மனோன்மணியம் சுந்தரனார் பல்கலைக்கழகம்

Manonmaniam Sundaranar University

Reaccredited with 'A' Grade (CGPA 3.13 Out of 4.0) by NAAC (3rd Cycle)

Tirunelveli - 627 012, Tamilnadu, India.

DIRECTORATE OF DISTANCE

& CONTINUING EDUCATION

RESEARCH METHODOLOGY

Prepared by

Dr.M.NISHANTHI M.Com., M. Phil., M.B.A., Ph.D.,

Assistant Professor (T) Department of Commerce

Manonmaniam Sundaranar University, Tirunelveli.

RESEARCH METHODOLOGY

Unit	Details
I	**Introduction to Research Methodology** Research: Definition – Objectives – Motivations for research – Types of research – Maintaining objectivity in research – Criteria of good research – Applications of research in business – Formulating a research problem – Literature Review – Reasons for review – Reference management tools - Identification of research gap – Framing of objectives.
II	**Hypothesis Testing and Research Design** Hypothesis – Formulation of hypothesis – Testing of hypothesis – Type I and Type II errors – Research design – Types of research design - Methods of data collection: Census, Sample survey, Case study – Sampling: Steps in sampling design, Methods of sampling – Testing of reliability and validity – Sampling errors.
III	**Data Collection** Variable**:** Meaning and types - Techniques of data collection – Primary data: Meaning, Advantages and limitations – Techniques: Interview, Schedule, Questionnaire, Observation – Secondary Data: Meaning and sources.
IV	**Data Analysis** Data Analysis – Uni-variate Analysis: Percentile, Mean, Median, Mode, Standard 64 deviation, Range, Minimum, Maximum, Independent sample t-test – Bi-variate analysis: Simple correlation, Simple Regression, Chi-square, Paired samples t-test, ANOVA, ManWhitney test – Wilcoxon signed rank test – Kruskal Wallis test (Simple problems) Multi Variate Analysis: Multiple Correlation, Multiple Regression, Factor Analysis, Friedman's test, Cluster analysis, Confirmatory Factor Analysis (CFA), Structural Equation Modelling (SEM), Multiple Discriminant Analysis.
V	**Preparation of Research Report** Report preparation – Guidelines and precautions for interpretation – Steps in Report writing - Style of research reports (APA, MLA, Anderson, Harvard) – Mechanics of report writing –Ethics in Research – Avoiding plagiarism – Plagiarism checker tools – Funding agencies for business research.

Text Books
Tripathi, (2014) "Research Methodology in Management and Social Sciences". SultanChand & Sons, New Delhi.
Kothari C.R and Gaurav Garg, (2020) "Research Methodology" – Methods and Techniques. New Age International (P) Limited, New Delhi.
Krishnaswami and Ranganathan, (2011) "Methodology of Research in Social Sciences",Himalaya Publishing House, Mumbai.

INDEX

UNIT – I

INTRODUCTION TO RESEARCH METHODOLOGY

MEANING:

Research is a very general term for an activity that involves finding out, in a more or less systematic way, things you did not know. A more academic interpretation is that research involves finding out about things that no-one else knew either. It is about advancing the frontiers of knowledge. Research is an academic activity and as such the term should be used in a technical sense. According to Clifford Woody research comprises defining and redefining problems, formulating hypothesis or suggested solutions; collecting, organizing and evaluating data; making deductions and reaching conclusions; and at last carefully testing the conclusions to determine whether they fit the formulating hypothesis.

DEFINITION:

"Research" may be defined as the systematic and objective analyze and recording of controlled observation that may lead to the developments or generalizations, principles or theories, resulting in prediction and possibility ultimate control of events". Sometimes research is defined as a movement, a movement from the known to the unknown. It is an effort to discover something. Some people say that research is a on effort to know "more and more about less and less". According to CLIFFORD WOODY, research comprises, defining and redefining problems formulating hypothesis or suggested solutions; collecting organizing and evaluating data making deductions and reaching conclusions; and at as carefully testing the conclusions to determine whether they fit the formulating a hypothesis. Research may also be defined." Any organized enquiry discussed and carried out to provide information for solving a problem".

OBJECTIVES OF RESEARCH:

Research is a conscious approach to find out the truth which is hidden and which has not been discovered by applying scientific procedure.

It develops Focus:

The research may be to understand for become familiar with some phenomena or to get to know more in depth it. For example, since the days of steam engine, the research continued to come up with more powerful locomotive which could be operated with alternative sources of energy like diesel, electricity etc.

It reveals characteristics:

To clearly reveal the characteristics of an individual or a situation or a group like a society is another type of research objective. For example in these days before a criminal is sentenced efforts are taken to study why he had turned criminal. This helps develops an approach to create opportunities for criminals to change themselves and join the main stream of life

It determines frequency of occurrence:

To determine the frequency with which something occurs or with which it associated with something else. In social research one of the major areas of repeated and continuous research is analysis of poverty and unemployment.

It tests hypothesis:

To test a hypothesis about the casual relationship between variable being studied. This type of research is mainly to determine the relationship between various factors so that necessary policy options could be framed. For example, the reasons for several malpractices adopted in public distribution outlets include low salary and absence of regulation of service of the staff in such outlets. This is turn make them to feel insecure and they resort to mal practices. Having found this the Govt., had taken a policy to improve the salary structure of these staff ad regularize their services. Hence the study of casual relationship might help in formulation the research should be honest in reporting the facts and revealing the flaws in the work.

MOTIVATION FOR RESEARCH:

Motivation for research can stem from a variety of sources and can be influenced by both personal and professional factors. Here are some common motivations:

Curiosity and Intellectual Interest: A desire to learn more about a specific topic or problem can drive individuals to conduct research. This intellectual curiosity often leads researchers to explore new ideas and expand their understanding.

Problem Solving: Research is often motivated by the need to address and solve specific problems or challenges. This can include practical issues in areas like medicine, engineering, or social sciences, where finding solutions can have significant impacts.

Advancement of Knowledge: The goal of contributing to the body of knowledge in a particular field can be a strong motivator. Researchers often seek to build on existing theories, discover new insights, and advance their discipline.

Professional Development: For many, research is a key aspect of career growth and professional development. Publishing research findings, obtaining grants, and contributing to academic or industry communities can enhance one's professional standing and open up new opportunities.

Societal Impact: The potential to make a positive impact on society or improve people's lives can drive research efforts. This motivation is often found in fields like public health, environmental science, and social work, where research can lead to beneficial changes.

Innovation and Discovery: The desire to create something new or innovative can be a powerful motivator. Research often leads to the development of new technologies, products, or methods that can revolutionize industries or improve daily life.

Personal Fulfillment: For some, the process of researching and discovering new information provides personal satisfaction and fulfillment. The joy of uncovering new findings or solving complex problems can be a strong intrinsic motivator.

Funding and Resources: Access to research funding or resources can also drive research efforts. Competitive grants, sponsorships, and institutional support can provide the means and incentives to pursue specific research projects.

Academic and Institutional Goals: Researchers may be motivated by the goals and priorities of their academic institutions or research organizations, which often set agendas or focus areas that align with broader strategic interests.Overall, the motivations for research are diverse and can be influenced by a combination of personal interests, professional aspirations, and societal needs.

TYPES OF RESEARCH:

Research can be classified into various categories depending on the perspective under which the research activity is initiated and conducted. The categorization depends on the following perspectives in general:

- Application of research study
- Objectives in undertaking the research
- Inquiry mode employed for research

1. Classification based on Application:

a. Pure / Basic / Fundamental Research:

As the term suggests a research activity taken up to look into some aspects of a problem or an issue for the first time is termed as basic or pure. It involves developing and testing theories and hypotheses that are intellectually challenging to the researcher but may or may not have practical application at the present time or in the future. The knowledge produced through pure research is sought in order to add to the existing body of research methods. Pure research is theoretical but has a universal nature. It is more focused on creating scientific knowledge and predictions for further studies.

b. Applied / Decisional Research:

Applied research is done on the basis of pure or fundamental research to solve specific, practical questions; for policy formulation, administration and understanding of a phenomenon. It can be exploratory, but is usually descriptive. The purpose of doing such research is to find solutions to an immediate issue, solving a particular problem, developing new technology and look into future advancements etc. This involves forecasting and assumes that the variables shall not change.

Key Differences between Basic and Applied Research

a) Basic Research can be explained as research that tries to expand the already existing scientific knowledge base. On the contrary, applied research is used to mean the scientific study that is helpful in solving real-life problems.

b) While basic research is purely theoretical, applied research has a practical approach.

c) The applicability of basic research is greater than the applied research, in the sense that the former is universally applicable whereas the latter can be applied only to the specific problem, for which it was carried out.

d) The primary concern of the basic research is to develop scientific knowledge and predictions. On the other hand, applied research stresses on the development of technology and technique with the help of basic science.

e) The fundamental goal of the basic research is to add some knowledge to the already existing one. Conversely, applied research is directed towards finding a solution to the problem under consideration.

2. Classification based on Objectives:

a. Descriptive Research:

This attempts to explain a situation, problem, phenomenon, service or programme, or provides information viz. living condition of a community, or describes attitudes towards an issue butthis is done systematically. It is used to answer questions of who, what, when, where, and how associated with a particular research question or problem. This type of research makes an attempt to collect any information that can be expressed in quantifiable terms that can be used to statistically analyze a target audience or a particular subject. Descriptive research is used to observe and describe a research subject or problem without influencing or manipulating the variables in any way. Thus, such studies are usually correlation or observational. This type of research is conclusive in nature, rather than inquisitive. E.g. explaining details of budget allocation changes to departmental heads in a meeting to assure clarity and understanding for reasons to bring in a change.

b. Co relational Research:

This 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. This is undertaken to discover or establish the existence of a relationship/ interdependence between two or more aspects of a situation. For example, the mind can memorize the bell of an ice cream seller or sugar candy vendor. Louder the bell sound, closer is the vendor to us. We draw this inference based on our memory and the taste of these delicious food items. This is specifically what co relational research is, establishing a relationship between two variables, ―bell sound‖ and ―distance of the vendor‖ in this particular example. Co relational research is looking for variables that seem to interact with each other so that when you see one variable changing, you have a fair idea how the other variable will change.

c. Explanatory:

It is the research whose primary purpose is to explain why events occur, to build, elaborate, extend or test a theory. It is more concerned with showcasing, explaining and presenting what

we already have. It is the process of turning over 100 rocks to find perhaps 1 or 2 precious gemstones. Explanatory survey research may look into the factors that contribute to customer satisfaction and determine the relative weight of each factor, or seek to model the variables that lead to people shifting to departmental stores from small shops from where they have been making purchases till now. An exploratory survey posted to a social networking site may uncover the fact that an organization‘s customers are unhappy thus helping the organization take up necessary corrective measures.

d. Exploratory Research:

Exploration has been the human kind‘s passion since the time immemorial. Looking out for new things, new destinations, new food, and new cultures has been the basis of most tourist and travel journeys. In the subjective terms exploratory research is conducted to find a solution for a problem that has not been studied more clearly, intended to establish priorities, develop operational definitions and improve the final research design. Exploratory research helps determine the best research design, data-collection method and selection of subjects. For such a research, a researcher starts with a general idea and uses this research as a medium to identify issues that can be the hub for future research. An important aspect here is that the researcher should be willing to change his/her direction subject to the revelation of new data or insight. Such a research is usually carried out when the problem is at a beginning stage. It is often referred to as grounded theory approach or interpretive research as it used to answer questions like what, why and how. For example: a fast food outlet owner feels that increasing the variety of snacks will enable increase in sales, however he is not sure and needs more information. Thus the owner starts studying local competition, talks to the existing customers, friends etc to find out what are their views about the current menu and what else do they wish to be included in the menu and also assess whether he would be able to generate higher revenues.

3. Classification based on Inquiry Mode:

a. Structured approach:

The structured approach to inquiry is usually classified as quantitative research. Here everything that forms the research process- objectives, design, sample, and the questions that you plan to ask of respondents- is predetermined. It is more appropriate to determine the extent of a problem, issue or phenomenon by quantifying the variation e.g. how many people have a particular problem? How many people hold a particular attitude? E.g. asking a guest to give feedback about the dishes served in a restaurant.

b. Unstructured approach: The unstructured approach to inquiry is usually classified as qualitative research. This approach allows flexibility in all aspects of the research process. It is more appropriate to explore the nature of a problem, issue or phenomenon without quantifying it. Main objective is to describe the variation in a phenomenon, situation or attitude e.g., description of an observed situation, the historical enumeration of events, an account of different opinions different people have about an issue, description of working condition in a particular industry. E.g. when guest is complaining about the room not being

comfortable and is demanding a discount the staff has to verify the claims empathically. In many studies you have to combine both qualitative and quantitative approaches. For example, suppose you have to find the types of cuisine / accommodation available in a city and the extent of their popularity. Types of cuisine are the qualitative aspect of the study as finding out about them entails description of the culture and cuisine. The extent of their popularity is the quantitative aspect as it involves estimating the number of people who visit restaurant serving such cuisine and calculating the other indicators that reflect the extent of popularity.

4. Other Types of Research:

(i) Descriptive v/s Analytical:

Descriptive research includes surveys and fact finding enquiries of different kinds. The major purpose of descriptive research is description of the state of affairs as it exists at any given time. The term Ex post facto research is used in social sciences and business research for descriptive research studies. The researcher only reports about the factors identified and cannot modify the details available thus it makes it clear that he does not have any control over such variables Most ex post facto research projects are used for descriptive studies in which the researcher strives to find out information about, for example, frequency of dining out, preferences of individuals, etc. Ex post facto studies also include attempts by researchers to discover causes even when they cannot control the variables. The methods of research utilized in descriptive research are survey methods of all kinds, including comparative and co relational methods. In analytical research, on the other hand, the researcher has to use facts or information already available, and analyze these to make a critical evaluation of the material.

(ii) Applied v/s Fundamental:

Research can either be applied (or action) research or fundamental (to basic or pure) research. Applied research aims at finding a solution for an immediate problem facing a society or an industrial/business organization, whereas fundamental research is mainly concerned with generalizations and with the formulation of a theory.

Gathering knowledge for knowledge‘s sake is termed ‘pure’ or ‘basic’ research. Research concerning some natural phenomenon or relating to pure mathematics are examples of fundamental research. Similarly, research studies, concerning human behavior carried on with a view to make generalizations about human behavior, are also examples of fundamental research, but research aimed at certain conclusions (say, a solution) facing a concrete social or business problem is an example of applied research. Research to identify social, economic or political trends that may affect a particular institution or the copy research (research to find out whether certain communications will be read and understood) or the marketing research or evaluation research are examples of applied research. Thus, the central aim of applied research is to discover a solution for some pressing practical problem, whereas basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge.

(iii) Quantitative v/s Qualitative:

Quantitative research is based on the measurement of quantity or amount. It is applicable to phenomena that can be expressed in terms of quantity. E.g. Studying the number of enquiries received for room bookings through different modes like internet, emails, calls, letters, or different sources like travel and tours operators, companies and government organizations etc.

Qualitative research, on the other hand, is concerned with qualitative phenomenon, i.e., phenomena relating to or involving quality or kind. E.g. studying the stress levels and reasons for variable performances of staff in different shifts in the same department of a hotel. The same individuals may perform differently with the change of shift timings. It can involve performing research about changing preferences of customers as per the change of season.

Another example is attitude or opinion research i.e. a research intended to find out how people feel or what they think about a particular subject or institution is also qualitative research. Through behavioral research we can evaluate the diverse factors which motivate people to behave in a particular manner or which make people like or dislike a particular thing. It is therefore important that to be relevant in qualitative research in practice the researcher should seek guidance from qualified individuals from the field opted.

(iv) Conceptual vs. Empirical:

Conceptual research is associated to some theoretical idea(s) or presupposition and is generally used by philosophers and thinkers to develop new concepts or to get a better understanding of an existing concept in practice.

On the other hand, Empirical research draws together the data based on experience or observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. It is also known as experimental research as it is essential to get facts firsthand, at their source, and actively to go about doing certain things to stimulate the production of desired information. Here the researcher develops a hypothesis and assimilates certain outcomes to start with followed by efforts to get adequate facts (data) to prove or disprove his hypothesis. An experimental design is then developed based on variables that can modify or concur the results to prove that he has given a valid statement. This also affirms that he has a reasonable control over the variables and can get different results by giving different values to them. Empirical research is appropriate when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies is today considered to be the most powerful support possible for a given hypothesis.

MAINTAINING OBJECTIVITY IN RESEARCH:

Maintaining objectivity in research means conducting studies in a way that is free from personal bias, ensuring that the results are based solely on the data and evidence, rather than the researcher's opinions or desires. It involves:

1. Clear and unbiased research questions: Formulating questions and hypotheses that don't imply a specific outcome.

2. Comprehensive literature review: Considering all relevant studies and perspectives, not just those that support your viewpoint.
3. Appropriate methodology: Choosing research methods that best answer the research questions, without bias.
4. Systematic data collection and analysis: Gathering and analyzing data in a consistent, unbiased manner.
5. Transparency: Documenting the research process in detail and being open about the limitations of the study.
6. Ethical practices: Following ethical guidelines to avoid conflicts of interest and ensure the integrity of the research.

By adhering to these principles, researchers ensure that their findings are valid, reliable, and contribute objectively to their field.

CRITERIA OF GOOD RESEARCH:

Whatever may be the types of research and studies; one thing that is important is that they all meet on the common ground of scientific method employed by them. One expects scientific research to satisfy the following criteria:

1. The purpose of the research should be clearly defined and common concepts be used.

2. The research procedure used should be described in sufficient detail to permit another researcher to repeat the research for further advancement, keeping the continuity of what has already been attained.

3. The procedural design of the research should be carefully planned to yield results that areas objective as possible.

4. The researcher should report with complete frankness, flaws in procedural design and estimate their effects upon the findings.

5. The analysis of data should be sufficiently adequate to reveal its significance and the methods of analysis used should be appropriate. The validity and reliability of the data should be checked carefully.

6. Conclusions should be confined to those justified by the data of the research and limited to those for which the data provide an adequate basis.

7. Greater confidence in research is warranted if the researcher is experienced, has a good reputation in research and is a person of integrity.

In other words, we can state the qualities of a good research as under:

1. Good research is systematic: It means that research is structured with specified steps to be taken in a specified sequence in accordance with the well-defined set of rules. Systematic

characteristic of the research does not rule out creative thinking but it certainly does reject the use of guessing and intuition in arriving at conclusions.

2. Good research is logical: This implies that research is guided by the rules of logical reasoning and the logical process of induction and deduction are of great value in carrying out research. Induction is the process of reasoning from a part to the whole whereas deduction is the process of reasoning from some premise to a conclusion which follows from that very premise. In fact, logical reasoning makes research more meaningful in the context of decision making.

3. Good research is empirical: It implies that research is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to research results.

4. Good research is replicable: This characteristic allows research results to be verified by replicating the study and thereby building a sound basis for decisions.

APPLICATIONS OF RESEARCH IN BUSINESS DECISIONS:

The role and significance of research in aiding business decision is very significant. The question one might ask here is about the critical importance of research in different areas of management. Is it most relevant in marketing? Do financial and production decisions really need research assistance? Does the method or process of research change with the functional area?

The answer to all the above questions is NO. Business managers in each field— whether human resources or production, marketing or finance—are constantly being confronted by problem situations that require effective and actionable decision making.

Applications of Research in Business Decisions

- Marketing Function
- Personnel and Human Resource Management
- Financial and Accounting Research
- Production and Operation Management
- Cross-Functional Research

Marketing Function:

This is one area of business where research is the lifeline and is carried out on a vast array of topics and is conducted both in-house by the organization itself and outsourced to external agencies. Broader industry- or product-category-specific studies are also carried out by market research agencies and sold as reports for assisting in business decisions. Studies like these could be:

- Market potential analysis; market segmentation analysis and demand estimation.
- Market structure analysis which includes market size, players and market share of the key players.

- Sales and retail audits of product categories by players and regions as well as national sales; consumer and business trend analysis—sometimes including short- and long-term forecasting.

Other than these, an organization also carries out researches related to all four Ps of marketing, such as:

Product Research: This would include new product research; product testing and development; product differentiation and positioning; testing and evaluating new products and packaging research; brand research—including equity to tracks and imaging studies.

Pricing Research: This includes price determination research; evaluating customer value; competitor pricing strategies; alternative pricing models and implications.

Promotional Research: This includes everything from designing of the communication mix to design of advertisements, copy testing, measuring the impact of alternative media vehicles, impact of competitors' strategy.

Place Research: This includes locational analysis, design and planning of distribution channels and measuring the effectiveness of the distribution network.

These days, with the onset of increased competition and the need to convert customers into committed customers, Customer Relationship Management (CRM), customer satisfaction, loyalty studies and lead user analysis are also areas in which significant research is being carried out.

