketing reports, financial reports, and a variety of other reports that are classified based on their intended use are included in this category. Almost all reports fit into at least one of these categories. A single report could be classified under several headings.

5.3.3 Components of a Research Report

Abstract or Summary

Summary or Abstract The abstract or summary informs the reader of the paper's main points and findings in a concise manner. This allows the reader to determine whether or not the paper will be of interest to them. When looking for papers that are relevant to your research, get into the habit of reading only the abstracts. Only read the body of a paper if you believe it will be of use to you.

Introduction

The introduction informs the reader about the paper's general topic, why it is important, and what to expect in the body of the paper. Introductions should flow

Notes

from broad concepts to the paper's specific topic. In some cases, introductions are incorporated into literature reviews.

Review of Literature

The literature review informs the reader about what other researchers have discovered about the topic of the paper or about other relevant research.

A literature review should influence how readers think about a topic by informing them about what the academic community has to say about it and its related issues.

Often, what students refer to as a "research paper" is nothing more than a literature review.

It states facts and ideas about the social world along the way, and it backs them up with credit for where they came from. The literature review makes it clear that the author is speculating if an idea cannot be substantiated by the community of scholars, and the logic of the speculation is detailed. Information that isn't relevant isn't discussed.

The literature review has its own distinct personality. The information sources aren't heavily quoted or "copied and pasted." Instead, the author rewrites facts and ideas in his or her own words, citing the source of the information. Consider how you tell your family about the exciting things you've learned in class... Consider how you talk about sociology at cocktail parties. You claim things in your own words... You don't copy and paste or quote word for word.

Research Methodology:

This is the most important section of the report, as it contains all of the crucial information. Readers can gain information about the topic while also evaluating the quality of the content provided, and the research can be approved by other market researchers. As a result, this section must be extremely informative, with each aspect of the research thoroughly discussed. Information must be presented in a chronological order based on its priority and significance. Researchers should include references if they used existing techniques to gather information.

Research Results:

This section of the results will include a brief description of the findings as well as the calculations used to achieve the goal. The exposition that follows data analysis is usually done in the report's discussion section.

Research Discussion:

In this section, the findings are discussed in great detail, as well as a comparison of reports that may or may not exist in the same domain. In the discussion section, any anomaly discovered during research will be discussed. When writing research reports, the researcher must connect the dots to show how the findings can be applied in the real world.

Research References and Conclusion:

Finish by summarising all of the research findings and mentioning each and every author, article, or other piece of content from which references were taken.

5.3.5 APA style Essentials

The American Psychological Association (APA) is a professional organisation of psychologists. You'll find answers to questions like "What is APA format?" in this guide. In terms of writing and organising your paper according to the standards of the American Psychological Association our APA citation page has instructions on how to properly cite sources. Our guide was based on the official American Psychological Association handbook, and we've included page numbers from it throughout. This page, however, is not affiliated with the organisation.

If your paper is about science, you'll almost certainly use the APA format. The standards and guidelines of this organisation are used by many behavioural and social sciences.

General Document Guidelines:

APA Style Essentials

http://www.vanguard.edu/uploaded/research/apa_style_guide/apastyleessentials.pdf

Last modified November 7, 2017

Douglas Degelman, PhD

Vanguard University of Southern California

Edition 7th

Margins: One inch on all sides (top, bottom, left, right)

Font Size and Type: 12-pt. Times New Roman font

Line Spacing: Double-space throughout the paper, including the title page and references. Spacing after Punctuation: Space once after commas, colons, and semicolons within sentences. Insert two spaces after punctuation marks that end sentences.

Alignment: Flush left (creating uneven right margin). Indent the first line of each paragraph
Pagination: The page number appears at the top right of every page. Title page is page 1
Running Head: The running head is a short title that appears at the top left of the pages of a paper or published article. The running head should not exceed 50 characters, including punctuation and spacing. Using most word processors, the running head and page number can be inserted into a header (in Microsoft word go to View - Header Footer), which then automatically appears on all pages.