Personnel and Human Resource Management:

Human Resources (HR) and organizational behaviour is an area which involves basic or fundamental research as a lot of academic, macro-level research may be adapted and implemented by organizations into their policies and programmes.

Applied HR research by contrast is more predictive and solution-oriented. Though there are a number of academic and organizational areas in which research is conducted, yet some key contemporary areas which seem to attract more research are as follows:

1. Performance Management: This includes leadership analysis development and evaluation; organizational climate and work environment studies; talent and aptitude analysis and management; organizational change implementation, management and effectiveness analysis.
2. Employee Selection and Staffing: This includes pre and on-the-job employee assessment and analysis; staffing studies.
3. Organizational Planning and Development: This includes culture assessment—either organization specific or the study of individual and merged culture analysis for mergers and acquisitions; manpower planning and development.
4. Incentive and Benefit Studies: These include job analysis and performance appraisal studies; recognition and reward studies, hierarchical compensation analysis; employee benefits and reward analysis, both within the organization and industry best practices.

5. Training and Development: These include training need gap analysis; training development modules; monitoring and assessing impact and effectiveness of training.
6. Other Areas: Other areas include employee relationship analysis; labour studies; negotiation and wage settlement studies; absenteeism and accident analysis; turnover and attrition studies and work-life balance analysis.

Financial and Accounting Research:

The area of financial and accounting research is so vast that it is difficult to provide a pen sketch of the research areas.

In this section, we are providing just a brief overview of some research topics:

1. Asset Pricing, Corporate Finance and Capital Markets: The focus here is on stock market response to corporate actions (IPOs or Initial Public Offerings, takeovers and mergers), financial reporting (earnings and firm specific announcements) and the impact of factors on returns, e.g., liquidity and volume.
2. Financial Derivatives and Interest Rate and Credit Risk Modeling: This includes analyzing interest rate derivatives, development and validation of corporate credit rating models and associated derivatives; analyzing corporate decision- making and investment risk appraisal.
3. Market Based Accounting Research: This includes analysis of corporate financial reporting behaviour; accounting-based valuations; evaluation and usage of accounting information by investors and evaluation of management compensation schemes.
4. Auditing and Accountability: This includes both private and public sector accounting studies, analysis of audit regulations; analysis of different audit methodologies; governance and accountability of audit committees.
5. Financial Econometrics: This includes modelling and forecasting involatility, risk estimation and analysis.
6. Other Areas: Other related areas of investigation are in merchant banking and insurance sector and business policy and economics areas.

Production and Operation Management:

This area of management is one in which quantifiable implementation of the research results takes on huge cost and process implications. Research in this area is highly focused and problem specific. The decision areas in which research studies are carried out are as follows:

- Operation planning which includes product/service design and development; resource allocation and capacity planning.
- Demand forecasting and decision analysis.
- Process planning which includes production scheduling and material requirement management; work design planning and monitoring. Production scheduling and material requirement management; work design planning and monitoring.
- Project management and maintenance management studies.
- Logistics and supply chain, and inventory management analysis.

- Quality estimation and assurance studies which include Total Quality Management (TQM) and quality certification analysis.

Cross-Functional Research:

Business management being an integrated amalgamation of all these and other areas sometimes requires a unified thought and approach to research. These studies require an open orientation where experts from across the disciplines contribute to and gain from the study.

For example, an area, such as new product development requires the commitment of the marketing, production and consumer insights team to exploit new opportunities.

Other areas requiring cross-functional efforts are:

- Corporate governance and ethics—the role of social values and ethics and their integration into a company's working is an area that is of critical significance to any organization.
- Technical support systems, enterprise resource planning systems, knowledge management, and data mining and warehousing are integrated areas requiring research on managing coordinated efforts across divisions.
- Ecological and environmental analysis; legal analysis of managerial actions; human rights and discrimination studies.

FORMULATING A RESEARCH PROBLEM:

Formulating a research problem is a critical first step in the research process, as it defines the focus and direction of the study. Here's how to effectively formulate a research problem:

1. Identify a Broad Topic:

Start with a general area of interest: Choose a broad subject related to your field of study or area of expertise.

Review existing literature: Conduct a thorough literature review to understand what has already been studied and identify gaps in the research.

2. Narrow Down the Topic:

Focus on a specific aspect: From the broad topic, identify a more specific issue, trend, or phenomenon that you find intriguing or that has not been extensively studied.

Consider feasibility: Ensure that the narrowed topic is manageable given your time, resources, and the scope of your study.

3. Identify the Problem Statement:

Express the issue clearly: Articulate the specific problem or question you want to address. This statement should clearly outline what is not known or what needs to be explored further.

Ensure relevance: The problem should be significant and relevant to the field, contributing to existing knowledge or addressing a practical concern.

4. Justify the Problem:

Explain the importance: Provide a rationale for why this problem is worth investigating. Highlight its significance to theory, practice, or both.

Link to literature: Show how the problem connects to existing research, identifying gaps or unresolved issues.

5. Define the Research Objectives:

Set clear objectives: Specify what you aim to achieve through your research. Objectives should be precise, measurable, and aligned with the research problem.

Formulate research questions: Develop specific research questions that your study will answer. These questions should guide your investigation and help address the research problem.

6. Consider the Scope and Constraints:

Set boundaries: Define the scope of your research by determining what aspects of the problem will be covered and what will be excluded.

Acknowledge limitations: Be aware of potential limitations, such as data availability, time constraints, or resource limitations, that may impact your research.

7. Ensure Clarity and Precision:

Use clear language: The research problem should be stated in a clear, concise, and unambiguous manner.

Avoid overly complex or broad problems: Ensure the problem is specific enough to be researched effectively within the available resources and time frame.

Example of a Research Problem Formulation:

Broad Topic: Online consumer behavior.

Narrowed Topic: The impact of social media advertising on consumer purchasing decisions.

Problem Statement: Despite the growing use of social media for marketing, there is limited understanding of how different types of social media advertisements influence consumer purchasing decisions.

Research Objective: To investigate the effectiveness of social media advertisements in influencing consumer purchasing behavior, focusing on different ad formats and consumer demographics.

Formulating a research problem effectively sets the foundation for a focused, well-structured, and meaningful study.

LITERATURE REVIEW:

What is a literature review?

A literature review is a compilation, classification, and evaluation of what other researchers have written on a particular topic. A literature review normally forms part of a research thesis but it can also stand alone as a self-contained review of writings on a subject. In either case, its purpose is to: Place each work in the context of its contribution to the subject under review; Describe the relationship of each work to the others under consideration; Identify new ways to interpret, and shed light on any gaps in, previous research; Resolve conflicts amongst seemingly contradictory previous studies; Identify areas of prior scholarship to prevent duplication of effort; Point the way forward for further research; Place an original piece of research in the context of existing literature. You can think of the above points as goals to be achieved in the process of writing a literature review. Before you can achieve any of these goals, however, you need to narrow down the possible subject areas into a relatively well-defined problem/issue, research question, or research objective.

Reasons for Literature Review:

A literature review is conducted for several important reasons in the context of academic and research work. Here are the key reasons:

1. Understanding the Research Context:

Establishing Background Knowledge: It helps researchers gain a deep understanding of the topic by summarizing what is already known.

Contextualizing Research: A literature review places your research within the broader academic field, showing how it fits into the existing body of knowledge.

2. Identifying Research Gaps:

Spotting Unanswered Questions: By reviewing existing literature, researchers can identify areas where knowledge is lacking or where previous studies have left questions unanswered.

Highlighting Inconsistencies: It can reveal contradictions or inconsistencies in previous research that need further exploration.

3. Refining Research Questions and Hypotheses:

Formulating Clear Objectives: A literature review helps refine research questions and hypotheses based on what has been previously studied and what gaps exist.

Avoiding Redundancy: It ensures that the research contributes new insights rather than duplicating what has already been done.

4. Supporting Research Design and Methodology:

Learning from Previous Studies: By examining how other researchers have approached similar topics, a literature review can guide the choice of research design, methods, and analytical techniques.

Enhancing Rigor: Understanding the strengths and weaknesses of past methodologies helps in designing a more robust and reliable study.

5. Building a Theoretical Framework:

Grounding Research in Theory: It helps to identify and adopt relevant theories or conceptual frameworks that underpin the research.

Linking to Established Concepts: Connecting your study to existing theories enhances the credibility and scholarly value of your research.

6. Demonstrating Knowledge of the Field:

Showing Expertise: A thorough literature review demonstrates to readers, reviewers, and academic peers that you are knowledgeable about the subject area.

Gaining Credibility: It establishes the researcher's credibility by showing familiarity with key studies, concepts, and debates within the field.

7. Supporting Arguments and Findings:

Providing Evidence: A literature review supports your arguments and findings by showing how they align with, differ from, or build upon previous research.

Justifying the Research: It helps justify the research's relevance, showing why the study is necessary based on the existing literature.

8. Facilitating Academic Dialogue:

Engaging with Other Researchers: By reviewing and discussing other scholars' work, you contribute to the ongoing academic conversation in your field.

Advancing Knowledge: A literature review helps to advance the field by synthesizing existing knowledge and suggesting directions for future research.

9. Avoiding Research Bias:

Ensuring Objectivity: By considering a wide range of studies and perspectives, a literature review helps minimize personal bias in the research.

Providing a Balanced Perspective: It ensures that the research is grounded in a comprehensive understanding of the topic, not just a selective or biased view.

10. Meeting Academic Requirements:

Fulfilling Research Protocols: Most academic and research institutions require a literature review as part of a thesis, dissertation, or research paper to demonstrate the research's foundation in existing knowledge.

Guiding Future Research: A well-conducted literature review can inform and guide subsequent studies by providing a solid starting point and direction.

In summary, a literature review is essential for situating your research within the existing body of knowledge, identifying gaps, refining your research questions, and demonstrating your expertise and credibility as a researcher.

REFERENCE MANAGEMENT TOOLS:

Reference management software:

Reference management software helps you to keep track of your reading and references and makes it easier to find referencing information to cite material in your work.

Using reference management software can save you time compiling and locating your references, and improves consistency and accuracy. However, it isn't a replacement for checking the accuracy of the references you use or for knowing how your references need to be written to comply with guidelines.

Here are some functions that can be performed by reference management tools for research:

- Create and store citations to efficiently generate an accurate bibliography: Integrate with your word processing program to insert citations and change journal format; can use a variety of styles and journal formats
- Search the literature: good reference management tools for research will recommend articles based on your library or written text; external search function
- Organize and store PDFs: use folders and tags to organize the reference library; search for documents by author name, keywords, text, notes; highlight passages and annotate PDF files
- Foster collaboration: allows you to share your library with colleagues
- Create mobility: can sync references across multiple devices

Reference management tools:

1. Zotero

Features: Zotero is a free, open-source tool that allows you to collect, organize, cite, and share research materials. It can automatically detect and import bibliographic information from web pages and other digital sources. Zotero also integrates with word processors like Microsoft Word and Google Docs.

Platforms: Windows, macOS, Linux, Web

2. Mendeley

Features: Mendeley is both a reference manager and an academic social network. It allows you to store and organize your references, collaborate with others online, and discover new research. Mendeley also has a PDF reader with annotation capabilities.

Platforms: Windows, macOS, Linux, Web, iOS, Android

3. EndNote

Features: EndNote is a powerful tool for managing bibliographies and references. It offers advanced features like searching for full-text articles, organizing references with tags and notes, and formatting citations and bibliographies in various styles. EndNote integrates with major word processors.

Platforms: Windows, macOS, iOS

4. RefWorks

Features: RefWorks is a web-based reference management tool often used by institutions and universities. It offers cloud storage for references, collaboration features, and citation management. RefWorks also provides integration with various databases and word processors.

Platforms: Web

5. JabRef

Features: JabRef is an open-source reference manager specifically designed for BibTeX users, making it popular among LaTeX users. It offers features like managing references, searching for full-text documents, and linking with external databases.

Platforms: Windows, macOS, Linux

6. Papers

Features: Papers is a reference manager that combines organizing references with discovering new research. It has a clean interface, powerful search features, and integrates with major citation styles. It also offers collaboration features.

Platforms: Windows, macOS, iOS

7. Citavi

Features: Citavi is a comprehensive tool for reference management, knowledge organization, and task planning. It allows users to manage references, create outlines, and organize research projects. Citavi supports both individual and team-based research.

Platforms: Windows, Web

8. BibDesk

Features: BibDesk is a reference management tool for macOS, designed specifically for BibTeX users. It allows users to manage bibliographic data and integrates well with LaTeX.

Platforms: macOS

9. Docear

Features: Docear is an academic literature management tool that integrates reference management with mind mapping. It allows users to organize their research materials visually and integrates with reference managers like Zotero.

Platforms: Windows, macOS, Linux

10. ReadCube Papers

Features: ReadCube Papers is a reference manager and citation tool with a focus on research discovery and PDF management. It offers cloud storage, citation management, and collaboration features.

Platforms: Windows, macOS, iOS, Android

Choosing the Right Tool

- For simple and intuitive use: Zotero or Mendeley.
- For advanced features and customization: EndNote or Citavi.
- For BibTeX/LaTeX users: JabRef or BibDesk.
- For collaborative projects: Mendeley, RefWorks, or Citavi.

Each tool has its strengths, so the best choice depends on your specific needs, the platforms you use, and the features you value most.

IDENTIFICATION OF RESEARCH GAP:

What is a Research Gap?

Today we are talking about the research gap: what is it, how to identify it, and how to make use of it so that you can pursue innovative research. Now, how many of you have ever felt you had discovered a new and exciting research question, only to find that it had already been written about? I have experienced this more times than I can count. Graduate studies come with pressure to add new knowledge to the field. We can contribute to the progress and knowledge of humanity. To do this, we need to first learn to identify research gaps in the existing literature.

A research gap is, simply, a topic or area for which missing or insufficient information limits the ability to reach a conclusion for a question. It should not be confused with a research question, however. For example, if we ask the research question of what the healthiest diet for humans is, we would find many studies and possible answers to this question. On the other hand, if we were to ask the research question of what are the effects of antidepressants on pregnant women, we would not find much-existing data. This is a research gap. When we identify a research gap, we identify a direction for potentially new and exciting research.

How to Identify Research Gap?

There are different techniques in various disciplines, but we can reduce most of them down to a few steps, which are:

- Identify your key motivating issue/question
- Identify key terms associated with this issue
- Review the literature, searching for these key terms and identifying relevant publications
- Review the literature cited by the key publications which you located in the above step
- Identify issues not addressed by the literature relating to your critical motivating issue

Different Types of Research Gaps:

Identifying research gaps is an essential step in conducting research, as it helps researchers to refine their research questions and to focus their research efforts on areas where there is a need for more knowledge or understanding.

1. Knowledge gaps

These are gaps in knowledge or understanding of a subject, where more research is needed to fill the gaps. For example, there may be a lack of understanding of the mechanisms behind a particular disease or how a specific technology works.

2. Conceptual gaps

These are gaps in the conceptual framework or theoretical understanding of a subject. For example, there may be a need for more research to understand the relationship between two concepts or to refine a theoretical framework.

3. Methodological gaps

These are gaps in the methods used to study a particular subject. For example, there may be a need for more research to develop new research methods or to refine existing methods to address specific research questions.

4. Data gaps

These are gaps in the data available on a particular subject. For example, there may be a need for more research to collect data on a specific population or to develop new measures to collect data on a particular construct.

5. Practical gaps

These are gaps in the application of research findings to practical situations. For example, there may be a need for more research to understand how to implement evidence-based practices in real-world settings or to identify barriers to implementing such practices.

WHAT ARE RESEARCH OBJECTIVES?

Research objectives are the guideposts that help you focus on project goals. They drive data collection, analysis and conclusions. Many research projects contain more than one research objective. Typically, research objectives appear either in the introduction of a research proposal or between the introduction and the research question.

Framing research objectives is a crucial step in the research process, as it guides the entire study and ensures that the research is focused and coherent. Here's a guide on how to frame effective research objectives:

1. Understand the Research Problem

Identify the Core Issue: Start by clearly defining the research problem or question. What gap in knowledge are you trying to address?

Contextualize the Problem: Understand the broader context of the problem, including existing research and theoretical frameworks.

2. Break Down the Problem

Specificity: Break the general problem down into smaller, more specific issues. This will help in identifying what exactly you need to explore.

Scope: Consider the scope of your study. What is feasible given your time, resources, and constraints?

3. Define Clear Objectives

Action-Oriented: Objectives should be framed as specific actions you will take to answer the research questions. Use action verbs like "determine," "analyze," "explore," "evaluate," or "compare."

Measurable: Ensure that each objective can be measured or assessed in some way. This allows you to evaluate whether you have met your objectives.

Realistic: Set achievable objectives considering the available resources, time, and scope of the research.

Relevant: Each objective should be directly related to addressing the research problem or question.

4. Structure of Research Objectives

Primary Objective: The main goal of the research, which directly addresses the research question.

Secondary Objectives: Sub-goals that support the primary objective. These could involve exploring sub-questions or examining related aspects of the problem.

5. Examples of Well-Framed Research Objectives

Primary Objective: To assess the impact of social media marketing on consumer purchasing behavior in the fashion industry.

Secondary Objectives:

To analyze the relationship between social media engagement and brand loyalty.

To evaluate the effectiveness of different social media platforms in influencing purchase decisions.

To explore the role of user-generated content in shaping consumer perceptions of fashion brands.

6. Review and Refine

Consistency: Ensure that your objectives are consistent with the overall research problem and the hypotheses (if applicable).

Clarity: Objectives should be clear and unambiguous, leaving no room for interpretation.

Feedback: Seek feedback from peers, advisors, or colleagues to refine and strengthen your objectives.

7. Linking Objectives to Methodology

Methods: Align each objective with the appropriate research methods (qualitative, quantitative, or mixed methods).

Data Collection: Define what data needs to be collected to meet each objective and how it will be analyzed.

By carefully framing your research objectives, you create a roadmap that guides your study, ensuring that it remains focused and productive. This foundation helps in developing a clear research design, choosing the right methodology, and ultimately, achieving meaningful results.

UNIT –II

HYPOTHESIS TESTING AND RESEARCH DESIGN

HYPOTHESIS TESTING

Hypothesis is usually considered as the principal instrument in research. The main goal in many research studies is to check whether the data collected support certain statements or predictions. A statistical hypothesis is an assertion or conjecture concerning one or more populations. Test of hypothesis is a process of testing of the significance regarding the parameters of the population on the basis of sample drawn from it. Thus, it is also termed as "Test of Significance'.

In short, hypothesis testing enables us to make probability statements about population parameter. The hypothesis may not be proved absolutely, but in practice it is accepted if it has withstood a critical testing.

Points to be considered while formulating Hypothesis

1. Hypothesis should be clear and precise.

2. Hypothesis should be capable of being tested.

3. Hypothesis should state relationship between variables.

4. Hypothesis should be limited in scope and must be specific.

5. Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned.

6. Hypothesis should be amenable to testing within a reasonable time.

7. Hypothesis must explain empirical reference.

FORMULATING A RESEARCH HYPOTHESIS:

Formulating a research hypothesis is a critical step in designing a study, as it provides a clear, testable prediction about the relationship between variables. Here's a structured approach to help you formulate a strong research hypothesis:

Identify the Research Problem

Example: Suppose you're interested in how sleep affects cognitive performance.

Conduct Preliminary Research

Gather information from existing literature to understand what is already known about the topic. This helps refine your focus and develop a more precise hypothesis.

Define the Variables

Independent Variable: The variable you manipulate (e.g., amount of sleep).

Dependent Variable: The variable you measure (e.g., cognitive performance).

Formulate the Hypothesis

Null Hypothesis (H_o): States that there is no effect or relationship. It's used for statistical testing.

Example: "There is no difference in cognitive performance between individuals who get 6 hours of sleep and those who get 8 hours."

Alternative Hypothesis (H_1 or Ha): States that there is an effect or relationship.

Example: "Individuals who get 8 hours of sleep will have better cognitive performance compared to those who get 6 hours."

Ensure the Hypothesis is Testable

It should be possible to collect data and analyze it to support or refute the hypothesis.

Example: Design an experiment where participants' sleep duration is controlled, and cognitive performance is measured through standardized tests.

Specify the Direction (if applicable)

Decide if your hypothesis is directional (predicting the direction of the effect) or non-directional.

Directional Hypothesis Example: "Increased sleep duration improves cognitive performance."

Non-Directional Hypothesis Example: "Sleep duration affects cognitive performance."

Refine and Revise

Make sure your hypothesis is clear, concise, and focused on a specific relationship. It should be grounded in theory and existing research.

Example Scenario

Research Problem: Does exercise impact mental health?

Variables:

Independent Variable: Exercise frequency

Dependent Variable: Levels of reported anxiety and depression

Null Hypothesis (H_o): "Exercise frequency has no effect on levels of anxiety and depression."

Alternative Hypothesis (H_1): "Increased exercise frequency reduces levels of anxiety and depression."