Order of Pages:

Title Page (page 1) includes:

Running Head: typed flush left (all uppercase) following "Running head:"

Centered on the page: Paper title, author, course name, professors name, date.

Body The body of the paper begins on a new page. Subsections of the body of the paper do not begin on new pages. The body of the paper includes:

Notes

Title: The title of the paper (in uppercase and lowercase letters) is centered at the top of the first page of the body of the paper (so before the introduction paragraph) on the first line below the running head.

Introduction: The introduction paragraph (which is not labelled with the word Introduction) begins on the line following the paper title.

Headings: Headings may be used to help organize the paper. Main headings would use Level 1 (centered, boldface), and subheadings would use Level 2 (flush left, boldface). Text citations: In text citations should be included throughout the body of the paper

Reference Page All sources included in the References section must be cited in the body of the paper (and all sources cited in the paper must be included in the References section).

Pagination: The References section begins on a new page.

Heading: The word References (centered on the first line below the running head)

Format: The references (with hanging indent - meaning the first line is not indented but the following lines are) begin on the line following the References heading. Entries are organized alphabetically by last names of first authors.

In-Text citations:

Source material must be documented in the body of the paper by citing the author(s) and date(s) of the sources. The underlying principle is that ideas and words of others must be formally acknowledged. The reader can obtain the full source citation from the list of references that follows the body of the paper.

- When the names of the authors of a source are part of the formal structure of the sentence, the year of publication appears in parentheses following the identification of the authors. Consider the following example: Wirth and Mitchell (1994) found that although there was a reduction in insulin dosage over a period of two weeks in the treatment condition compared to the control condition, the difference was not statistically significant. [Note: here, and is used when multiple authors are identified as part of the formal structure of the sentence.]
- When the authors of a source are not part of the formal structure of the sentence, both the authors and year of publication appear in parentheses. Consider the following example: Reviews of research on religion and health have concluded that at least some types of religious behaviours are related to higher levels of physical and mental health (Gartner, Larson, & Allen, 1991; Koenig, 1990; Levin & Vanderpool, 1991; Maton & Pargament, 1987; Paloma & Pendleton, 1991; Payne, Bergin, Bielema, & Jenkins, 1991). [Note: & is used when multiple authors are identified in parenthetical material. Note also that when several sources are cited parenthetically, they are ordered alphabetically by first authors' surnames and separated by semicolons.]
- When a source that has two authors is cited, both authors are included every time the source is cited.
- When a source that has three, four, or five authors is cited, all authors are

included the first time the source is cited. (Payne, Bergin, Bielema, & Jenkins, 1991). When that source is cited again, the first author's surname and "et al." are used. Payne et al. (1991) showed that ... When a source that has six or more authors is cited, the first author's surname and "et al." are used every time the source is cited (including the first-time).

- Every effort should be made to cite only sources that you have actually read. When it is necessary to cite a source that you have not read ("Grayson" in the following example) that is cited in a source that you have read ("Murzynski&Degelman" in the following example), use the following format for the text citation and list only the source you have read in the References list: Grayson (as cited in Murzynski&Degelman, 1996) identified four components of body language that were related to judgments of vulnerability.
- To cite a personal communication (including letters, emails, and telephone interviews), include initials, surname, and as exact a date as possible. Because a personal communication is not "recoverable" information, it is not included in the References section. For the text citation, use the following format: B. F. Skinner (personal communication, February 12, 1978) claimed...
- To cite a Web document, use the author-date format. If no author is identified, use the first few words of the title in place of the author. If no date is provided, use "n.d." in place of the date. Consider the following examples: Degelman (2009) summarizes guidelines for the use of APA writing style. Changes in Americans' views of gender status differences have been documented (Gender and Society, n.d.).
- To cite the Bible, provide the book, chapter, and verse. The first time the Bible is cited in the text, identify the version used. Consider the following example: "You are forgiving and good, O Lord, abounding in love to all who call to you" (Psalm 86:5, New International Version). [Note: No entry in the References list is needed for the Bible.]

Quotations: It is almost always better to put information into your own words, but when a direct quotation is used, always include the author, year, and page number as part of the citation.

A quotation of fewer than 40 words should be enclosed in double quotation marks and should be incorporated into the formal structure of the sentence. Consider the following example:

Patients receiving prayer "required less diuretic and antibiotic therapy, had fewer episodes of pneumonia, had fewer cardiac arrests, and were less frequently intubated and ventilated" (Byrd, 1988, p. 829).

A lengthier quotation of 40 or more words should appear (without quotation marks) apart from the surrounding text, in block format, with each line indented five spaces from the left margin.

Formatting References on the Reference Page:

Most reference entries have the following components:

- **Authors:** Authors are listed in the same order as specified in the source, using each author's last name and first initial. Commas separate all authors. If no

Notes

author is identified, the title of the document begins the reference.

- Year of Publication: In parentheses following authors, with a period following the closing parenthesis. If no publication date is identified, use n.d.
- Source Reference: Includes title, journal, volume, pages (for journal article) or title, city of publication, publisher (for book). Italicize titles of books, titles of periodicals, and periodical volume numbers.
- Electronic Retrieval Information: Electronic retrieval information may include digital object identifiers (DOIs) or uniform resource locators (URLs). DOIs are unique alphanumeric identifiers that lead users to digital source material. To learn if an article has been assigned a DOI, go to <http://www.crossref.org/guestquery/> HYPERLINK "http://www.crossref.org/guestquery/".

Examples of sources (Note: On the reference page you do not write the type of source as is done below in bold - it is just listed here to show you different types of references. Also remember that on the reference page, references need to be double spaced.)

Book:

Paloutzian, R. F. (1996). *Invitation to the psychology of religion* (2nd ed.). Boston, MA: Allyn and Bacon.

Journal article with DOI:

Murzynski, J., & Degeiman, D. (1996). Body language of women and judgments of vulnerability to sexual assault. *Journal of Applied Social Psychology*, 26, 1617-1626. doi: HYPERLINK "http://dx.doi.org/10.1111/j.1559-1816.1996.tb00088.x"10.1111/j.1559-1816.1996.tb00088.x

Journal article without DOI, print version:

Koenig, H. G. (1990). Research on religion and mental health in later life: A review and commentary. *Journal of Geriatric Psychiatry*, 23, 23-53.

Journal article without DOI, retrieved online [Note: For articles retrieved from databases, include the URL of the journal home page. Database information is not needed. Do not include the date of retrieval.]

Aldridge, D. (1991). Spirituality, healing and medicine. *British Journal of General Practice*, 41, 425-427. Retrieved from <http://www.rcgp.org.uk/publications/bjgp.aspx>

Informally published Web document:

Degelman, D. (2009). APA style essentials. Retrieved from http://www.vanguard.edu/faculty/ddegelman/detail.aspx?doc_id=796

Informally published Web document (no date):

Nielsen, M. E. (n.d.). Notable people in psychology of religion. Retrieved from <http://www.psywww.com/psyrelig/psyrelpr.htm>

Informally published Web document (no author, no date):

Gender and society.(n.d.). Retrieved from <http://www.trinity.edu/~mkearl/gender.html>

For more information on citing other source try visiting <http://www.mhc.ab.ca/library/howtoguides.html>

Notes

5.3.6 Citing & Referencing Sources

According to the Harvard system uses the author's name and data of publication to identify cited documents within the text. For example:

- It has been shown that... (Saunders, 1993).
- When referring generally to work by different authors on the subject, place the authors in alphabetical order: (Baker, 1991; Lewis, 1991; Thornhill, 1993).
- When referring to dual authors: (Saunders and Cooper, 1993).
- When there are more than two authors: (Bryce et al., 1991).
- For corporate authors, for instance a company report: (Hanson Trust Plc., 1990).
- For publications with no obvious author; for example, an employment gazette: (Employment Gazette, 1993).

Referencing in the Text

When using footnotes, a number shows references within the research report. For example: 'Recent research indicates that...' This number refers directly to the references.