By following these steps, you can create a hypothesis that provides a clear direction for your research and sets the stage for rigorous testing and analysis.

TYPES OF HYPOTHESIS:

Hypotheses come in various types, depending on the nature of the research and the specific objectives of the study. Here's a breakdown of the main types:

1. Null Hypothesis (H_o)

Definition: States that there is no effect or no difference between groups or variables. It serves as a baseline or default position that assumes no relationship between the independent and dependent variables.

Purpose: To provide a statement that can be tested and potentially rejected. It helps in determining if there is enough evidence to support an alternative hypothesis.

Example: "There is no significant difference in growth rates between tomato plants exposed to full sunlight and those exposed to partial sunlight."

2. Alternative Hypothesis (H_1 or Ha)

Definition: States that there is an effect or a difference between groups or variables. It is what researchers typically aim to support through their research.

Purpose: To present a specific prediction or claim that researchers are trying to prove.

Example: "Tomato plants exposed to full sunlight will grow significantly faster than those exposed to partial sunlight."

3. Directional Hypothesis

Definition: Predicts the direction of the effect or relationship between variables. It specifies not only that there is a difference but also the nature of that difference.

Purpose: To provide a clear prediction of how one variable will affect another.

Example: "Increasing the amount of sunlight will result in a higher growth rate in tomato plants."

4. Non-Directional Hypothesis

Definition: Predicts that there will be a difference or effect, but does not specify the direction of the difference or relationship.

Purpose: To indicate that there is an expected effect without detailing the nature of that effect.

Example: "There will be a difference in growth rates between tomato plants exposed to full sunlight and those exposed to partial sunlight."

5. Complex Hypothesis

Definition: Involves multiple variables or interactions between them. It predicts the relationship between two or more independent variables and two or more dependent variables.

Purpose: To address more complex research questions involving multiple factors and their interactions.

Example: "The growth rate of tomato plants will be affected by both the amount of sunlight and the type of soil, with the greatest growth occurring in full sunlight and nutrient-rich soil."

6. Simple Hypothesis

Definition: Predicts the relationship between a single independent variable and a single dependent variable.

Purpose: To address a straightforward research question with a clear cause-and-effect relationship.

Example: "Tomato plants exposed to full sunlight will grow taller than those exposed to partial sunlight."

7. Causal Hypothesis

Definition: Suggests a cause-and-effect relationship between variables. It posits that changes in one variable will cause changes in another.

Purpose: To determine the causal impact of one variable on another.

Example: "Increasing the amount of water provided to tomato plants will result in increased growth rates."

8. Associative Hypothesis

Definition: Suggests a correlation or association between variables without necessarily implying causation. It indicates that two variables are related, but does not specify which one influences the other.

Purpose: To explore the relationships or correlations between variables.

Example: "There is an association between the amount of sunlight received and the growth rate of tomato plants."

9. Statistical Hypothesis

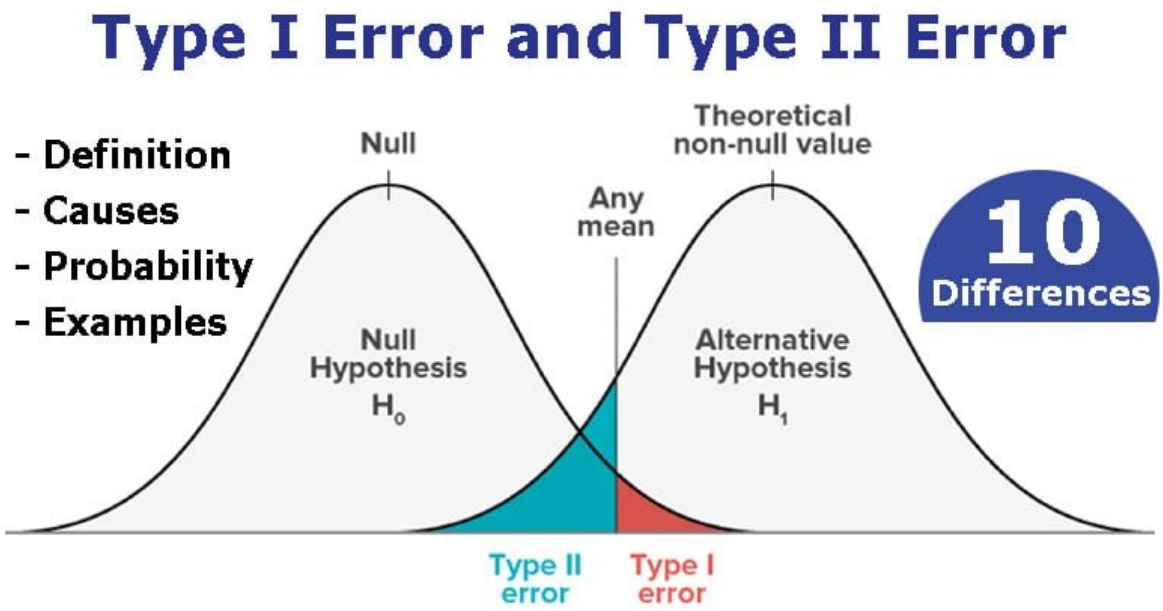
Definition: Specific to statistical testing, it includes the null and alternative hypotheses used in hypothesis testing.

Purpose: To apply statistical methods to determine if there is enough evidence to reject the null hypothesis.

Example: "The average growth rate of tomato plants in full sunlight is equal to the average growth rate of those in partial sunlight" (null hypothesis).

Each type of hypothesis serves a different role in the research process and helps guide the design, analysis, and interpretation of studies.

Type I Error and Type II Error:



Type 1 error

- Type 1 error, in statistical hypothesis testing, is the error caused by rejecting a null hypothesis when it is true.
- Type 1 error is caused when the hypothesis that should have been accepted is rejected.
- Type I error is denoted by α (alpha), known as an error, also called the level of significance of the test.
- This type of error is a false positive error where the null hypothesis is rejected based on some error during the testing.
- The null hypothesis is set to state that there is no relationship between two variables and the cause-effect relationship between two variables, if present, is caused by chance.
- Type 1 error occurs when the null hypothesis is rejected even when there is no relationship between the variables.
- As a result of this error, the researcher might believe that the hypothesis works even when it doesn't.

Type 1 error causes

- Type 1 error is caused when something other than the variable affects the other variable, which results in an outcome that supports the rejection of the null hypothesis.
- Under such conditions, the outcome appears to have happened due to some causes than chance when it is caused by chance.

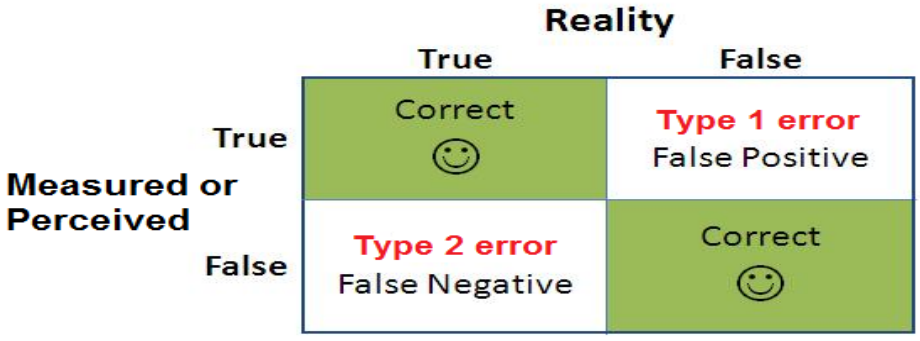
- Before a hypothesis is tested, a probability is set as a level of significance which means that the hypothesis is being tested while taking a chance where the null hypothesis is rejected even when it is true.
- Thus, type 1 error might be due to the chance/ level of significance set before the test without considering the test duration and sample size.

Probability of type 1 error

- The probability of Type I error is usually determined in advance and is understood as the significance level of testing the hypothesis.
- If the Type I error is fixed at 5 percent, there are about five chances in 100 that the null hypothesis, H0, will be rejected when it is true.
- The rate or probability of type 1 error is symbolized by α and is also termed the level of significance in a test.
- It is possible to reduce type 1 error at a fixed size of the sample; however, while doing so, the probability of type II error increases.
- There is a trade-off between the two errors where decreasing the probability of one error increases the probability of another. It is not possible to reduce both errors simultaneously.
- Thus, depending on the type and nature of the test, the researchers need to decide the appropriate level of type 1 error after evaluating the consequences of the errors.

Type 1 error examples

- For this, let us take a hypothesis where a player is trying to find the relationship between him wearing new shoes and the number of wins for his team.
- Here, if the number of wins for his team is more when he was wearing his new shoes is more than the number of wins for his team otherwise, he might accept the alternative hypothesis and determine that there is a relationship.
- However, the winning of his team might be influenced by just chance rather than his shoes which results in a type 1 error.
- In this case, he should've accepted the null hypothesis because the winning of a team might happen due to chance or luck.



Type II error

- Type II error is the error that occurs when the null hypothesis is accepted when it is not true.
- In simple words, Type II error means accepting the hypothesis when it should not have been accepted.
- The type II error results in a false negative result.
- In other words, type II is the error of failing to accept an alternative hypothesis when the researcher doesn't have adequate power.
- The Type II error is denoted by β (beta) and is also termed the beta error.
- The null hypothesis states that there is no relationship between two variables, and the cause-effect relationship between two variables, if present, is caused by chance.
- Type II error occurs when the null hypothesis is acceptable considering that the relationship between the variables is because of chance or luck, and even when there is a relationship between the variables.
- As a result of this error, the researcher might believe that the hypothesis doesn't work even when it should.

Type II error causes

- The primary cause of type II error, like a Type II error, is the low power of the statistical test.
- This occurs when the statistical is not powerful and thus results in a Type II error.
- Other factors, like the sample size, might also affect the test results.
- When small sample size is selected, the relationship between the two variables being tested might not be significant even when it does exist.
- The researcher might assume the relationship is due to chance and thus reject the alternative hypothesis even when it is true.
- There it is important to select an appropriate size of the sample before beginning the test.

Probability of type II error

- The probability of committing a Type II error is calculated by subtracting the power of the test from 1.
- If Type II error is fixed at 2 percent, there are about two chances in 100 that the null hypothesis, H0, will be accepted when it is not true.
- The rate or probability of type II error is symbolized by β and is also termed the error of the second type.
- It is possible to reduce the probability of Type II error by increasing the significance level.
- In this case, the probability of rejecting the null hypothesis even when it is true also increases, decreasing the chances of accepting the null hypothesis when it is not true.

- However, because type I and Type II error are interconnected, reducing one tends to increase the probability of the other.
- Therefore, depending on the nature of the test, it is important to determine which one of the errors is less detrimental to the test.
- For this, if a type I error involves the time and effort of retesting the chemicals used in medicine that should have been accepted. In contrast, the type II error involves the chances of several users of this medicine being poisoned, and it is wise to accept the type I error over type II.

Type II error examples

- For this, let us take a hypothesis where a shepherd thinks there is no wolf in the village, and he wakes up all night for five nights to determine the wolf's existence.
- If he sees no wolf for five nights, he might assume that there is no wolf in the village where the wolf might exist and attack the sixth night.
- In this case, when the shepherd accepts that no wolf exists, a type II error results where he agrees with the null hypothesis even when it is not true.

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

RESEARCH DESIGN:

MEANING:

Research design serves as a blueprint or framework that guides researchers in carrying out their study. It involves making crucial decisions about how to collect data, what tools to use, how to measure outcomes, and how to analyze results. By specifying the methods and techniques, research design helps to ensure that the research process is systematic, logical, and coherent.

DEFINITION:

Research design can be defined as "a detailed plan or strategy outlining how a research project will be conducted, including the methods of data collection, measurement, and analysis. It acts as a framework that ensures the study addresses the research problem, tests hypotheses, or answers research questions in a systematic and scientific manner."

THE PROCESS OF RESEARCH DESIGN:

The research design process is a systematic and structured approach to conducting research. The process is essential to ensure that the study is valid, reliable, and produces meaningful results.

Consider your aims and approaches: Determine the research questions and objectives, and identify the theoretical framework and methodology for the study.

Choose a type of Research Design: Select the appropriate research design, such as experimental, correlational, survey, case study, or ethnographic, based on the research questions and objectives.

Identify your population and sampling method: Determine the target population and sample size, and choose the sampling method, such as random, stratified random sampling, or convenience sampling.

Choose your data collection methods: Decide on the data collection methods, such as surveys, interviews, observations, or experiments, and select the appropriate instruments or tools for collecting data.

Plan your data collection procedures: Develop a plan for data collection, including the timeframe, location, and personnel involved, and ensure ethical considerations.

Decide on your data analysis strategies: Select the appropriate data analysis techniques, such as statistical analysis, content analysis, or discourse analysis, and plan how to interpret the results.

The process of research design is a critical step in conducting research. By following the steps of research design, researchers can ensure that their study is well-planned, ethical, and rigorous.

RESEARCH DESIGN ELEMENTS:

Impactful research usually creates a minimum bias in data and increases trust in the accuracy of collected data. A design that produces the slightest margin of error in experimental research is generally considered the desired outcome. The essential elements are:

- Accurate purpose statement
- Techniques to be implemented for collecting and analyzing research
- The method applied for analyzing collected details
- Type of research methodology
- Probable objections to research

- Settings for the research study
- Timeline
- Measurement of analysis

CHARACTERISTICS OF RESEARCH DESIGN:

A proper design sets your study up for success. Successful research studies provide insights that are accurate and unbiased. You'll need to create a survey that meets all of the main characteristics of a design. There are four key characteristics:

Neutrality: When you set up your study, you may have to make assumptions about the data you expect to collect. The results projected in the research should be free from research bias and neutral. Understand opinions about the final evaluated scores and conclusions from multiple individuals and consider those who agree with the results.

Reliability: With regularly conducted research, the researcher expects similar results every time. You'll only be able to reach the desired results if your design is reliable. Your plan should indicate how to form research questions to ensure the standard of results.

Validity: There are multiple measuring tools available. However, the only correct measuring tools are those which help a researcher in gauging results according to the objective of the research. The questionnaire developed from this design will then be valid.

Generalization: The outcome of your design should apply to a population and not just a restricted sample. A generalized method implies that your survey can be conducted on any part of a population with similar accuracy.

TYPES OF RESEARCH DESIGN:

A researcher must clearly understand the various types to select which model to implement for a study. Like the research itself, the design of your analysis can be broadly classified into quantitative and qualitative.

Qualitative research

Qualitative research determines relationships between collected data and observations based on mathematical calculations. Statistical methods can prove or disprove theories related to a naturally existing phenomenon. Researchers rely on qualitative observation research methods that conclude "why" a particular theory exists and "what" respondents have to say about it.

Quantitative research

Quantitative research is for cases where statistical conclusions to collect actionable insights are essential. Numbers provide a better perspective for making critical business decisions. Quantitative research methods are necessary for the growth of any organization. Insights drawn from complex numerical data and analysis prove to be highly effective when making decisions about the business's future.

Qualitative Research vs. Quantitative Research:

Here is a chart that highlights the major differences between qualitative and quantitative research:

Qualitative Research	Quantitative Research
Focus on explaining and understanding experiences and perspectives.	Focus on quantifying and measuring phenomena.
Use of non-numerical data, such as words, images, and observations.	Use of numerical data, such as statistics and surveys.
Usually uses small sample sizes.	Usually uses larger sample sizes.
Typically emphasizes in-depth exploration and interpretation.	Typically emphasizes precision and objectivity.
Data analysis involves interpretation and narrative analysis.	Data analysis involves statistical analysis and hypothesis testing.
Results are presented descriptively.	Results are presented numerically and statistically.

1. Descriptive:

In a descriptive composition, a researcher is solely interested in describing the situation or case under their research study. It is a theory-based design method created by gathering, analyzing, and presenting collected data. This allows a researcher to provide insights into the why and how of research. Descriptive design helps others better understand the need for the research. If the problem statement is not clear, you can conduct exploratory research.

2. Experimental:

Experimental research establishes a relationship between the cause and effect of a situation. It is a causal research design where one observes the impact caused by the independent variable on the dependent variable. For example, one monitors the influence of an independent variable such as a price on a dependent variable such as customer satisfaction or brand loyalty. It is an efficient research method as it contributes to solving a problem.

The independent variables are manipulated to monitor the change it has on the dependent variable. Social sciences often use it to observe human behavior by analyzing two groups.

Researchers can have participants change their actions and study how the people around them react to understand social psychology better.

3. Correlational research:

Correlational research is a non-experimental research technique. It helps researchers establish a relationship between two closely connected variables. There is no assumption while evaluating a relationship between two other variables, and statistical analysis techniques calculate the relationship between them. This type of research requires two different groups.

A correlation coefficient determines the correlation between two variables whose values range between -1 and +1. If the correlation coefficient is towards +1, it indicates a positive relationship between the variables, and -1 means a negative relationship between the two variables.

4. Diagnostic research:

In diagnostic design, the researcher is looking to evaluate the underlying cause of a specific topic or phenomenon. This method helps one learn more about the factors that create troublesome situations.

This design has three parts of the research:

- Inception of the issue
- Diagnosis of the issue
- Solution for the issue

5. Explanatory research:

Explanatory design uses a researcher's ideas and thoughts on a subject to further explore their theories. The study explains unexplored aspects of a subject and details the research questions' what, how, and why.

Benefits of Research Design:

There are several benefits of having a well-designed research plan. Including:

Clarity of research objectives: Research design provides a clear understanding of the research objectives and the desired outcomes.

Increased validity and reliability: To ensure the validity and reliability of results, research design help to minimize the risk of bias and helps to control extraneous variables.

Improved data collection: Research design helps to ensure that the proper data is collected and data is collected systematically and consistently.

Better data analysis: Research design helps ensure that the collected data can be analyzed effectively, providing meaningful insights and conclusions.

Improved communication: A well-designed research helps ensure the results are clean and influential within the research team and external stakeholders.

Efficient use of resources: reducing the risk of waste and maximizing the impact of the research, research design helps to ensure that resources are used efficiently.

A well-designed research plan is essential for successful research, providing clear and meaningful insights and ensuring that resources are practical.

DATA COLLECTION:

Data collection through a census typically involves several methods to ensure comprehensive and accurate information is gathered from the population. Here are some common methods used:

Data collection refers to the process of gathering information to analyze, interpret, and make decisions. It is a crucial part of research, surveys, and studies in various fields such as business, healthcare, social sciences, and more. There are several methods of data collection, categorized into several methods:

A. CENSUS METHOD:

A census is a collection of information from all units in the population or a 'complete enumeration' of the population. We use a census when we want accurate information for many subdivisions of the population. Such a survey usually requires a very large sample size and often a census offers the best solution.

1. Self-Enumeration (Self-Reporting)

Online Surveys: People fill out census forms through secure online portals.

Paper Forms: Census forms are mailed to households, and residents fill them out and return them by post.

2. Face-to-Face Interviews

Door-to-Door Enumeration: Census workers visit households in person to collect data, especially in areas with low literacy or internet access.

Interviewing Enumerators: Enumerators may be trained to ask specific questions and record responses directly on paper forms or tablets.

3. Telephone Interviews

In areas where face-to-face enumeration is difficult, or during pandemic situations, phone interviews can be used. Census workers call households and ask questions from the census form.

4. Administrative Records

Government and institutional records, like tax records, social security databases, school enrollment data, and healthcare records, can be used to supplement or verify data collected from households.

5. Mobile Data Collection Tools

Tablets or smartphones with preloaded forms and data-collection software are used to gather census data. This allows for quicker, more accurate data collection and real-time syncing with central databases.

6. Observation

In some cases, enumerators may use observation for specific data points (e.g., estimating the number of people in a house) when direct responses are unavailable.

7. Mail-Out, Mail-Back Method

Households receive census forms via postal mail and are asked to fill them out and return them by mail. This is commonly followed up by enumerators if responses are not received.

8. Proxy Reporting

When certain individuals are unavailable to provide data (e.g., traveling or incapacitated), someone else (a family member or neighbor) may be allowed to provide census information on their behalf.

The combination of these methods ensures a more inclusive and accurate data collection, as it helps to address various challenges like literacy barriers, lack of access to digital devices, or remote locations.

Advantages

Accurate data: Census data is generally reliable and accurate.

Sampling frame: Census data can be used as a sampling frame for future studies and surveys.

Nation-building: Census data can be a useful tool for nation-building by involving the entire population.

Small area data: Census data can provide data for small areas like counties and districts, which is useful for planning services.

Limitations of Census Method

The expenditure incurred during the census is much higher because of the sheer size of the population. Also, data is collected from each unit of a sample population, which requires additional costing.

Owing to the huge volume of data that is collated, a greater number of the workforce (as well as man-hours) is required for completion.

Costly Method: Census method is a very costly method of data collection.

Time Consuming: Census method consumes more time and labor to complete data collecting tasks.

Unsuitability: Census method is not applicable or suitable if the universe is large. This method is suitable only for a small universe.

Chance of Errors: There is a comparatively higher chance of statistical errors in this method.

B. SAMPLE SURVEY METHOD:

Data collection through a sample survey involves gathering information from a subset of the population, called a sample, rather than surveying the entire population. This method is widely used when it is impractical, time-consuming, or too expensive to collect data from everyone. Sample surveys are particularly useful for making inferences about a population based on the responses from a representative group.

1. Define the Population

The first step is to clearly define the population that the sample will represent. This could be a group of people, organizations, or any other entities of interest.

For example, if a company wants to know customer preferences, the population might be all its customers.

2. Determine the Sampling Frame

The sampling frame is a list of individuals or elements from the population that will be used to select the sample.

It is important to ensure the sampling frame is comprehensive and up-to-date, as any omissions or errors may lead to biased results.

3. Select a Sampling Method

Several sampling methods can be used to choose a representative subset of the population:

Simple Random Sampling: Every member of the population has an equal chance of being selected. This is often done using random number generators or lotteries.

Advantages: Unbiased and easy to implement.

Disadvantages: Requires a complete sampling frame and can be time-consuming for large populations.

Stratified Sampling: The population is divided into subgroups (strata) based on specific characteristics (e.g., age, income level). A random sample is then taken from each subgroup.

Advantages: Ensures that important subgroups are represented.

Disadvantages: More complex to implement.

Systematic Sampling: Every nth element is selected from the sampling frame after a random starting point.

Advantages: Easier to implement than simple random sampling.

Disadvantages: Risk of bias if there is a hidden pattern in the list.

Cluster Sampling: The population is divided into clusters, and a random sample of clusters is selected. All individuals within the chosen clusters are surveyed.

Advantages: Cost-effective for geographically dispersed populations.

Disadvantages: May lead to higher sampling error if the clusters are not representative.

Convenience Sampling: Individuals who are easiest to reach are selected.

Advantages: Quick and inexpensive.

Disadvantages: High potential for bias; not representative of the entire population.

Quota Sampling: Non-random sampling in which quotas are set for different subgroups, and individuals are chosen based on meeting these quotas.

Advantages: Ensures certain groups are included.

Disadvantages: Not random, so may introduce bias.

4. Design the Questionnaire

Structured Questions: Close-ended questions with predetermined response options (e.g., Yes/No, Likert scale).

Unstructured Questions: Open-ended questions allowing respondents to provide detailed answers in their own words.

Balanced Questions: Ensure that questions are neutral and do not lead respondents toward specific answers.

The questionnaire should be clear, concise, and relevant to the survey objectives. Pre-testing (pilot survey) is also recommended to check for any issues before conducting the full survey.

5. Collect the Data

Methods of Collection:

Online Surveys: Distributed through emails, websites, or social media platforms.

Phone Surveys: Interviews conducted over the phone.

Face-to-Face Surveys: Interviewers collect data in person, often used when a personal touch is needed.

Mail Surveys: Questionnaires are sent via postal mail, and respondents return them after completion.

The choice of data collection method depends on the target population, available resources, and the type of data being collected.

6. Analyze the Data

After the data is collected, statistical methods are applied to analyze the results.

Inferences about the entire population are made based on the sample data, using techniques like confidence intervals and hypothesis testing.

Statistical software (e.g., SPSS, R, Excel) is often used to manage and analyze survey data efficiently.

7. Interpret and Report Results

The final step is to interpret the data and provide insights relevant to the study's objectives.

It is important to highlight any limitations of the sample survey (e.g., sampling bias, non-response bias) and discuss how these might affect the conclusions.

Advantages of Sample Surveys

Cost-Effective: Surveying a sample is much less expensive than conducting a full census.

Faster Data Collection: Data can be gathered and analyzed more quickly.

Manageable: Easier to handle and process smaller datasets.

Flexibility: Sample surveys can be easily adapted to different research needs or population groups.

Disadvantages of Sample Surveys

Sampling Error: There is always the possibility that the sample may not perfectly represent the population, leading to sampling error.

Bias: If the sample is not properly chosen or there is non-response, the results may be biased.

Generalization: While inferences are made about the population, they may not always be fully accurate due to limitations in sample size or method.

Parameter	**Census**	**Sample**
Definition	A method of data collection that involves gathering information from every member of a population.	A method of data collection that involves gathering information from a selected group of individuals or units within a population.
Scope	Comprehensive, covering the entire population.	Limited, covering only a part of the population.
Cost	High, due to the exhaustive nature of data collection.	Lower, as it involves collecting data from a subset of the population.

Time Required	Longer, can take months to years depending on the population size.	Shorter, as data is collected from a smaller group.
Accuracy	Higher accuracy as it covers the entire population.	Potentially less accurate due to sampling error, but accuracy can be increased with proper sampling techniques.
Practicality	Less practical for large populations due to the high costs and time required.	More practical, especially for studies requiring quick results or when dealing with large populations.
Error Type	Subject to non-sampling errors such as errors in data collection or processing.	Subject to sampling errors, which can be estimated and adjusted for.
Resource Intensity	Very resource-intensive in terms of manpower, financial investment, and time.	Less resource-intensive, making it feasible for most research projects.
Data Collection	Data is collected from every individual unit.	Data is collected from a representative sample.
Suitability	Suitable for small populations or when detailed information is needed for every unit.	Suitable for large populations or when time and resources are limited.

c. Case study:

Data collection through case studies is a qualitative research method that involves an in-depth, detailed examination of a specific case or a few cases within a real-world context. A "case" can be an individual, organization, event, or phenomenon. This method is used to explore complex issues where the boundaries between the phenomenon and the context are not clear. It is commonly used in social sciences, psychology, business, and education.

Features of Case Study Data Collection

In-Depth Focus: Case studies provide detailed insights into a specific case rather than generalizing to a larger population.

Contextual Understanding: They emphasize understanding the context in which a phenomenon occurs.

Multiple Data Sources: Case studies often incorporate multiple data collection methods to build a comprehensive picture of the case.

Steps in Data Collection through Case Study:

1. Define the Case

The first step is identifying the specific case to be studied. This could be a person, a company, an event, or a process.

Define the boundaries of the case: What will be included, and what will be excluded? For example, if studying a company's response to a crisis, the focus might be on key decision-makers and strategies used.

2. Determine the Research Objectives

What are the goals of the case study? What specific issues, questions, or hypotheses are being explored?

Clearly define what the researcher wants to learn from the case, whether it is understanding a process, explaining an outcome, or exploring a phenomenon.

3. Select Data Collection Methods

Case studies often use multiple methods to gather rich, diverse data. The most common methods include:

Interviews: Semi-structured or unstructured interviews are used to gather qualitative data from individuals involved in the case.

Advantages: Provides in-depth, personal insights into the case.

Disadvantages: Time-consuming, and interviewer bias may influence responses.

Observations: Direct observation of behaviors, processes, or events as they occur in the real world. Researchers may take detailed notes or use video recordings.

Advantages: Offers real-time data on the case and its context.

Disadvantages: Observer bias and the possibility that the presence of the observer may alter behavior.

Documents and Archival Records: Collecting data from written or recorded materials such as reports, memos, meeting minutes, emails, and official records related to the case.

Advantages: Provides historical context and objective information.

Disadvantages: Documents may not cover all aspects of the case, and their accuracy must be verified.

Surveys or Questionnaires: Structured tools to collect information from people involved in or affected by the case.

Advantages: Can gather quantitative or qualitative data from multiple stakeholders.

Disadvantages: Limited by the design of the survey and may not capture deep insights.

Audio-Visual Materials: Photographs, videos, or audio recordings related to the case (e.g., footage of events or meetings).

Advantages: Can capture non-verbal cues and environmental context.

Disadvantages: May require careful interpretation, and some materials might be sensitive or hard to access.

4. Triangulation of Data

Triangulation is the process of using multiple data sources or methods to cross-verify findings. This helps ensure the accuracy and reliability of the case study by looking at the same phenomenon from different angles.

For example, findings from interviews may be cross-checked with observational data or archival documents to confirm their validity.

5. Conduct Data Collection

Engage in the collection of data by using the selected methods. Since case studies typically deal with real-world scenarios, the researcher needs to be adaptable and flexible during data collection to respond to changing circumstances.

Detailed field notes, audio recordings, or video documentation may be used to capture interactions and observations accurately.

6. Analyze the Data

Thematic Analysis: Common themes or patterns across the data are identified. For example, in a business case study, the researcher might identify key factors influencing decision-making processes.

Narrative Analysis: This involves telling the story of the case, focusing on individual accounts and experiences.

Content Analysis: Examining written, visual, or audio data for specific content, such as recurring ideas, concepts, or language.

7. Interpret the Findings

Interpreting case study findings requires the researcher to connect the data to the broader research questions or objectives. The aim is to explain the findings in a way that provides insights into the case and its broader implications.

The analysis must consider the context, relationships between variables, and alternative explanations.

8. Write the Case Study Report

The final report includes a detailed account of the case, data collected, analysis, and conclusions. It often follows a narrative format that tells the story of the case and integrates various data points.

The report may highlight key themes, insights, and potential implications for theory, practice, or further research.

Types of Case Studies

Exploratory Case Study: Used to explore a phenomenon in depth when it is not well understood. It often leads to the generation of hypotheses or theories.

Descriptive Case Study: Provides a detailed description of a case or situation without necessarily drawing broader conclusions or theory development.

Explanatory Case Study: Seeks to explain how or why something happened by examining causal relationships.

Intrinsic Case Study: Focuses on understanding the unique aspects of a particular case, often chosen for its uniqueness or interest.

Multiple (or Collective) Case Study: Involves studying multiple cases to identify commonalities and differences, often used for comparative analysis.

Advantages of Case Study Data Collection

Rich, Detailed Data: Case studies allow researchers to delve deep into a case, uncovering complex relationships and processes that other methods may miss.

Contextual Insights: They provide a thorough understanding of the context in which the case occurs, making them useful for real-world applications.

Flexibility: Researchers can adapt data collection methods as the study progresses, which is helpful in dynamic or unpredictable situations.

Holistic Approach: Case studies capture the complexity and interconnectedness of different variables in real-life scenarios.

Disadvantages of Case Study Data Collection

Time-Consuming: Case studies require significant time for data collection, analysis, and interpretation.

Limited Generalizability: Since case studies focus on specific cases, findings may not be easily generalized to broader populations or contexts.

Potential for Researcher Bias: The researcher's subjectivity and close involvement with the case may influence the data collection or interpretation.

Complexity in Data Analysis: Handling multiple data sources and integrating them into a coherent analysis can be challenging.

SAMPLING:

SAMPLE:

Asample is group of people, objects or items that are taken from a large population for a measurement. The sample should be representative of the population to ensure that we can generalize the findings from the research sample to the population as a whole.

SAMPLING:

Sampling is the act, process, or technique of selecting a suitable sample, or a representative part of a population for the purpose of determining parameters or characteristics of the whole population.

PURPOSE OF SAMPLING:

The purpose of sampling is to gather data from a subset of a population to make inferences or draw conclusions about the entire population without needing to collect data from every individual or element. Sampling is widely used in research, surveys, and statistical analysis because it is more efficient, cost-effective, and manageable than studying the whole population. Below are the key purposes and benefits of sampling:

1. Cost and Time Efficiency

Purpose: Sampling reduces the resources (money, time, and effort) required to collect data.

Explanation: Gathering data from an entire population (census) can be prohibitively expensive and time-consuming, especially for large populations. Sampling allows researchers to obtain relevant data in a shorter timeframe and at a lower cost by focusing on a representative portion of the population.

2. Feasibility

Purpose: Sampling makes it feasible to conduct research in situations where studying the entire population is not practical.

Explanation: In cases where the population is too large, geographically dispersed, or inaccessible, it may not be possible to survey everyone. A well-chosen sample provides a practical way to obtain useful information while maintaining study integrity.

3. Accuracy and Precision

Purpose: Sampling allows for precise and accurate data collection when the sample is carefully designed and selected.

Explanation: A representative sample can provide highly accurate insights into the population's characteristics, behaviors, or opinions, often as effectively as a full census. The key is to use proper sampling techniques to minimize bias and sampling error.

4. Manageability of Data Collection

Purpose: Sampling simplifies data collection, processing, and analysis.

Explanation: Handling large datasets can be cumbersome and difficult to manage. Sampling reduces the volume of data to a manageable size, making it easier to process, analyze, and interpret findings without overwhelming resources or analytical tools.

5. Generalizability

Purpose: Sampling helps researchers make inferences about a population based on data from a subset of that population.

Explanation: A carefully selected sample allows researchers to generalize the results to the entire population with a certain level of confidence. This is especially important when the goal is to understand broader trends, behaviors, or opinions without having to measure every individual.

6. Understanding Variability

Purpose: Sampling helps in understanding the variability and patterns within a population.

Explanation: By selecting a representative sample, researchers can study how different subgroups within the population behave, their characteristics, and how these differ or converge. This variability is key in fields like marketing, healthcare, and social sciences, where understanding different segments is essential.

7. Testing Hypotheses

Purpose: Sampling enables researchers to test hypotheses about a population without needing data from everyone.

Explanation: By studying a sample, researchers can draw conclusions and test hypotheses about the population with a degree of statistical confidence. This is especially useful in scientific experiments, public health studies, or political polling, where statistical methods are applied to validate or refute hypotheses.

8. Reducing Data Redundancy

Purpose: Sampling reduces data redundancy and avoids unnecessary collection of similar data.

Explanation: In many populations, there may be redundancy or repetition in the data. Sampling ensures that data collected is diverse and meaningful, avoiding unnecessary duplication of information that would result from studying the entire population.

9. Ethical and Practical Considerations

Purpose: Sampling allows for ethical and practical handling of sensitive or limited data.

Explanation: In fields such as medical research, it may be ethically or practically impossible to collect data from the entire population (e.g., due to risks, costs, or patient confidentiality). Sampling ensures that researchers can still conduct studies without overburdening participants or exposing them to risks unnecessarily.

STEPS IN SAMPLING DESIGN:

1. Define the Population

Population must be defined in terms of elements, sampling units, extent and time. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.

For ex, if a kitchen appliances firm wants to conduct a survey to ascertain the demand for its micro ovens, it may define the population as ‘all women above the age of 20 who cook (assuming that very few men cook)’. However this definition is too broad and will include every household in the country, in the population that is to be covered by the survey. Therefore the definition can be further refined and defined at the sampling unit level, that, all

women above the age 20, who cook and whose monthly household income exceeds Rs.20,000. This reduces the target population size and makes the research more focused. The population definition can be refined further by specifying the area from where the researcher has to draw his sample, that is, households located in Hyderabad.

2. Specifying the Sampling Frame

Once the definition of the population is clear a researcher should decide on the sampling frame. A sampling frame is the list of elements from which the sample may be drawn. Continuing with the micro oven ex, an ideal sampling frame would be a database that contains all the households that have a monthly income above Rs.20, 000. However, in practice it is difficult to get an exhaustive sampling frame that exactly fits the requirements of a particular research. In general, researchers use easily available sampling frames like telephone directories and lists of credit card and mobile phone users. Various private players provide databases developed along various demographic and economic variables. Sometimes, maps and aerial pictures are also used as sampling frames. Whatever may be the case, an ideal sampling frame is one that entire population and lists the names of its elements only once.

3. Specifying the Sampling Unit

Asampling unit is a basic unit that contains a single element or a group of elements of the population to be sampled. In this case, a household becomes a sampling unit and all women above the age of 20 years living in that particular house become the sampling elements. If it is possible to identify the exact target audience of the business research

4. Selection of the Sampling Method

The sampling method outlines the way in which the sample units are to be selected. The choice of the sampling method is influenced by the objectives of the business research, availability of financial resources, time constraints, and the nature of the problem to be investigated

5.Determination of Sample Size

The sample size plays a crucial role in the sampling process. There are various ways of classifying the techniques used in determining the sample size. A couple those hold primary importance and are worth mentioning are whether the technique deals with fixed or sequential sampling and whether its logic is based.

6. Specifying the Sampling Plan

In this step, the specifications and decisions regarding the implementation of the research process are outlined. Suppose, blocks in a city are the sampling units and the households are the sampling elements.

7. Selecting the Sample

This is the final step in the sampling process, where the actual selection of the sample elements is carried out. At this stage, it is necessary that the interviewers stick to the rules

outlined for the smooth implementation of the business research. This step involves implementing the sampling plan to select the sampling plan to select a sample required for the survey.

Techniques of Sampling:

❖ Probability Sampling Techniques

❖ Non Probability Sampling Probability Sampling Techniques

A probability sampling technique is one in which one can specify for each element of population, the probability of its being included in the sample. Every probability can be expressed in the form of a proportion e.g. the probability of getting a head in testing a coin is 1/2 or 1 chance in 2 trials. Thus, probability samples are characteristised by the fact that the probability of selection of each unit is known. In the sample of example each of the elements has the same probability of being included as in random sampling method. An essential quality of a probability sample is that it makes possible representative sampling plans. It also provides an estimate of the extent to which the sample characteristics or findings are likely to differ from the total population.

Major Forms of Probability Sampling Methods are:

❖ Simple random sampling method,

In a day to day business, the term random is frequently used for careless, unpremeditated, casual haphazard activity or process. Which means that a random samples is drawn carelessly in unplanned manner, without a definite aim or deliberate purpose. This concept is not correct. Random sampling correctly means the arranging of conditions in such a manner that every item of the whole universe from which we are to select the sample shall have the same chance of being selected as any other item. Random sampling, therefore, involves careful planning and orderly procedure.

Steps of Simple Random Sampling

➢ Involves listing or cataloguing of all the elements in the population and assigning them consecutive numbers.

➢Deciding upon the desired sample size.

➢Using any method of sampling, a certain number of elements from the list is selected.

Advantages of Random Sampling Technique

➢ Most basic, simple and easy method

➢Provides a representative sample.

Disadvantages

➢In most cases it is difficult to find data list of all units of the population to be sampled.

➢The task of numbering every unit before the sample is chosen is time consuming and expensive.

➢The units need not only to be numbered but also arranged in a specified order.

➢The possibility of obtaining a poor or misleading sample is always present when random selection is used.

Methods of Drawing, Sample in Random Method

Lottery Method:

The numbers of all the elements of the universe are written on different tickets or pieces of paper of equal size shape and colour. Which are then shuffled thoroughly in a box, or a container. Then tickets are then drawn randomly their numbers are noted and the corresponding individuals or objects are studied.

Tippets Numbers:

It was first developed by Prof L. H. C. Tippet and since then is known by his name. He developed a list of 10,400 sets of numbers randomly, each set being of four digits there numbers are written on several pages in unsystematic order.

Grid Method:

This method is applied in selection of the areas. Suppose we have to select any number of areas from a town or any number of towns from a province for survey. For selection, first a map of the whole area is prepared. The area is often divided into different blocks. A transparent plate is made equivalent to the size of the map that consists of several holes in it which carries different numbers. By random sampling method it is decided as to which numbers are to be included in the sample.

Systematic Sampling Method

In this method first of all a list is prepared of all the elements of the universe on the basis of a selection criterion. A list may be prepared in alphabetical order, as given in the telephone directory. Then from the list every third, every tenth every twentieth or any number in the like manner can be selected. For the application of this method, preparing a list of all the elements and numbering them is essential. Secondly, the population needs to be homogenous in nature. Social phenomenon is variable in nature and individuals are heterogeneous. However on their social characteristics they are homogenous viz. we may decide to cover only the students, the professors, the slum dwellers etc. The characteristics to be selected for this purpose must be relevant to the problem under study. Advantages

➢ It is frequently used because it is simple, direct and in- expensive.

➢When a list of names or items is available, systematic sampling is often an efficient approach. Disadvantages

➢One should not use systematic sampling in case of exploring unfamiliar areas because listing of elements is not possible

➢When there is a periodic fluctuation in the characteristic under examination in relation to the order in which the items appear, the methods is ineffective

❖ Stratified random sampling method

When the population is divided into different strata or groups and then samples are selected from each stratum by simple random sampling procedure or by regular interval method, we call it as stratified random sampling method. According to the nature of the problem relevant criteria are selected for stratification. Among the possible stratifying criteria, cum age, sex, family income, number of years of education, occupation, religion, race, place of residence etc. On the basis of characteristics universe can be divided into different strata or stratum, Each stratum has to be homogeneous from within such a division can be done on the basis of any single criterion. e.g. on the basis of age we can divide people into below 25 and above 25 groups, on the basis of education into matriculates and non matriculates etc. Stratification can also be done on the basis of a combination of any two or more criteria viz. on the basis of sex and education, we can divide the people into four groups.