These list the referenced publications sequentially in the order they are referred to in our research report. This can be useful as it enables us to include comments and footnotes as well as references.

- The layout of individual references in the bibliography is the same as that for the Harvard system.
- If you find that you refer to the same item more than once you can use standard bibliographic abbreviations to save repeating the references in full.
- The publications referred to only include those you have cited in your report. They should therefore be headed 'References' rather than 'Bibliography' as shown below

Abbreviation	Explanation
Op. cit. (opereciato)	Meaning, in the work cited. This refers to a work previously referenced and so you must give the author and date and if necessary the page number, like: Robson (1993) op. cit. pp. 23-4.
Loc. Cit. (loco ciato)	Meaning, in the place cited. This refers to the same page of a work previously referenced. So you must give the author and date, like: Robson (1993) loc. Cit.
Ibid. (ibidem)	Meaning, the same work given immediately before. This refers to the work referenced immediately before and replaces all details of the previous reference other than a page number if necessary.

Notes

5.3.7 Footnotes

Researchers must insert footnotes in the appropriate places. These fulfil two purposes:

- The proper identification of materials used in quotations in the report.
- The footnotes provide supplementary value to the main body of the text. Based on the footnotes' description, one can easily refer the cross references, citation of authorities and sources, acknowledgement and elucidation or explanation of a point of view. The recent trend is to avoid footnotes. Some people feel that they enhance display of the scholarship of the researchers. But it is neither an end nor a means of displaying scholarship.

5.3.8 Key Considerations/factors

Using proper report form (for longer reports): the "form" filter

Report cover

- Title page (Includes: title of report, for whom the report is prepared, by whom it is prepared, release date. If the title does not contain the recommendation, it normally indicates what problem the report tries to solve: Ways to Market Communication Consulting Services)
- Table of Contents (List headings exactly as they appear in the body of the report, along with page numbers.)
- List of Illustrations (Tables are numbered independently from figures (pie charts, bar charts, drawings, etc.))
- Executive Summary (A good summary can be understood by itself. It summarizes the recommendation of the report, reasons for the recommendation or describes the topics the report discusses and indicates the depth of the discussion.)
- Introduction (Orients the reader.) Usually has subheadings for Purpose (identifies the problem the report addresses and if its purpose is to inform, to recommend, etc.) and Scope (identifies how broad an area the report covers if a company is losing money on its line of radios, does your report investigate the quality of the radios? The advertising campaign? The cost of manufacturing? The demand for radios?). Depending on the situation, may also have: Limitations (problems or factors that limit the validity of your recommendations), Assumptions (statements whose truth you assume and which you use to prove your final point), Methods (an explanation of how you gathered your data), Criteria (factors you used to weigh in the decision), Definitions (if you have terms to define), Background/History of the Problem (Serves as a record for later readers of the report. For most of your cases, this will not be necessary. However, in business reports, this is often a useful component of a longer, formal report.)
- Body (Presents and interprets information in words and visuals. Analyzes causes of the problem and evaluates possible solutions Specific headings will depend on the topic of the report. Uses a logical organizational format to present information)

- Conclusions (Summarizes main points of report. The most widely read part of reports. No new information should be included in the Conclusions. Conclusions are usually presented in paragraphs. but you could also use a numbered or bulleted list.)
- Recommendations (Recommends actions to solve the problem. May be combined with Conclusions, may be put at beginning of body rather than at the end (for direct order). Number the recommendations to make it easy for people to discuss them. If they seem difficult or controversial give a brief paragraph of rationale after each recommendation. The recommendations will also be in the Exec. Summary.)
- References (Document sources cited in the report. Use appropriate form for citations)
- Appendixes (Provide additional materials that the reader may want copies of questionnaires, interviews, computer printouts, previous reports, etc. Number and title them for example, Appendix A Copy of Survey. Appendix B: Sample Breakfast Menu Board, etc.)

Introduction: Explain clearly the decision problem and research objective. The background information should be provided on the product and services provided by the organisation which is under study.