❖ Educated women

❖ Un-educated women

❖ Education men

❖Un educated men

Elements are then selected from each stratum through simple a random sampling method. An estimate is made for each stratum separately. These estimates are combined to provide an estimate for the entire population. Purpose: The primary purpose is to increase the representatives of the sample without increasing the size of the sample on the basis of having greater knowledge of the population characteristics.

Advantages

❖The population is first stratified into different groups and then the elements of the sample are selected from each group. Therefore, the different groups are sure to have representation in the sample. In case of random sample, there is possibility that bigger groups have greater representation and the smaller groups are often eliminated or under represented.

❖With more homogenous population greater precision can be achieved with fewer cases. This saves time in collecting and processing of the data when detailed study about population characteristics are wanted it is more effective.

❖ As compared to random samples, stratified samples are geographically more concentrated and thus save time, money and energy, in money from one address to another.

Disadvantages

❖Unless there are extreme differences between the strata, the expected proportional representation would be small. Here a random sampling may give a nearly proportional representation.

❖Even after stratification, the sample is selected from each stratum either by simple random sampling method or by systematic sampling method; as such the draw backs of both methods can be present.

❖For application of the stratified method, one must know the characteristics of the specified population in which the study is to be made. He must also know as to which characteristics are related to the subject under investigation and therefore can be considered as relevant for stratification.

❖The process of stratification becomes more and more complicated and difficult as the numbers of characteristics to be used for stratification are increased.

Cluster Sampling

In cluster sampling the stratification is done in a manner that the groups are heterogeneous in nature rather than homogenous. Here the elements are not selected from each stratum as is done in stratified sampling, rather the elements are obtained by taking a sample of group and not from within groups. That means that out of several clusters or groups, one, two or more number of clusters are selected by simple or stratified random method and their elements are studied. All the elements in these clusters are not to be included in the sample; the ultimate selection from within the clusters is also carried out on simple or stratified sampling basis. Purpose: The purpose of a cluster sample is to reduce cost and not essentially to increase percussion.

Advantage

❖In cluster sampling the cost per element is greatly reduced.

❖It becomes possible to take a larger sample and regain the amount of precision

❖It can be used in situations where it is impossible to obtain sample by other methods. Disadvantage

❖It is a complicated sample design the researcher has to be highly skilled in sampling.

❖Its standard errors are almost inevitably larger then those of sample random sampling.

Multi-stage sampling

The method is used in selecting a sample from a very large area. As the name suggests m.s. sampling refers to a sampling technique which is carried out in various stages. Normally a

multi-stage sampling is the one that combines cluster and random sampling methods. Eg., if we want to study the socio-economic background, attitudes and motivations of slum dwellers, we can first make a list of the cities which would thus make our clusters. From these clusters we can select any number of cities. Then each city or cluster would be stratified into different slum areas. Thus our cities can be called as primary sampling units and the slum areas as secondary sampling units.

Non Probability Sampling

In non-probability sampling techniques one cannot estimate beforehand the probability of each element being included in the sample. It does not also assure that every element has a chance of being included. In probability sampling, one has to prepare or know at least all the elements of the total population from which the sample is to be drawn. This makes the sampling procedure costlier and more time consuming.

The major forms of non-probability samples are:

❖ Accidental samples and; ❖ Purposive samples

Accidental sampling means selecting the units on the basis of easy approaches. Here one selects the sample that fall to hand easily. E.g. suppose one is studying the political socialization and political participation among university and college students of A.U. and his sample size is 100. He would go to the university campus and would select the first hundred students whom he happens to meet, whether in class room, or in students common room or in field. Such type of sampling is easy to do and saves time and money. But the chores of bias are also great.

❖ Quota samples

In quota sampling the interviewers are interested to interview a specified number of persons from each category. The required numbers of elements from each category are determined in the office ahead of time according to the number of elements in each category. Thus an interviewer would need to contact a specified number of men and specified number of women, from different age categories from different religious or social groups etc. The basic purpose of quota sampling is the selection of a sample that no true replace of the population about which one wants to generalize. Advantage

❖If properly planned and executed, a quota sample is most likely to give maximum representative sample of the population.

❖In purposive sampling one picks up the cases that are considered to be typical of the population in which to one is interested.

❖The cases are judged to be typical on the basis of the need of the researcher.

❖Since the selection of elements is based upon the judgment of the researcher, the purposive sampling as called judgment sample.

❖The researcher trees in his sample to match the universe in some of the important known characteristics.

Disadvantage

❖The defect with this method is that the researcher can easily make esser in judging as to which cases are typical.

Purposive Sampling "Deliberate Sampling" or "Judgment Sampling".

❖When the researcher deliberately selects certain units from the universe, it is known as purposive sampling.

❖ However, it must be kept in mind that the units selected must be representative of the universe. ❖That, the names may be selected from a Telephone Diretory, Automobile Registration Records (RTOs) etc.

Advantage

❖ Quote sampling is a stratified cum purposive sampling and thus enjoys the benefits of both samplings.

❖It proper controls or checks are imposed, it is likely to give accurate results.

❖It is only useful method when no sample frame is available.

Convenience Sampling

It is known as unsystematic, careless, accidental or opportunistic sampling. Under this a sample is selected according to the convenience of the investigator. May be use when

❖ Universe is not clearly defined

❖ Sampling units are not clear

❖Complete source list is not available

Testing of Reliability:

Reliability is the consistency of your measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. In short, it is the repeatability of your measurement. A measure is considered reliable if a person's score on the same test given twice is similar. It is important to remember that reliability is not measured, it is estimated.

There are two ways that reliability is usually estimated: Test/Retest and Internal Consistency.

Test/Retest: Test/retest is the more conservative method to estimate reliability. Simply put, the idea behind test/retest is that you should get the same score on test 1 as you do on test 2. The three main components to this method are as follows

- Implement your measurement instrument at two separate times for each subject;
- Compute the correlation between the two separate measurements; and
- Assume there is no change in the underlying condition (or trait you are trying to measure) between test 1 and test 2.

Internal Consistency:

Internal consistency estimates reliability by grouping questions in a questionnaire that measure the same concept. For example, you could write two sets of three questions that measure the same concept (say class participation) and after collecting the responses, run a correlation between those two groups of three questions to determine if your instrument is reliably measuring that concept.

The primary difference between test/retest and internal consistency estimates of reliability is that test/retest involves two administrations of the measurement instrument, whereas the internal consistency method involves only one administration of that instrument.

Validity:

Validity is the strength of our conclusions, inferences or propositions. More formally, Cook and Campbell (1979) define it as the "best available approximation to the truth or falsity of a given inference, proposition or conclusion." In short, were we right? Let's look at a simple example. Say we are studying the effect of strict attendance policies on class participation. In our case, we saw that class participation did increase after the policy was established. Each type of validity would highlight a different aspect of the relationship between our treatment (strict attendance policy) and our observed outcome (increased class participation).

Types of Validity;

There are four types of validity commonly examined in social research:

Conclusion validity asks is there a relationship between the programme and the observed outcome? Or, in our example, is there a connection between the attendance policy and the increased participation we saw?

Internal Validity asks if there is a relationship between the programme and the outcome we saw, is it a causal relationship? For example, did the attendance policy cause class participation to increase?

Construct validity is the hardest to understand in my opinion. It asks if there is there a relationship between how I operationalized my concepts in this study to the actual causal relationship I'm trying to study? Or in our example, did our treatment (attendance policy)

reflect the construct of attendance, and did our measured outcome – increased class participation – reflect the construct of participation? Overall, we are trying to generalize our conceptualized treatment and outcomes to broader constructs of the same concepts.

External validity refers to our ability to generalize the results of our study to other settings. In our example, could we generalize our results to other classrooms?

Comparison b/w Validity and Reliability

The real difference between reliability and validity is mostly a matter of definition. Reliability estimates the consistency of your measurement, or more simply the degree to which an instrument measures the same way each time it is used in under the same conditions with the same subjects. Validity, on the other hand, involves the degree to which you are measuring what you are supposed to, more simply, the accuracy of your measurement. It is my belief that validity is more important than reliability because if an instrument does not accurately measure what it is supposed to, there is no reason to use it even if it measures consistently (reliably).

So what is the relationship between validity and reliability? The two do not necessarily go hand-in hand. At best, we have a measure that has both high validity and high reliability. It yields consistent results in repeated application and it accurately reflects what we hope to represent.

It is possible to have a measure that has high reliability but low validity – one that is consistent in getting bad information or consistent in missing the mark. It is also possible to have one that has low reliability and low validity – inconsistent and not on target.

Finally, it is not possible to have a measure that has low reliability and high validity – you can't really get at what you want or what you're interested in, if your measure fluctuates wildly.

SAMPLING AND NON SAMPLING ERRORS

❖ The error assign out due to drawing inferences about population on the basis of few observations (sampling), is termed 'sampling error'.

❖In the complete enumeration survey since the whole population is surveyed, sampling error in this sense in non-existent. However, the mainly arising at the stage of ascertainment and processing of data, which are termed non-sampling errors, are common both in complete enumeration and sample surveys.

Sampling Errors:

Even if utmost care has been taken in selecting a sample, the results derived from a sample study may not be exactly equal to the true value in the population. The reason is that estimate is based on a part and not on the whole and samples are seldom, if ever, perfect miniature of the population. Hence sampling gives rise to certain errors known as sampling errors. However, the errors can be controlled. The modern sampling theory helps in designing the

survey in such a manner that the sampling errors can be made small. Sampling errors are of two types:

❖ biased, and

❖Un-biased

Biased Errors:

These errors arise from any bias in selection, estimation, etc. For example, if in place of simple random sampling, deliberate sampling has been used in a particular case some bias is introduced is the result and hence such errors are called sampling errors.

Un-biased Errors: These errors arise due to "chance" differences between the members of the population included in the sample and those not included. An error in statistics is the difference between the value of a statistic and that of the corresponding parameter.

❖Thus the total sampling error is made up of errors due to bias, if any and the random sampling error.

❖The bias error, forms a constant component of error that does not decrease in large population as the number of sample increases. Such error is also known as cumulative or non-compensating error. The random sampling error, on the other hand, decreases, on an average, as the size of sample increases. Such errors are, therefore, known as non-cumulative or compensating error.

Causes of Bias: Bias may arise due to:

❖ Faulty process of selection;

❖ Faulty work during the collection; and

❖ Faulty methods of analysis

Faulty Selection: Deliberate selection of a 'representative' sample. Substitution: Substitution of an item in place of one chosen in random sample sometimes lead to bias.

Non response: It all the items to be included in the sample are not covered then there will be bias even though no substitution has been attempted. An appeal to the variety of the person questioned may give rise to yet another kind of bias. For example, the question. Are you a good student? is such that most of the students would succumb to variety and answer 'Yes'.

Bias Due to Faulty Collection of Data: Any consistent error in measurement will give rise to bias whether the measurements are carried out on a sample or on all units of the population. The danger of error is, however, likely to be greater in sampling work. Bias may arise due to improper formulation of the decision, problem or strongly defining the population etc. Bias observation may result from poorly designed questionnaire, ill-trained interviewer, failure of a respondent's memory.

Bias in Analysis: In addition to bias, which arises from faulty process of selection and faulty collection of information, faulty methods of analysis may also introduce such bias. Such bias can be avoided by adopting the proper method of analysis.

Avoidance of Bias: If the possibility of bias exists, fully objective conclusion cannot be drawn. The first essential of any sampling or census procedure must, therefore, be the elimination of all sources of bias.

Method of Reducing Sampling Errors

- Once the absence of bias has been ensured, attention should be given to the random sampling errors. Such errors must be reduced to the minimum so as to attain the desired accuracy.
- Apart from reducing errors of bias, the simplest way of increasing the accuracy of a sample is to increase its size. The sampling error usually decreases with increase in sample size, and in fact in many situations the decrease is inversely proportional to the square root of the sample size.
- From this diagram it is clear that though the reduction in sampling error is substantial for initial increases in sample size, it becomes marginal after a certain stage. In other words, considerably greater effort is needed after a certain stage to decrease the sampling error this is the initial instance.
- From this point of view it could be said that there is a strong case for resorting to a sample survey to provide estimates within permissible margins of error instead of a complete enumeration survey.

Non Sampling Errors

- As regards non-sampling errors they are likely to be more in case of complete enumeration survey than in case of a sample survey. When a complete enumeration of units in the universe is needs, one would expect that it would give rise to date free from errors. However, in practice it is not so. For example, it is difficult to completely avoid errors of observation or ascertainment. Similarly, in the processing of data, tabulation errors may be committed, affecting the final result. Errors arising in this manner are termed as non-sampling errors. Non-sampling error can occur at every stage of planning and execution of census or survey. Such errors can arise due to a number of causes such as defective methods of data collection, and tabulation, faulty definition, incomplete coverage etc. More specifically, non-sampling errors may arise from one or more of the following factors:
  Data specification may be inadequate and inconsistent with respect to the objectives of the study.
- Inaccurate or inappropriate method of interview, observation or measurement with inadequate on ambiguous schedules.
- Lack of trained and experienced investigators.
- Lack of inadequate inspection and supervision of primacy staff.
- Errors due to non-response.
- Errors in data processing operations.
- Errors committed during presentation and printing of tabulated results.

Control of Non Sampling Errors: In some situations the non-sampling errors may be large and deserve greater attention than sampling errors. While, in general, sampling error decrease with increase in sample size, non-sampling error tends to increase with the sample size.

Increase of complete enumeration non-sampling errors and incase of sample surveys both sampling and non-sampling errors require to be controlled and reduced at a level at which their presence does not vitiate the use of final result.

Reliability of Samples:

The reliability of samples can be tested in the following ways. More samples of the same size should be taken from the same universe and their results be compared. If the results are similar, the sample will be reliable. If the measurements of the universe are known, then they should be compared with the measurements of the sample. In case of similarity of measurements, the sample will be reliable.

UNIT – III

DATA COLLECTION

VARIABLE

Meaning:

In research, a variable is a characteristic, quantity, or number that can be measured or quantified, and that can change or vary. Variables are essential to research because they allow researchers to:

- Frame research questions
- Formulate hypotheses
- Interpret results
- Gain insights into relationships, causes, and effects

Variables can be categorized in different ways, including:

- Type of data: Whether the variable is quantitative or categorical
- Role in the study: Whether the variable is independent, dependent, controlled, or confounding
- Relationship to other variables: Whether the variable is confounding or controlled

Some examples of variables include:

- Age
- Height
- Satisfaction levels
- Economic status
- Time it takes for something to occur
- Whether or not an object is used within a study

TYPES OF VARIABLE:

Qualitative Variables

Qualitative variables are those that express a qualitative attribute, such as hair color, religion, race, gender, social status, method of payment, and so on. The values of a qualitative variable do not imply a meaningful numerical ordering.

The value of the variable 'religion' (Muslim, Hindu.., etc..) differs qualitatively; no ordering of religion is implied. Qualitative variables are sometimes referred to as categorical variables.

For example, the variable sex has two distinct categories: 'male' and 'female.' Since the values of this variable are expressed in categories, we refer to this as a categorical variable.

Similarly, the place of residence may be categorized as urban and rural and thus is a categorical variable.

Categorical variables may again be described as nominaland ordinal.

Ordinal variables can be logically ordered or ranked higher or lower than another but do not necessarily establish a numeric difference between each category, such as examination grades (A+, A, B+, etc., and clothing size (Extra large, large, medium, small).

Nominal variables are those that can neither be ranked nor logically ordered, such as religion, sex, etc.

A qualitative variable is a characteristic that is not capable of being measured but can be categorized as possessing or not possessing some characteristics. ▸ iedunote.com/variables

Quantitative Variables

Quantitative variables, also called numeric variables, are those variables that are measured in terms of numbers. A simple example of a quantitative variable is a person's age.

Age can take on different values because a person can be 20 years old, 35 years old, and so on. Likewise, family size is a quantitative variable because a family might be comprised of one, two, or three members, and so on.

Each of these properties or characteristics referred to above varies or differs from one individual to another. Note that these variables are expressed in numbers, for which we call quantitative or sometimes numeric variables.

A quantitative variable is one for which the resulting observations are numeric and thus possess a natural ordering or ranking.

Discrete and Continuous Variables

Quantitative variables are again of two types: discrete and continuous.

Variables such as some children in a household or the number of defective items in a box are discrete variables since the possible scores are discrete on the scale.

For example, a household could have three or five children, but not 4.52 children.

Other variables, such as 'time required to complete an MCQ test' and 'waiting time in a queue in front of a bank counter,' are continuous variables.

The time required in the above examples is a continuous variable, which could be, for example, 1.65 minutes or 1.6584795214 minutes.

Of course, the practicalities of measurement preclude most measured variables from being continuous.

Discrete Variable

A discrete variable, restricted to certain values, usually (but not necessarily) consists of whole numbers, such as the family size and a number of defective items in a box. They are often the results of enumeration or counting.

A few more examples are;

- The number of accidents in the twelve months.
- The number of mobile cards sold in a store within seven days.
- The number of patients admitted to a hospital over a specified period.
- The number of new branches of a bank opened annually during 2001- 2007.
- The number of weekly visits made by health personnel in the last 12 months.

Continuous Variable

A continuous variable may take on an infinite number of intermediate values along a specified interval. Examples are:

- The sugar level in the human body;
- Blood pressure reading;
- Temperature;

- Height or weight of the human body;
- Rate of bank interest;
- Internal rate of return (IRR),
- Earning ratio (ER);
- Current ratio (CR)

No matter how close two observations might be, if the instrument of measurement is precise enough, a third observation can be found, falling between the first two.

A continuous variable generally results from measurement and can assume countless values in the specified range.

Dependent Variables and Independent Variable

In many research settings, two specific classes of variables need to be distinguished from one another: independent variable and dependent variable.

Many research studies aim to reveal and understand the causes of underlying phenomena or problems with the ultimate goal of establishing a causal relationship between them.

Look at the following statements:

- Low intake of food causes underweight.
- Smoking enhances the risk of lung cancer.
- Level of education influences job satisfaction.
- Advertisement helps in sales promotion.
- The drug causes improvement of health problems.
- Nursing intervention causes more rapid recovery.
- Previous job experiences determine the initial salary.
- Blueberries slow down aging.
- The dividend per share determines share prices.

In each of the above queries, we have two independent and dependent variables. In the first example, 'low intake of food' is believed to have caused the 'problem of being underweight.'

It is thus the so-called independent variable. Underweight is the dependent variable because we believe this 'problem' (the problem of being underweight) has been caused by 'the low intake of food' (the factor).

Similarly, smoking, dividend, and advertisement are all independent variables, and lung cancer, job satisfaction, and sales are dependent variables.

In general, an independent variable is manipulated by the experimenter or researcher, and its effects on the dependent variable are measured.

Independent Variable

The variable that is used to describe or measure the factor that is assumed to cause or at least to influence the problem or outcome is called an independent variable.

The definition implies that the experimenter uses the independent variable to describe or explain its influence or effect of it on the dependent variable.

Variability in the dependent variable is presumed to depend on variability in the independent variable.

Depending on the context, an independent variable is sometimes called a predictor variable, regressor, controlled variable, manipulated variable, explanatory variable, exposure variable (as used in reliability theory), risk factor (as used in medical statistics), feature (as used in machine learning and pattern recognition) or input variable. ▸ iedunote.com/variables

The explanatory variable is preferred by some authors over the independent variable when the quantities treated as independent variables may not be statistically independent or independently manipulable by the researcher.

If the independent variable is referred to as an explanatory variable, then the term response variable is preferred by some authors for the dependent variable.

Dependent Variable

The variable used to describe or measure the problem or outcome under study is called a dependent variable.

In a causal relationship, the cause is the independent variable, and the effect is the dependent variable. If we hypothesize that smoking causes lung cancer, 'smoking' is the independent variable and cancer the dependent variable.

A business researcher may find it useful to include the dividend in determining the share prices. Here dividend is the independent variable, while the share price is the dependent variable.

The dependent variable usually is the variable the researcher is interested in understanding, explaining, or predicting.

In lung cancer research, the carcinoma is of real interest to the researcher, not smoking behavior per se. The independent variable is the presumed cause of, antecedent to, or influence on the dependent variable.

Depending on the context, a dependent variable is sometimes called a response variable, regressand, predicted variable, measured variable, explained variable, experimental variable, responding variable, outcome variable, output variable, or label.

An explained variable is preferred by some authors over the dependent variable when the quantities treated as dependent variables may not be statistically dependent.

If the dependent variable is referred to as an explained variable, then the term predictor variable is preferred by some authors for the independent variable.

Levels of an Independent Variable

If an experimenter compares an experimental treatment with a control treatment, then the independent variable (a type of treatment) has two levels: experimental and control.

If an experiment were to compare five types of diets, then the independent variables (types of diet) would have five levels.

In general, the number of levels of an independent variable is the number of experimental conditions.

Background Variable

In almost every study, we collect information such as age, sex, educational attainment, socioeconomic status, marital status, religion, place of birth, and the like. These variables are referred to as background variables.

These variables are often related to many independent variables, so they indirectly influence the problem. Hence they are called background variables.

The background variables should be measured if they are important to the study. However, we should try to keep the number of background variables as few as possible in the interest of the economy.

Moderating Variable

In any statement of relationships of variables, it is normally hypothesized that in some way, the independent variable 'causes' the dependent variable to occur.

In simple relationships, all other variables are extraneous and are ignored.

In actual study situations, such a simple one-to-one relationship needs to be revised to take other variables into account to explain the relationship better.

This emphasizes the need to consider a second independent variable that is expected to have a significant contributory or contingent effect on the originally stated dependent-independent relationship.

Such a variable is termed a moderating variable.

Suppose you are studying the impact of field-based and classroom-based training on the work performance of health and family planning workers. You consider the type of training as the independent variable.

If you are focusing on the relationship between the age of the trainees and work performance, you might use 'type of training' as a moderating variable.

Extraneous Variable

Most studies concern the identification of a single independent variable and measuring its effect on the dependent variable.

But still, several variables might conceivably affect our hypothesized independent-dependent variable relationship, thereby distorting the study. These variables are referred to as extraneous variables.

Extraneous variables are not necessarily part of the study. They exert a confounding effect on the dependent-independent relationship and thus need to be eliminated or controlled for.

An example may illustrate the concept of extraneous variables. Suppose we are interested in examining the relationship between the work status of mothers and breastfeeding duration.

It is not unreasonable in this instance to presume that the level of education of mothers as it influences work status might have an impact on breastfeeding duration too.

Education is treated here as an extraneous variable. In any attempt to eliminate or control the effect of this variable, we may consider this variable a confounding variable.

An appropriate way of dealing with confounding variables is to follow the stratification procedure, which involves a separate analysis of the different levels of lies in confounding variables.

For this purpose, one can construct two cross tables for illiterate mothers and the other for literate mothers.

Suppose we find a similar association between work status and duration of breast-feeding in both the groups of mothers. In that case, we conclude that mothers' educational level is not a confounding variable.

Intervening Variable

Often an apparent relationship between two variables is caused by a third variable.

For example, variables X and Y may be highly correlated, but only because X causes the third variable, Z, which in turn causes Y. In this case, Z is the intervening variable.

An intervening variable theoretically affects the observed phenomena but cannot be seen, measured, or manipulated directly; its effects can only be inferred from the effects of the independent and moderating variables on the observed phenomena.

We might view motivation or counselling as the intervening variable in the work-status and breastfeeding relationship.

Thus, motive, job satisfaction, responsibility, behaviour, and justice are some of the examples of intervening variables.

Suppressor Variable

In many cases, we have good reasons to believe that the variables of interest have a relationship, but our data fail to establish any such relationship. Some hidden factors may suppress the true relationship between the two original variables.

Such a factor is referred to as a suppressor variable because it suppresses the relationship between the other two variables.

The suppressor variable suppresses the relationship by being positively correlated with one of the variables in the relationship and negatively correlated with the other. The true relationship between the two variables will reappear when the suppressor variable is controlled for.

Thus, for example, low age may pull education up but income down. In contrast, a high age may pull income up but education down, effectively cancelling the relationship between education and income unless age is controlled.

TECHNIQUES OF DATA COLLECTION

Data collection is the process of collecting, measuring and analysing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyze them to make critical business decisions. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses. There are two main methods of data collection in research based on the information that is required, namely:

- Primary Data Collection
- Secondary Data Collection

1. PRIMARY DATA COLLECTION

Primary data refers to original data collected directly from its source for a specific research or analysis purpose. This information has not been previously gathered, processed, or interpreted by anyone else. It is the data that researchers or analysts collect first-hand. Primary data collection methods include surveys, interviews, experiments, observations, or direct measurements.

Primary data is often contrasted with secondary data, which others have already collected and analysed for a different purpose. It is valuable because it can be tailored to address specific research questions or objectives and is typically more reliable and relevant to the study.

For example, if a company conducts a customer satisfaction survey to gather customer feedback, the responses it collects from the survey would be considered primary data. The company gathers this data directly for its use and analysis.

Advantages:

Primary data collection has many advantages over traditional data collection methods. Primary data is collected directly from the people who are experiencing the problem or issue you're trying to solve. This means that it's more accurate and reliable than other types of data, which can be collected through surveys or interviews.

Traditional data collection methods rely on asking questions of a sample of people in order to get an understanding of the problem or issue. However, this method isn't always as accurate as primary data because it doesn't allow for feedback from the people who are experiencing the problem.

By collecting primary data, you're able to gather information from those who are affected by the issue at hand. This allows you to get a better understanding of how they're feeling and what needs to be done in order to address their concerns.

It also makes research more efficient since there's no need for a large number of respondents—just enough people who have experienced the issue firsthand will do fine.

Plus, primary data is often more relevant because it considers all aspects of an individual's experience rather than just one aspect (like with survey results). Ultimately, this leads to better solutions that reflect everyone's reality accurately and efficiently.

Limitations of Primary data:

1. Time-consuming

Primary data collection can be a lengthy process, from designing data collection tools to analyzing the results.

2. Expensive

Primary data collection can be costly, requiring resources for material creation, data gathering, personnel, and data analysis.

3. Limited scope

Primary data collection is usually focused on a specific research question or context, which may limit the breadth of the data.

4. Bias

The process of collecting primary data can be susceptible to various biases, which can compromise the data's accuracy and reliability.

5. Ethical, legal, or logistical challenges

There may be challenges in accessing and contacting the target population.

6. Invasive and disruptive

Primary data collection can be invasive and disruptive, often requiring people to take time away from their normal activities.

7. May not be representative

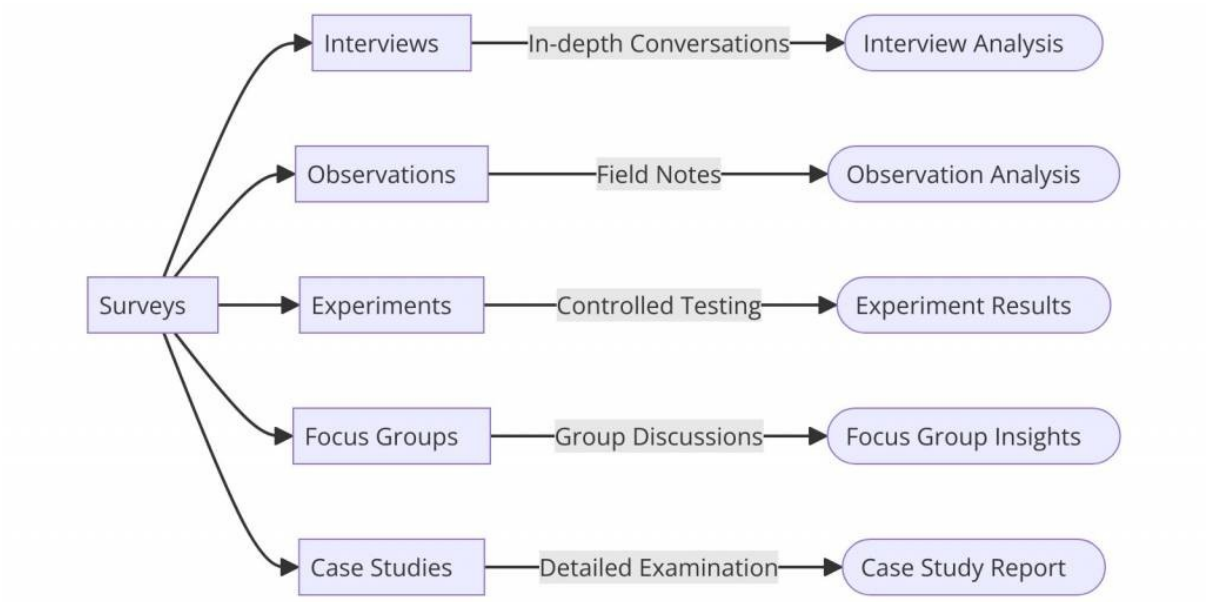
Primary data may not be representative of the entire audience.

8. Incomplete, inconsistent, or inaccurate

Primary sources can be incomplete, inconsistent, or inaccurate due to gaps, errors, contradictions, or distortions.

Methods of Primary Data Collection

Primary data collection involves gathering first-hand information directly from the source for specific research purposes. This process includes various methods, allowing researchers to obtain relevant and accurate data tailored to their study's objectives.



1. INTERVIEW:

Interviews are a direct method of data collection. It is simply a process in which the interviewer asks questions and the interviewee responds to them. It provides a high degree of flexibility because questions can be adjusted and changed anytime according to the situation.

Techniques of Interview:

Here are some common techniques used in research interviews:

Structured interviews:

In a structured interview, the interviewer asks a set of standard, predetermined questions about particular topics, in a specific order. The respondents need to select their answers from a list of options. The interviewer may provide clarification on some questions. Structured Interviews are typically used in surveys (see our “Survey Research Methods” Tip Sheet for more information).

Semi-structured interviews:

In a semi-structured interview, the interviewer uses a set of predetermined questions and the respondents answer in their own words. Some interviewers use a topic guide that serves as a checklist to ensure that all respondents provide information on the same topics. The interviewer can probe areas based on the respondent's answers or ask supplementary questions for clarification. Semi-structured interviews are useful when there is a need to collect in-depth information in a systematic manner from a number of respondents or interviewees (e.g., teachers, community leaders).

Unstructured interviews:

In an unstructured interview, the interviewer has no specific guidelines, restrictions, predetermined questions, or list of options. The interviewer asks a few broad questions to engage the respondent in an open, informal, and spontaneous discussion. The interviewer also probes with further questions and/or explores inconsistencies to gather more in-depth information on the topic. Unstructured interviews are particularly useful for getting the stories behind respondents' experiences or when there is little information about a topic.

Steps in Conducting an Interview:

Before the Interview:

1. Define your objectives → identify what you want to achieve and the information you need to gather. Make sure an interview is the appropriate way to meet your objectives.

2. Choose the type of interview → Review your required information, budget, time, and potential respondents and decide whether you need to conduct structured, semi-structured, or unstructured interviews.

3. Choose the appropriate respondents → Depending on the type of interview, decide on the characteristics of interviewees and the number of interviews required.

4. Decide how you will conduct the interviews → Consider telephone or face-to face interviews. For large surveys, consider computer-aided interviewing and recording.

5. Decide how to recruit your respondent's → Obtain contact information for a number of respondents larger than the number of interviews you need, since some may not respond. Contact them by phone, e-mail, or regular mail and introduce yourself, your organization, and your project. Explain the purpose of the interview, the importance of their participation, and set up an appointment.

6. Decide how you will record the interviews → Depending on the type of interview, you may fill in a prepared form, use written notes, voice recorders, or computer aided devices.

7. Make a list of questions and test them with a sample of respondent's → the questions must be aligned with the type of interview. If you are running structured interviews, see our Tip Sheets on "Questionnaire Design" and Survey Research Methods" for more information.

8. Decide who will conduct the interviews → develop an information kit that includes an introduction to the research topic and instructions. For unstructured interviews, you may need to hire skilled interviewers.

During the interview:

1. Introduce yourself and initiate a friendly but professional conversation.

2. Explain the purpose of your project, the importance of their participation, and the expected duration of the interview.

3. Be prepared to reschedule the interview if a respondent has a problem with the timing.

4. Explain the format of the interview.

5. Tell respondents how the interview will be recorded and how the collected information will be used → if possible, obtain their written consent to participate.

6. Ask respondents if they have any questions.

7. Control your tone of voice and language → remain as neutral as possible when asking questions or probing on issues.

8. Keep the focus on the topic of inquiry and complete the interview within the agreed time limit.

9. Ensure proper recording → without distracting the respondent, check your notes and voice recorder regularly.

10. Complete the session → make sure all questions were asked, explain again how you will use the data, thank the respondent, and ask them if they have any questions.

After the interview

1. Make sure the interview was properly recorded → make additional notes, if needed.

2. Organize your interview responses → responses from unstructured and semi-structured interviews need to be transcribed.

3. Responses from structured interviews need to be entered into a data analysis program. Get ready for data analysis → search for resources for analyzing qualitative and/or quantitative data.

SCHEDULE:

"The schedule is nothing more than a list of questions which it seems necessary to test the hypothesis". A schedule is a structure of set of questions on a given topic which are asked by an interviewer is investigated personally. Schedule is the most important tool. It is similar to a questionnaire. It is administered by the researcher in person and it is filled up by the researcher. As the schedule is presented in person it need not be attractive. The techniques of preparing questionnaire are also applied to framing an interview schedule. Here are the key techniques of schedule sampling:

1. Fixed Interval Sampling

Definition: Data is collected at regular, predetermined time intervals (e.g., every hour, day, or week).

Purpose: To ensure systematic and consistent data collection across a fixed period.

Advantages: Simplifies scheduling, provides regular data points for analysis.

Disadvantages: May miss key events if they fall between sampling intervals.

Example: Monitoring employee productivity every two hours during a workday.

2. Random Interval Sampling

Definition: Data collection occurs at random intervals within a specified timeframe.

Purpose: To avoid patterns or biases that may arise from fixed intervals.

Advantages: Captures more natural and unbiased behaviors.

Disadvantages: Requires careful planning to ensure randomness; may be harder to schedule.

Example: Observing customer behavior in a store at random times throughout the day.

3. Event-Based Sampling

Definition: Data is collected when a specific event or trigger occurs, rather than based on time.

Purpose: To capture data related to a particular activity or phenomenon.

Advantages: Ensures data is relevant to the event of interest, focused data collection.

Disadvantages: If the event occurs infrequently, data collection may be sparse.

Example: Recording physiological responses during a stressful situation or customer reactions when a product is introduced.

4. Time-Point Sampling

Definition: Data is collected at particular moments or time points, often selected based on relevance to the study (e.g., before, during, or after an event).

Purpose: To analyze changes or trends at critical time points.

Advantages: Focused and relevant to key moments, efficient use of time and resources.

Disadvantages: May miss data between time points, limiting the understanding of transitions.

Example: Measuring mood before, during, and after an exam.

5. Signal-Contingent Sampling

Definition: Participants are signalled to provide data at certain times, either through electronic prompts (like alarms) or messages.

Purpose: To gather real-time data at specific times during the day.

Advantages: Reduces recall bias, ensures data is captured in real-time.

Disadvantages: Relies on participants' availability and responsiveness.

Example: An app sends a signal for participants to report their mood or activity every few hours.

6. Experience Sampling Method (ESM)

Definition: Participants report on their experiences, thoughts, or behaviors in real-time, often prompted by electronic devices at random or scheduled times.

Purpose: To capture data on participants' real-time experiences and contexts.

Advantages: Reduces recall bias and collects data in a natural environment.

Disadvantages: Can be disruptive to participants and may lead to incomplete data if participants miss prompts.

Example: Participants use a smartphone app to answer questions about their emotions at random times throughout the day.

7. Diary Sampling

Definition: Participants keep a diary and record events or experiences at specified times or after specific activities.

Purpose: To collect detailed and self-reported data over time.

Advantages: Offers rich, personalized data and can capture subjective experiences.

Disadvantages: Relies on participants' consistency and accuracy, prone to recall bias.

Example: Asking participants to record their daily stress levels and activities each evening.

8. Ecological Momentary Assessment (EMA)

Definition: A type of sampling where participants are asked to record their thoughts, feelings, or behaviors as they happen in their natural environment.

Purpose: To capture data in real-time in natural settings.

Advantages: Reduces memory bias and captures experiences in context.

Disadvantages: Intrusive for participants, requiring frequent data input.

Example: Participants reporting their emotional state via an app whenever they feel a mood change.

9. Interval-Contingent Sampling

Definition: Data collection happens at predetermined intervals (like daily or weekly) based on the nature of the study.

Purpose: To gather regular updates from participants over a specified period.

Advantages: Ensures systematic data collection, good for longitudinal studies.

Disadvantages: Rigid schedule may miss significant fluctuations in experiences.

Example: Participants filling out a health survey every morning.

10. Hybrid Sampling

Definition: A combination of different sampling techniques (e.g., mixing event-based and time-based sampling).

Purpose: To increase data richness and flexibility.

Advantages: Maximizes data collection opportunities, allows for comprehensive analysis.

Disadvantages: More complex to design and analyze.

Example: Observing behavior both at random times and during specific events, such as meetings or social interactions.

QUESTIONNAIRE

A questionnaire is a form prepared and distributed to secure responses to certain questions. It is a device for securing answers to questions by using a form which the respondent fills by himself. It is a systematic complication of questions and organised series of questions that are to be sent to the population samples. It is an important instrument in normative-survey research, being used to gather information from widely scattered sources. The questionnaire procedure normally comes into use where one cannot readily see personally all of the people from whom the research desires responses or where there is no particular reason to see them personally.

Purpose of questionnaire is twofold: i) to collect information from the respondents who are scattered in a vast area, ii) to achieve success in collecting reliable and dependable data.

FORMS OF QUESTIONNAIRE

Structured vs. non-structured:

The structured contains definite, concrete and direct questions, whereas non-structured may consist of partially completed questions or statements.

A non-structured questionnaire is often used as the interview guide, which is non-directive. The interviewer possesses only a blueprint of the enquiries and he is largely free to arrange the form or statements of the questions. The enquiries framed in a general form beforehand are given a specific form during the actual process of interview.

Closed form vs. open form:

The question that call for short check responses are known as restricted or closed form type. They provide for making a yes or no, a short response, or checking an item out of a list of given responses. It restricts the choice of response for the respondent. He has simply to select a response out of supplied responses and has not to frame his response in his own way.

It is easy to fill out, takes less time, keeps the respondent on the subject, is relatively more objective, more acceptable and convenient to the respondent, and is fairly easy to tabulate and analyse.

The open-form, open-end or unrestricted type questionnaire calls for a free response in the respondent's own words. The respondent frames and supplies his own response. No clues are provided. It probably provides for greater depth of response. The subject reveals his mind, gives his frame of reference and possibility the reasons for his responses

The mixed questionnaire:

The mixed questionnaire consists of both close and open type questionnaires. For social research, this method is very useful. Many questionnaires include both open and closed type items. Each type has its specific merits and limitations and the research worker has to decide which type is more likely to supply the information he wants.

Fact and opinion questionnaire:

The questionnaire are also classified as: i) questionnaire of fact, which requires certain information of facts from the respondent without any reference to his opinion or attitude about them and ii) questionnaire of opinion and attitude in which the informant's opinion, attitude or preference regarding some phenomena is sought.

Pictorial questionnaire:

In the pictorial questionnaire, pictures are used to promote interest in answering questions. It is used extensively in studies of social attitudes and prejudices in children or illiterate persons. In a pictorial questionnaire, the selected alternative answers in the form of pictures are given and the respondent is required to tick the picture concerned. This questionnaire may be very useful for collecting data in a developing country like India, specially from the rural masses who are mostly illiterate and less knowledgeable. The serious limitation of this questionnaire is that it is lengthy in form. Also it is highly expensive.

In the questionnaire technique, great reliance is placed on the respondent's verbal report for data on the stimuli or experiences to which he is exposed and for knowledge of his behaviour. The questionnaire is effective only when the respondent is able or willing to express his reactions clearly. A good questionnaire can elicit cooperation of the respondent to get frank answers on almost any subject, even such personal matters as sex and income. Thus, it is clear that the respondent can judge the study only by what he can see. The questionnaire, by its very nature, is an impersonal technique and it is several pieces of paper appeals/persuades the respondent that he ought to participate.

- RULES FOR CONSTRUCTING A QUESTIONNAIRE:
- Clearly worded items(eg: avoid usually ,most)
- Short questions
- Only one idea(concept)per question
- Avoid negatively worded items
- Avoid technical language and jargon
- Avoid leading questions(cues to answers)
- Avoid lengthy questions

The following factors are to be considered before drafting the questionnaire.

Covering letter: The person conducting the survey should introduce himself to the respondents through a covering letter. In this covering letter, one can state the objectives of his study along with a formal request to fill up the questionnaire.

Number of questions: As far as possible, the number of questions should be limited. There should be no repetition of questions. The response from the respondents will be poor if the numbers of questions are too many. Hence, care must be taken to minimise the number of questions.

Sequence of questions: The questions should be arranged in a logical order. The sequential arrangement of questions makes it easy for the respondents to make a spontaneous reply. Eg: It is irrelevant to ask the number of children before asking whether the respondent is married or not.

Ambiguity of questions: The questions should be unambiguous, that is, questions should give only one meaning. There should be only one answer to a question. The question should focus on only one point.

Questions should be short and simple: The questions should not be lengthy. They must be short and easily understandable by the respondent. As far as possible technical terms should be avoided.

Personal questions: As far as possible, questions of personal and pecuniary nature should not be asked. Eg: Questions about income, sales tax paid and the like may not be answered by respondents in writing. If this information is essential, if must be obtained by personal interviews. Therefore, these questions are to be avoided unless otherwise the study actually requires it.