Methodology: How we collect the data is described in this section.

For example, was primary data collected or secondary data used? Was a questionnaire used? What was the sample size and sampling plan and method of analysis? Was the design exploratory or conclusive?

Limitations: Every report will have some limitation such as time, geographical area, the methodology adopted, correctness of the responses, etc.

Analysis and interpretations: collected data will be tabulated. Statistical tools if any will be applied to make analysis and to take decisions.

Conclusion and Recommendation:

- (a) What was the conclusion drawn from the study?
- (b) Based on the study, what recommendation do you make?

Bibliography: If the report are based on secondary data, use a bibliography section to list the publications or sources that you have consulted. The bibliography should include, title of the book, name of the journal in case of article, volume number, page number, edition, etc.

Appendix: In an appendix used to provide a place for material which is not absolutely essential to the body of the report. The appendix will contain copies of data collection forms called questionnaires, details of the annual report of the company, details of graphs/charts, photographs, CDs, interviewers' instructions. Following are the items to be placed in this section.

- Data collection forms
- Project related paper cuttings
- Pictures and diagrams related to project

Notes

- Any other relevant things.

Summary

- Types of Report
- Discuss about Correlation & Regression
- Meaning of Research Report
- The Usefulness APA style essentials
- Key Considerations/factors

Questions

1. How many types of report are there discuss about them.
2. Which are the key factors for a report writing?
3. Discuss Key features of the report writing.

Exercises:

1. The research report will differ based on theof the particular managers using the report.
 - a) need
 - b) position
 - c) designation
 - d) none of the above
2. Accuracy refers to the degree to which information reflects.....
 - a) reality
 - b) light
 - c) unreality
 - d) none of the above
3. Availability refers to the communication process between researcher and the.....
 - a) decision maker
 - b) trainees
 - c) other researchers
 - d) none of the above
4.refers to the time span between completion of the research project and presentation of the research report to management
 - a) Currency
 - b) custom
 - c) taxation

- d) none of the above
5.is regarded as a major component of the research study
- a) Research report
 - b) final report
 - c) formal report
 - d) none of the above
6. Writing of report is thestep in a research study and requires a set of skills somewhat different from those called for in respect of the former stages of research.
- a) final
 - b) semifinal
 - c) primary
 - d) none of the above
7.means bringing out the meaning of data.
- a) Interpretation
 - b) translation
 - c) transformation
 - d) none of the above
8. Successful interpretation depends on how well the data is.....
- a) analysed
 - b) collected
 - c) interpreted
 - d) none of the above
9. In themethod, one starts from observed data and then generalisation is done
- a) induction
 - b) conduction
 - c) coronation
 - d) invention
10. In an oral presentation,plays a big role.
- a) communication
 - b) presentation
 - c) visual effects
 - d) none of the above
11.report presents the outcome of the research in detail.
- a) Long

Notes

- b) short
 - c) medium
 - d) none of the above
12. Thestatement should explain the nature of the project, how it came about and what was attempted.
- a) opening
 - b) Closing
 - c) Starting
 - d) ending
13. Theshould indicate the various parts or sections of the report.
- a) table of contents
 - b) chair of contents
 - c) stool of contents
 - d) none of the above
14.Page should indicate the topic on which the report is prepared.
- a) Title
 - b) introduction
 - c) conclusion
 - d) none of the above
15. A selected bibliography lists the items which the author thinks are of interest to the reader.
- a) primary
 - b) secondary
 - c) no
 - d) none of the above
16. In a report there must bein margins and spacing.
- a) Consistency
 - b) inconsistent
 - c) aligned
 - d) none of the above
17. Aim must be logical andin the report presentation
- a) systematic
 - b) unsystematic
 - c) illogical
 - d) none of the above

Answers:

1. need
2. reality
3. decision maker
4. Currency
5. Research report
6. final
7. Interpretation
8. analysed
9. induction
10. communication
11. Long
12. opening
13. table of contents
14. Title
15. primary
16. Consistency
17. systematic

Notes