Instructions to the informants: The questionnaire should provide necessary instructions to the informants. For example, it should specify the time within which it should be sent back and the address to which it should be sent. Instructions necessary to fill up the questions can also be given in the questionnaire.

Type of answer: As far as possible the answers for the questions should be objective type, that is 'Yes' or 'No' type questions are most welcome. However, when the alternative is not clear cut, the 'Yes' or 'No' questions should be avoided.

Questions requiring calculations: Questions requiring calculation of ratios, percentages, and totals should not be asked as it may take much time and the respondents may feel reluctant.

Attraction: A questionnaire should be made to look as attractive as possible. The printing and paper used should be neat and qualitative. Enough space should be left for answering the questions.

QUESTIONNAIRE DESIGN PROCESS:

STEP I: Determine survey objectives, resources and constraints
STEP II: Determine the data collection method
STEP III: Determine the question response format
STEP IV: Decide on the question wording
STEP V: Establish questionnaire flow and layout
STEP VI: Evaluate the questionnaire
STEP VII: obtain approval of all relevant parties
STEP VIII: Pre-test And Revise
STEP IX: Prepare and final copy
STEP X: Implement the survey

TYPES OF QUESTIONNAIRES:

Exploratory questionnaire (qualitative)

Exploratory questionnaires are Structured questionnaire analysis used to collect the qualitative data that information can be observed and recorded but not in a numerical form. It's used to obtain approximate and characterize the data. A case of personal information would be somebody giving your input about your composition. They may specify things about the tone, clearness, word decision, and so forth, it causes you to order your essay. However, you can't connect a number to the criticism in the questionnaire development in research. Exploratory surveys are ideal when you're in the beginning phases and need to become familiar with a subject before planning an answer or theory. For instance, in case you're in the beginning phases of item improvement. You don't think enough about the market; at that point, exploratory questionnaires are ideal using the questionnaire hypothesis survey.

Formal standardized questionnaire (quantitative)

They're otherwise called organized surveys. These are utilized to gather quantitative information which is data recorded as a check or mathematical worth. The data is quantifiable, which implies it very well may be utilized for numerical counts or factual investigation. It addresses the topic of how much, the number of, or how frequently. A case of quantitative information would be the response to the accompanying inquiry, "how old are you?" which requires a mathematical answer. Normalized surveys are best utilized when you've framed underlying speculation or worked out a model for an item. You'll utilize it to stretch test your suspicions, plans, use cases, and so forth before going further with item advancement. Because of its reasonable centre, the inquiries you pose are limited in scope and request detailed data.

Similarly, as essential as the survey type are the inquiry types you pick. Not all inquiry types are ideal in each circumstance. That is the reason it's vital to comprehend the kind of poll you're making first. With that data, it gets simpler to pick the right sorts of inquiry useful for questionnaire design business research

Open-ended questionnaire

As the name states, these questions are open for the respondent to answer with more freedom. Instead of presenting a set of answers choices, the respondent writes as much is as little as they want. It is ideal for exploratory questionnaires which collect Qualitative data analysis.

Closed questionnaire

Closed questionnaires structure the appropriate response by just permitting reactions which fit into pre-chosen classes. Information that can be put into a classification is called ostensible information. The classification can be limited to as not many as two choices, i.e., dichotomous (e.g., 'yes' or 'no,' 'male' or 'female'), or incorporate very unpredictable arrangements of choices from which the respondent can pick (e.g., multiple choices). Closed questionnaires can likewise give ordinal information (which can be positioned). This type frequently includes utilizing a persistent rating scale to gauge the quality of perspectives or feelings and useful in business survey questionnaire design. For instance, emphatically concur/concur/nonpartisan/differ/firmly differ/incapable to reply.

Multiple-choice questionnaire

This inquiry gives the respondent top-notch of answer choices, and they can choose at least one. The test with numerous decision questions is giving fragmented answer choices. For instance, you may ask what industry accomplish your work in and rattle off 5 of the most widely recognized enterprises. There are more than five ventures on the planet so that a few people won't be spoken to in this circumstance. A basic answer to this issue is adding an "other" choice.

Dichotomous questionnaire

A question with only two possible answers is Dichotomous questionnaire. It often solves a yes or no problem, but it can also be something like agree/disagree or true/false. Use this when all you need is necessary validation without going too deeply into the motivations.

Scaled questionnaire

Scaled questions are common in questionnaires, and they are mainly used to judge the degree of a feeling. Both exploratory and standardized questionnaires can be used because there are many different types of scaled questions such as:

- Rating scale
- Likert scale
- Semantic differential scale

Pictorial questionnaire

Images are the final type of question used in questionnaires substitutes' text. Respondents are asked a question and allowed to choose pictures. It usually has a greater response rate than other question types.

OBSERVATIONS

Technique is to watch what they do, to record this in some way and then to describe, analyse and interpret what we have observed. It is commonly used in an exploratory phase, typically in an unstructured form, to seek to find out what is going on a situation as a precursor to subsequent testing out of the insights obtained.

Observation can also be used as a supportive or supplementary method to collect data that may complement or set in perspective data obtained by other means. Suppose that the main effort in a particular study is devoted to a series of interviews, observation might then be used to validate or corroborate the messages obtained in the interviews.

Two popular types are participant observation and non-participant observation. Participant observation is an essentially qualitative style which has been used in variety of disciplines including in the legal profession. Participant observation is a widely used method in flexible designs, particularly those which follow an ethnographic approach. Non-participant observation can be structured but is more usually unstructured and informal.

Observation as a data collection method can be structured or unstructured. Structured or systematic observation, data collection is conducted using specific variables and according to a pre-defined schedule. Unstructured observation, is conducted in an open and free manner in a sense that there would be no pre-determined variables or objectives.

Participant observation: A key feature of participant observation is that the observer seeks to become some kind of a member of the observed group. This involves not only a physical presence and a sharing of life experiences but also entry into their social and 'symbolic' world through learning their social conventions and habits, their use of language and non-verbal communication, and so on. The observer also has to establish some role within the group. The primary data are the interpretations by the observer of what is going on around him. The observer is the research instrument, and hence great sensitivity and personal skills are called for if worthwhile data are to be collected. Participant observation might be useful in a small project: with small groups, for events or processes that take a reasonably short time, for frequent events, for activities that are accessible to observers, when your prime motivation is to find out what is going on, and when you are not short of time.

The complete participant: The complete participant role involves the observer concealing that she is an observer, acting as naturally as possible and seeking to become a full member of the group.

The participant as observer: It is a feasible alternative to have the participant as observer role. The fact that the observer is an observer is made clear to the group from the start. The observer then tries to establish close relationships with members of the group. This stance means that as well as observing through participating in activities, the observer can ask members to explain various aspects of what is going on. It is important to gain the trust of key members of the group. It would appear that this role would have more of a disturbing effect on the phenomena observed than that of the complete participant, and several experienced participant observers have documented this. However, one effect may be that members of the group are led to more analytical reflection about processes and other aspects of the group's functioning.

The marginal participant: In some situations, it may be feasible and advantageous to have a lower degree of participation than that envisaged in the preceding sections. This can be done by adopting the role of a larger passive, though completely accepted, participant - a passenger in a train or bus, or a member of the audience at a concert or sports meeting.

The observer as participant: This is someone who takes no part in the activity but whose status as a researcher is known to the participants. Such a state is aspired to by many researchers using systematic observation. However, it is questionable whether anyone who is known to be a researcher can be said not to take part in the activity - in the sense that their role is now one of the roles within the larger group that includes the researcher.

SECONDARY DATA COLLECTION

The next techniques of data collection is Secondary data collection which involves using existing data collected by someone else for a purpose different from the original intent. Researchers analyze and interpret this data to extract relevant information. Secondary data can be obtained from various sources, including:

a. Published Sources: Researchers refer to books, academic journals, magazines, newspapers, government reports, and other published materials that contain relevant data.

b. Online Databases: Numerous online databases provide access to a wide range of secondary data, such as research articles, statistical information, economic data, and social surveys.

c. Government and Institutional Records: Government agencies, research institutions, and organizations often maintain databases or records that can be used for research purposes.

d. Publicly Available Data: Data shared by individuals, organizations, or communities on public platforms, websites, or social media can be accessed and utilized for research.

e. Past Research Studies: Previous research studies and their findings can serve as valuable secondary data sources. Researchers can review and analyze the data to gain insights or build upon existing knowledge.

UNIT – IV

DATA ANALYSIS

Data Analysis

Data analysis is the systematic process of inspecting, cleaning, transforming, and interpreting data with the objective of discovering valuable insights and drawing meaningful conclusions. This process involves several steps:

1. Inspecting: Initial examination of data to understand its structure, quality, and completeness.
2. Cleaning: Removing errors, inconsistencies, or irrelevant information to ensure accurate analysis.
3. Transforming: Converting data into a format suitable for analysis, such as normalization or aggregation.
4. Interpreting: Analyzing the transformed data to identify patterns, trends, and relationships.

Types of Data Analysis Techniques in Research

Data analysis techniques in research are categorized into qualitative and quantitative methods, each with its specific approaches and tools. These techniques are instrumental in extracting meaningful insights, patterns, and relationships from data to support informed decision-making, validate hypotheses, and derive actionable recommendations. Below is an in-depth exploration of the various types of data analysis techniques commonly employed in research:

1) Qualitative Analysis:

Definition: Qualitative analysis focuses on understanding non-numerical data, such as opinions, concepts, or experiences, to derive insights into human behavior, attitudes, and perceptions.

Content Analysis: Examines textual data, such as interview transcripts, articles, or open-ended survey responses, to identify themes, patterns, or trends.

Narrative Analysis: Analyzes personal stories or narratives to understand individuals' experiences, emotions, or perspectives.

Ethnographic Studies: Involves observing and analyzing cultural practices, behaviors, and norms within specific communities or settings.

2) Quantitative Analysis:

Quantitative analysis emphasizes numerical data and employs statistical methods to explore relationships, patterns, and trends. It encompasses several approaches:

a) Descriptive Analysis:

Frequency Distribution: Represents the number of occurrences of distinct values within a dataset.

Central Tendency: Measures such as mean, median, and mode provide insights into the central values of a dataset.

Dispersion: Techniques like variance and standard deviation indicate the spread or variability of data.

b) Diagnostic Analysis:

Regression Analysis: Assesses the relationship between dependent and independent variables, enabling prediction or understanding causality.

ANOVA (Analysis of Variance): Examines differences between groups to identify significant variations or effects.

c) Predictive Analysis:

Time Series Forecasting: Uses historical data points to predict future trends or outcomes.

Machine Learning Algorithms: Techniques like decision trees, random forests, and neural networks predict outcomes based on patterns in data.

d) Prescriptive Analysis:

Optimization Models: Utilizes linear programming, integer programming, or other optimization techniques to identify the best solutions or strategies.

Simulation: Mimics real-world scenarios to evaluate various strategies or decisions and determine optimal outcomes.

e) Specific Techniques:

Monte Carlo Simulation: Models probabilistic outcomes to assess risk and uncertainty.

Factor Analysis: Reduces the dimensionality of data by identifying underlying factors or components.

Cohort Analysis: Studies specific groups or cohorts over time to understand trends, behaviors, or patterns within these groups.

Cluster Analysis: Classifies objects or individuals into homogeneous groups or clusters based on similarities or attributes.

Sentiment Analysis: Uses natural language processing and machine learning techniques to determine sentiment, emotions, or opinions from textual data.

UNIVARIATE ANALYSIS

It involves examining a single variable at a time. This approach allows you to summarize and describe the distribution of that variable without considering relationships with other factors.

In univariate analysis, you focus on measures such as:

- Central tendency (mean, median, mode)
- Dispersion (range, standard deviation, variance)

- Distribution shape (skewness, kurtosis)

To visualize univariate data, you can use tools like histograms, box plots, or violin plots. These graphical representations help you understand the spread of your data and identify potential outliers.

I Central Tendency

One of the important objectives of statistics is to find out various numerical values which explain the inherent characteristics of a frequency distribution. The first of such measures is averages. The averages are the measures which condense a huge unwieldy set of numerical data into single numerical values which represent the entire distribution. The inherent inability of the human mind to remember a large body of numerical data compels us to few constants that will describe the data. Averages provide us the gist and give a bird's eye view of the huge mass of unwieldy numerical data. Averages are the typical values around which other items of the distribution congregate. This value lies between the two extreme observations of the distribution and give us an idea about the concentration of the values in the central part of the distribution. They are called the measures of central tendency.

Averages are also called measures of location since they enable us to locate the position or place of the distribution in question. Averages are statistical constants which enables us to comprehend in a single value the significance of the whole group. According to Croxlon and Cowden, an average value is a single value within the range of the data that is used to represent all the values in that series. Since an average is somewhere within the range of data, it is sometimes called a measure of central value. An average is the most typical representative item of the group to which it belongs and which is capable of revealing all important characteristics of that group or distribution.

Measures of central tendency, Mean, Median, Mode, etc., indicate the central position of a series. They indicate the general magnitude of the data but fail to reveal all the peculiarities and characteristics of the series. In other words, they fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. This can be explained by certain other measures, known as „Measures of Dispersion" or Variation.

The study of statistics does not show much interest in things which are constant. The total area of the Earth may not be very important to a research-minded, person but the area covered by different crops, forests, residential and commercial buildings are figures of great importance, because these figures keep on changing from time to time and from place to place. Many experts are engaged in the study of changing phenomena.

Experts working in different countries keep a watch on forces which are responsible for bringing changes in the fields of human interest. Agricultural, industrial and mineral production and their transportation from one area to other parts of the world are of great interest to economists, statisticians, and other experts. Changes in human populations, changes in standards of living, changes in literacy rates and changes in prices attract experts to perform detailed studies and then correlate these changes to human life. Thus variability or variation is connected with human life and its study is very important for mankind. Objects of Central Tendency: The most important object of calculating an average or measuring central

tendency is to determine a single figure which may be used to represent a whole series involving magnitudes of the same variable. Second object is that an average represents the empire data; it facilitates comparison within one group or between groups of data. Thus, the performance of the members of a group can be compared with the average performance of different groups. Third object is that an average helps in computing various other statistical measures such as dispersion, skewness, kurtosis etc.

Different methods of measuring "Central Tendency" provide us with different kinds of averages. The following are the main types of averages that are commonly used:

1. Mean

2. Median

3. Mode

1. Arithmetic Mean:

Arithmetic mean is the most commonly used average or measure of the central tendency applicable only in case of quantitative data; it is also simply called the "mean". Arithmetic mean is defined as: "Arithmetic mean is a quotient of sum of the given values and number of the given values". Arithmetic mean can be computed for both ungrouped data (raw data: data without any statistical treatment) and grouped data (data arranged in tabular form containing different groups).

Pros and Cons of Arithmetic Mean:

Pros:

- It is rigidly defined
- It is easy to calculate and simple to follow
- It is based on all the observations
- It is determined for almost every kind of data
- It is finite and not indefinite
- It is readily used in algebraic treatment
- It is least affected by fluctuations of sampling

Cons:

- The arithmetic mean is highly affected by extreme values
- It cannot average the ratios and percentages properly
- It is not an appropriate average for highly skewed distributions
- It cannot be computed accurately if any item is missing
- The mean sometimes does not coincide with any of the observed values

Formula:Mean$(\mathbf{x^-})=\sum fi\sum fixi$

Calculate the mean height for the following data using the direct method.

Height (in inches)	**60 – 62**	**62 – 64**	**64 – 66**	**66 – 68**	**68 – 70**	**70 – 72**
Frequency	**3**	**6**	**9**	**12**	**8**	**2**

As, $\bar{x}=\sum fixi\sum fi\bar{x}=\sum fi\sum fixi$

Height (in inches)	**Frequency(fi)**	**Midpoint (xi)**	**fi × xi**
60 – 62	3	61	183
62 – 64	6	63	378
64 – 66	9	65	585
66 – 68	12	67	804
68 – 70	8	69	552
70 – 72	2	71	142
	∑fi = 40		**∑fi xi = 2644**

⇒ Mean = 2644/40 = 66.1
Thus, mean height is 66.1 inches.

2. Median

The median is that value of the variable which divides the group in two equal parts. One part comprising the values greater than and the other all values less than median. Median of a distribution may be defined as that value of the variable which exceeds and is exceeded by the same number of observation. It is the value such that the number of observations above it is equal to the number of observations below it. Thus we know that the arithmetic mean is based on all items of thedistribution, the median is positional average, i.e., it depends upon the position occupied by a value in the frequency distribution. When the items of a series are arranged in ascending or descending order of magnitude the value of the middle item in the series is known as median in the case of individual observation.

Symbolically, Median = size of nth item

It the number of items is even, and then there is no value exactly in the middle of the series. In such a situation the median is arbitrarily taken to be halfway between the two middle items.

Advantages of Median:

1. It is very simple to understand and easy to calculate. In some cases it is obtained simply by inspection.
2. Median lies at the middle part of the series and hence it is not affected by the extreme values.
3. It is a special average used in qualitative phenomena like intelligence or beauty which are not quantified but ranks are given. Thus we can locate the person whose intelligence or beauty is the average.
4. In grouped frequency distribution it can be graphically located by drawing gives.
5. It is specially useful in open-ended distributions since the position rather than the value of item that matters in median.

Disadvantages of Median:

1. In simple series, the item values have to be arranged. If the series contains large number of items, then the process becomes tedious.
2. It is a less representative average because it does not depend on all the items in the series.
3. It is not capable of further algebraic treatment. For example, we cannot find a combined median of two or more groups if the median of different groups are given.
4. It is affected more by sampling fluctuations than the mean as it is concerned with on1y one item i.e. the middle item.
5. It is not rigidly defined. In simple series having even number of items, median cannot be exactly found. Moreover, the interpolation formula applied in the continuous series is based on the unrealistic assumption that the frequency of the median class is evenly spread over the magnitude of the class interval of the median group.

Median for Ungrouped or Raw data:

Problem:

The number of rooms in the seven five stars hotel in Chennai city is 71, 30, 61, 59, 31, 40 and 29. Find the median number of rooms

Solution:

Arrange the data in ascending order 29, 30, 31, 40, 59, 61, 71

n = 7 (odd)

Median = 7+1 / 2 = 4th positional value

Median = 40 rooms

Median for Discrete grouped data:

Problem:

The following data are the weights of students in a class. Find the median weights of the students

Weight(kg)	10	20	30	40	50	60	70
Number of Students	4	7	12	15	13	5	4

Weight (kg) x	Frequency f	Cumulative Frequency $c.f$
10	4	4
20	7	11
30	12	23
40	15	38
50	13	51
60	5	56
70	4	60
Total	N = 60	

Here, $N = \sum f = 60$

Solution: $\frac{N+1}{2} = 30.5$

The cumulative frequency greater than 30.5 is 38.The value of x corresponding to 38 is 40. The median weight of the students is 40 kgs

Median for Continuous grouped data

In this case, the data is given in the form of a frequency table with class-interval etc., The following formula is used to calculate the median.

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Where

l = Lower limit of the median class

N = Total Numbers of frequencies

f = Frequency of the median class

m = Cumulative frequency of the class preceding the median class

c = the class interval of the median class.

Problem:

From the formula, it is clear that one has to find the median class first. Median class is, that class which correspond to the cumulative frequency just greater than N/2.

The following data attained from a garden records of certain period Calculate the median weight of the apple

Weight in grams	410 – 420	420 – 430	430 – 440	440 – 450	450 – 460	460 – 470	470 – 480
Number of apples	14	20	42	54	45	18	7

Solution:

Weight in grams	Number of apples	Cumulative Frequency
410 – 420	14	14
420 – 430	20	34
430 – 440	42	76
440 – 450	54	130
450 – 460	45	175
460 – 470	18	193
470 – 480	7	200
Total	**N = 200**	

$$\frac{N}{2} = \frac{200}{2} = 100.$$

Median class is 440 – 450

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$l = 440,\ \frac{N}{2} = 100,\ m = 76,\ f = 54,\ c = 10$

$$\text{Median} = 440 + \frac{100 - 76}{54} \times 10$$

$$= 440 + \frac{24}{54} \times 10 = 440 + 4.44 = 444.44$$

The median weight of the apple is 444.44 grams

Problem:

The following table shows age distribution of persons in a particular region:

Age (years)	No. of persons (in thousands)
Below 10	2
Below20	5
Below30	9
Below 40	12
Below 50	14
Below 60	15
Below 70	15.5
Below 80	15.6

Find the median age.

Solution:

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 10, hence the width of the class interval is equal to 10

Age groups	No. of persons (in thousands) f	cf
0 – 10	2	2
10 – 20	3	5
20 – 30	4	9
30 – 40	3	12
40 – 50	2	14
50 – 60	1	15
60 – 70	0.5	15.5
70 – 80	0.1	15.6
Total	**N = 15.6**	

$$\left(\frac{N}{2}\right) = \frac{15.6}{2} = 7.8$$

Median lies in the 20 – 30 age group

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

$$= 20 + \frac{7.8 - 5}{4} \times 10$$

$$\text{Median} = 27 \text{ years}$$

3. Mode

Mode is that value of the variable which occurs or repeats itself maximum number of item. The mode is most "fashionable" size in the sense that it is the most common and typical and is defined by Zizek as "the value occurring most frequently in series of items and around which the other items are distributed most densely." In the words of Croxton and Cowden, the mode of a distribution is the value at the point where the items tend to be most heavily concentrated. According to A.M. Tuttle, Mode is the value which has the greater frequency density in its immediate neighbourhood. In the case of individual observations, the mode is that value which is repeated the maximum number of times in the series. The value of mode can be denoted by the alphabet z also.

Graphic Location of Mode:

Since mode is a positional average it can be located graphically by the following process:

- A histogram of the frequency distribution is drawn.
- In the histogram, the highest rectangle represents the model class.
- The top left corner of the highest rectangle is joined with the top left corner of the following rectangle and the top right corner of the highest rectangle is joined with the top right corner of the preceding rectangle respectively.
- From the point of intersection of both the lines a perpendicular is drawn on the X-axis, and check that point on the X-axis. This will be the required value of mode.

Advantages and Disadvantages of Mode:

Advantages:

- It is easy to understand and simple to calculate.
- It is not affected by extremely large or small values.
- It can be located just by inspection in ungrouped data and discrete frequency distribution.
- It can be useful for qualitative data.
- It can be computed in an open-end frequency table.
- It can be located graphically.

Disadvantages:

- It is not well defined.
- It is not based on all the values.
- It is stable for large values so it will not be well defined if the data consists of a small number of values.
- It is not capable of further mathematical treatment.
- Sometimes the data has one or more than one mode and sometimes the data has no mode at all.

For Ungrouped or Raw Data:

Problem:

The following are the marks scored by 20 students in the class. Find the mode 90, 70, 50, 30, 40, 86, 65, 73, 68, 90, 90, 10, 73, 25, 35, 88, 67, 80, 74, 46

Solution:

Since the marks 90 occurs the maximum number of times, three times compared with the other numbers, mode is 90.

Problem:

A doctor who checked 9 patients' sugar level is given below. Find the mode value of the sugar levels. 80, 112, 110, 115, 124, 130, 100, 90, 150, 180

Solution:

Since each values occurs only once, there is no mode.

Mode for Continuous data:

The mode or modal value of the distribution is that value of the variate for which the frequency is maximum. It is the value around which the items or observations tend to be most heavily concentrated. The mode is computed by the formula.

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

Modal class is the class which has maximum frequency.
$f1$ = frequency of the modal class
$f0$ = frequency of the class preceding the modal class
$f2$ = frequency of the class succeeding the modal class
c = width of the class limits

Remarks

(i) If $(2f_1\text{-}f_0\text{-}f_2)$ comes out to be zero, then mode is obtained by the following formula taking absolute differences $M_0 = l + \left(\frac{(f_1 - f_0)}{|f_1 - f_0| + |f_1 - f_2|} \times C \right)$

(ii) If mode lies in the first class interval, then f_0 is taken as zero.

(iii) The computation of mode poses problem when the modal value lies in the open-ended class.

Problem:

The following data relates to the daily income of families in an urban area. Find the modal income of the families.

Income (`)	0-100	100-200	200-300	300-400	400-500	500-600	600-700
No.of persons	5	7	12	18	16	10	5

Solution:

Income (`)	No.of persons (f)	
0-100	5	
100-200	7	
200-300	12	f_0
300-400	18	f_1
400-500	16	f_2
500-600	10	
600-700	5	

$$\text{Mode} = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C$$

The highest frequency is 18, the modal class is 300-400

Here, $l = 300$, $f_0 = 12$, $f_1 = 18$, $f_2 = 16$,

$$\text{Mode} = 300 + \frac{18 - 12}{2 \times 18 - 12 - 16} \times 100$$

$$= 300 + \frac{6}{36 - 28} \times 100$$

$$= 300 + \frac{6}{8} \times 100$$

$$= 300 + \frac{600}{8} = 300 + 75 = 375$$

The modal income of the families is 375.

II Dispersion:

The word dispersion has a technical meaning in statistics. The average measures the center of the data, and it is one aspect of observation. Another feature of the observation is how the observations are spread about the center. The observations may be close to the center or they may be spread away from the center. If the observations are close to the center (usually the arithmetic mean or median), we say that dispersion, scatter or variation is small. If the observations are spread away from the center, we say dispersion is large.

The study of dispersion is very important in statistical data. If in a certain factory there is consistency in the wages of workers, the workers will be satisfied. But if some workers have high wages and some have low wages, there will be unrest among the low paid workers and they might go on strike and arrange demonstrations. If in a certain country some people are very poor and some are very rich, we say there is economic disparity. This means that dispersion is large.

The idea of dispersion is important in the study of workers' wages, price of commodities, standards of living of different people, distribution of wealth, distribution of land among framers, and many other fields of life. Some brief definitions of dispersion are:

- The degree to which numerical data tend to spread about an average value is called the dispersion or variation of the data.
- Dispersion or variation may be defined as a statistic signifying the extent of the scatteredness of items around a measure of central tendency.
- Dispersion or variation is the measurement of the size of the scatter of items in a series about the average.

Properties of a good measure of Dispersion:

There are certain pre-requisites for a good measure of dispersion:

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each individual item of the distribution.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should not be unduly affected by the extreme items.

For the study of dispersion, we need some measures which show whether the dispersion is small or large. There are two types of measure of dispersion, which are:

(a) Absolute Measures of Dispersion

(b) Relative Measures of Dispersion

(a) Absolute Measures of Dispersion:

These measures give us an idea about the amount of dispersion in a set of observations. They give the answers in the same units as the units of the original observations. When the observations are in kilograms, the absolute measure is also in kilograms. If we have two sets of observations, we cannot always use the absolute measures to compare their dispersions. We shall explain later as to when the absolute measures can be used for comparison of dispersion in two or more sets of data. The absolute measures which are commonly used are:

1. The Range

Range is a simplest method of studying dispersion. It takes lesser time to compute the „absolute" and „relative" range. Range does not take into account all the values of a series, i.e. it considers only the extreme items and middle items are not given any importance. Therefore, Range cannot tell us anything about the character of the distribution. Range cannot be computed in the case of "open ends" distribution i.e., a distribution where the lower limit of the first group and upper limit of the higher group is not given. The concept of range is useful in the field of quality control and to study the variations in the prices of the shares etc.

Problem:

Find the range and coefficient of range of the following data: 25, 67, 48, 53, 18, 39, 44.

Solution Largest value $L = 67$; Smallest value $S = 18$

Range $R = L - S = 67 - 18 = 49$

Coefficient of range $= (L - S) / (L + S)$

Coefficient of range $= (67 - 18) / (67 + 18) = 49/85$

$= 0.576$

2. The Quartile Deviation

The quartile deviation is a slightly better measure of absolute dispersion than the range, but it ignores the observations on the tails. If we take difference samples from a population and calculate their quartile deviations, their values are quite likely to be sufficiently different. This is called sampling fluctuation, and it is not a popular measure of dispersion. The quartile deviation calculated from the sample data does not help us to draw any conclusion (inference) about the quartile deviation in the population.

Advantages of Quartile Deviation:

- It is easy to calculate. We are required simply to find the values of Q1 and Q3 and then apply the formula of absolute and coefficient of quartile deviation.
- It has better results than range method. While calculating range, we consider only the extreme values that make dispersion erratic, in the case of quartile deviation; we take into account middle 50% items.
- The quartile deviation is not affected by the extreme items.

Disadvantages:

- It is completely dependent on the central items. If these values are irregular and abnormal the result is bound to be affected.
- All the items of the frequency distribution are not given equal importance in finding the values of Q1 and Q3.
- Because it does not take into account all the items of the series, considered to be inaccurate.

Quartile Deviation = (Third Quartile – First Quartile) / 2

Quartile Deviation =$(Q_3 – Q_1) / 2$

Quartile deviation can be calculated for both the grouped data and the ungrouped data. Quartile deviation measures the absolute level of dispersion and is not affected by the extreme values. And the relative measure with reference to quartile deviation is known as the coefficient of quartile deviation.

Coefficient of Quartile Deviation = (Q3 – Q1) / (Q3 + Q1)

Problem:

Problem:

Find the quartile deviation and the coefficient of quartile deviation for the following given data.

23, 8, 5, 16, 33, 7, 24, 5, 30, 33, 37, 30, 9, 11, 26, 32

Solution:

The given data points are 23, 8, 5, 16, 33, 7, 24, 5, 30, 33, 37, 30, 9, 11, 26, 32

Let us arrange this data in the following ascending order.

5, 5, 7, 8, 9, 11, 16, 23, 24, 26, 30, 30, 32, 33, 33, 37

From the above data we have $Q_1 = (8 + 9)/2 = 17/2 = 8.5$, and $Q_3 = (30 + 32)/2 = 62/2 = 31$

Quartile Deviation = Q3−Q1/2=31−8.5/2=22.5/2=11.25

Q3−Q1/2=31−8.5/2=22.5/2=11.25

Coefficient of Quartile Deviation = Q3−Q1÷Q3+Q1=31−8.5÷31+8.5=22.5÷39.5=0.57

Therefore, the quartile deviation is 11.25, and the coefficient of quartile deviation is 0.57.

3. The Mean Deviation

Average deviation is defined as a value which is obtained by taking the average of the deviations of various items from a measure of central tendency Mean or Median or Mode, ignoring negative signs. Generally, the measure of central tendency from which the deviations arc taken, is specified in the problem. If nothing is mentioned regarding the measure of central tendency specified than deviations are taken from median because the sum of the deviations (after ignoring negative signs) is minimum. This method is more effective during the reports presented to the general public or to groups who are not familiar with statistical methods.

Steps to Compute Average Deviation:

1. Calculate the value of Mean or Median or Mode
2. Take deviations from the given measure of central-tendency
3. Ignore the negative signs of the deviation.
4. Apply the formula to get Average Deviation about Mean or Median or Mode.

Advantages of Average Deviations

- Average deviation takes into account all the items of a series and hence, it provides sufficiently representative results.
- It simplifies calculations since all signs of the deviations are taken as positive.
- Average Deviation may be calculated either by taking deviations from Mean or Median or Mode.
- Average Deviation is not affected by extreme items.
- It is easy to calculate and understand.
- Average deviation is used to make healthy comparisons.

Disadvantages of Average Deviations

- It is illogical and mathematically unsound to assume all negative signs as positive signs.
- Because the method is not mathematically sound, the results obtained by this method are not reliable.

This method is unsuitable for making comparisons either of the series or structure of the series.

Problem:

Calculate the mean deviation from the median and the co-efficient of mean deviation from the following data:

Marks of the students: 86, 25, 87, 65, 58, 45, 12, 71, 35.

Solution: Arrange the data in ascending order: 12, 25, 35, 45, 58, 65, 71, 86, 87.

Median = Value of the $\frac{n+1}{2}$thterm

= Value of the $\frac{9+1}{2}^{th}$term=58

Calculation of mean deviation:

X	|X–M|
12	46
25	33
35	23
45	13
58	0
65	7
71	13
86	28
87	29
N = 9	∑|X–M|=460

M.D. = $\frac{\sum|X-M|}{N}$

= 4609

= 51.11

Co-efficient of Mean Deviation from Median = $\frac{\text{M.D}}{\text{M}}$

= $\frac{51.11}{58}$

= 0.881

Problem:

Calculate the mean deviation from mean for the following data.

x	12	9	6	18	10
f	7	3	8	1	2

Answer.

x	**f**	**x.f**	**$\|x - \mu\|$**	**f. $\|x - \mu\|$**
12	7	84	2.619	18.33
9	3	27	0.381	1.143
6	8	48	3.381	27.048
18	1	18	8.619	8.619
10	2	20	0.619	1.238
Total	**21**	**197**		**56.378**

We first find the Mean of the given dataset,

$$\text{Mean } (\mu) = \frac{\sum_1^5 f_i x_i}{\sum_1^5 f_i} = \frac{197}{21} = 9.381$$

Finally, we substitute values in the mean deviation about mean formula,

$$\text{Mean Deviation} = \frac{\sum_1^5 f_i |x_i - \mu|}{\sum_1^5 f_i} = \frac{56.378}{21} = 2.684$$

Hence, the mean deviation about the mean is found to be 2.684

4. The Standard Deviation and Variance

The standard deviation, which is shown by greek letter s (read as sigma) is extremely useful in judging the representativeness of the mean. The concept of standard deviation, which was introduced by Karl Pearson has a practical significance because it is free from all defects, which exists in a range, quartile deviation or average deviation. Standard deviation is calculated as the square root of average of squared deviations taken from actual mean. It is also called root mean square deviation. The square of standard deviation i.e., S2 is called „variance".

Calculation of standard deviation in case of raw data:

There are four ways of calculating standard deviation for raw data:

1. When actual values are considered;
2. When deviations are taken from actual mean;
3. When deviations are taken from assumed mean; and
4. When „step deviations" are taken from assumed mean.

Advantages of Standard Deviation:

- Standard deviation is the best measure of dispersion because it takes into account all the items and is capable of future algebraic treatment and statistical analysis.
- It is possible to calculate standard deviation for two or more series.
- This measure is most suitable for making comparisons among two or more series about variability.

Disadvantages:

- It is difficult to compute.
- It assigns more weights to extreme items and less weight to items that are nearer to mean. It is because of this fact that the squares of the deviations which are large in size would be proportionately greater than the squares of those deviations which are comparatively small.

(b) Relative Measures of Dispersion:

These measures are calculated for the comparison of dispersion in two or more sets of observations. These measures are free of the units in which the original data is measured. If the original data is in dollars or kilometer's, we do not use these units with relative measures of dispersion. These measures are a sort of ratio and are called coefficients. Each absolute measure of dispersion can be converted into its relative measure. Thus the relative measures of dispersion are:

1. Coefficient of Range or Coefficient of Dispersion

Coefficient of range

The ratio of the difference between the highest and lowest value in a data set to the sum of those values.

2. Coefficient of Quartile Deviation or Quartile

   The coefficient of quartile deviation (CQD) is a relative measure of how spread out data is in the upper and lower halves of a set. It is calculated by dividing the difference between the first and third quartiles by the mean and the standard deviation. The formula for the coefficient of quartile deviation is:

   Formula

   Coefficient of quartile deviation (CQD)= Q_3 -Q_1 /Mean × (Std. Dev.)

   The quartile deviation is the absolute measure of dispersion, and is calculated by dividing the difference between the first and third quartiles by 2. The formula for quartile deviation is:

   Formula

   Quartile deviation(QD)= (Q3 – Q1)/2

3. Coefficient of Dispersion Coefficient of Mean Deviation or Mean Deviation of Dispersion

   The coefficient of mean deviation is a relative measure of dispersion that is calculated by dividing the mean deviation by the average value from which it is calculated.

   Coefficient of mean deviation=Mean deviation / average value from which it is calculated

4. Coefficient of Standard Deviation or Standard Coefficient of Dispersion

   The coefficient of variation (CV) is a measure that compares the standard deviation of a data set to its mean, while the coefficient of dispersion (COD) is a measure of average deviation from the median:

   Coefficient of variation (CV)

   The CV is the ratio of the standard deviation to the mean, and is used to compare data sets with different units or means.

   Coefficient of dispersion (COD)

   The COD is calculated by taking the difference between the ratio for each data point and the median ratio, adding these differences, and dividing by the number of observations. The result is then divided by the median to express it as a percentage of the median.

5. Coefficient of Variation (a special case of Standard Coefficient of Dispersion)

Coefficient of Variation The most important of all the relative measures of dispersion is the coefficient of variation. This word is variation not variance. There is no such thing as coefficient of variance. Thus CV is the value of SD when mean is assumed equal to 100. It is a pure number and the unit of observation is not mentioned with its value. It is written in percentage form like 20% or 25%. When its value is 20%, it means that when the mean of the observations is assumed equal to 100, their standard deviation will be 20. The C.V is used to compare the dispersion in different sets of data particularly the data which differ in their means or differ in their units of measurement. The wages of workers may be in dollars and the consumption of meat in families may be in kilograms. The standard deviation of wages in dollars cannot be compared with the standard deviation of amount of meat in kilograms. Both the standard deviations need to be converted into a coefficient of variation for comparison.

Suppose the value of C.V for wages is 10% and the values of C.V for kilograms of meat are 25%. This means that the wages of workers are consistent. Their wages are close to the overall average of their wages. But the families consume meat in quite different quantities. Some families consume very small quantities of meat and some others consume large quantities of meat. We say that there is greater variation in their consumption of meat. The observations about the quantity of meat are more dispersed or more variant.

Uses of Coefficient of Variation

- Coefficient of variation is used to know the consistency of the data. By consistency we mean the uniformity in the values of the data/distribution from the arithmetic mean of the data/distribution. A distribution with a smaller C.V than the other is taken as more consistent than the other.
- C.V is also very useful when comparing two or more sets of data that are measured in different units of measurement.

Problem:

Calculate variance and standard deviation for the following data:

x	2	4	6	8	10
f	3	5	9	5	3

Ans:

x	f	fx	D	D^2	fD^2
2	3	6	-4	16	48
4	5	20	-2	4	20
6	9	54	0	0	0
8	5	40	2	4	20
10	3	30	4	16	48
	$\sum$= 25	$\sum$= 150			$\sum$= 136

Mean= $\sum fx/\sum f$= 150/25= 6
Hence, variance= $\sum fD^2/N$= 136/25= 5.44

And standard deviation= $\sqrt{(5.44)}$= 2.33

Independent sample t-test:
The independent samples t-test (also known as the two-sample t-test) is a statistical test used to determine whether there is a significant difference between the means of two independent groups. This test is commonly used in research to compare two different groups on the same variable, especially when trying to see if a treatment or condition has an effect.

Purpose of the Test:
The independent samples t-test is used when:

- You have two groups that are independent (i.e., different subjects in each group).
- You want to compare the means of these two groups.
- You want to know if the difference between the group means is statistically significant.

Assumptions of Independent Samples t-Test:

- Independence of Observations: The two groups should consist of different subjects, and observations should not be related.
- Normality: The data should approximately follow a normal distribution. This assumption becomes less critical if sample sizes are large due to the Central Limit Theorem.
- Homogeneity of Variance: The variance of the two groups should be approximately equal. This can be tested using Levene's Test for Equality of Variances.

Problem:

A researcher wants to know if there is a significant difference in the weight of newborn babies between two hospitals in a city. The researcher randomly selects 20 newborns from Hospital A and 20 newborns from Hospital B and records their weights in pounds. The mean weight for the Hospital A group is 7.5, with a standard deviation of 0.8. The mean weight for the Hospital B group is 7.1, with a standard deviation of 1.2. Is there a significant difference between the two hospitals?

Solution: This is an independent samples t-test problem since the two groups being compared are independent of each other.

Using the formula for the t-value, we get:

$$t = \frac{X_1 - X_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t = (7.5 - 7.1) / (sqrt((0.8^2/20) + (1.2^2/20)))

t = 1.77

Assuming a significance level of 0.05 and df = 38, the critical t-value is 2.024.

Since the calculated t-value of 1.77 is less than the critical t-value of 2.024, we can conclude that there is not a significant difference in the weight of newborn babies between the two hospitals in the city.

Problem:

A professor wants to know if her introductory statistics class has a good grasp of basic math. Six students are chosen at random from the class and given a math proficiency test. The professor wants the class to be able to score above 70 on the test. The six students get scores of 62, 92, 75, 68, 83, and 95. Can the professor have 90 percent confidence that the mean score for the class on the test would be above 70?

Ans. Null hypothesis: $H\,0$: $\mu = 70$

Alternative hypothesis: $H\,a$: $\mu > 70$

First, calculate the mean and standard deviation:

$$\begin{array}{r} 62 \\ 92 \\ 75 \\ 68 \\ 83 \\ +\ 95 \\ \hline 475 \end{array} \qquad \bar{x} = \frac{475}{6} = 79.17$$

$$s = 13.17$$

Next, calculate the *t*-value:

$$t = \frac{79.17 - 70}{\frac{13.17}{\sqrt{6}}} = \frac{9.17}{5.38} = 1.71$$

For testing the hypothesis, the computed *t*-value of 1.71 needs to be compared with the critical value in the *t*-table. If the calculated *t*-value is larger than the critical *t*-value from the table, the null hypothesis will be discarded.

Problem:

A Little League baseball coach wants to know if his team is representative of other teams in scoring runs. Nationally, the average number of runs scored by a Little League team in a game is 5.7. He chooses five games at random in which his team scored 5, 9, 4, 11, and 8 runs. Is it likely that his team's scores could have come from the national distribution? Assume an alpha level of 0.05.

Ans. Null hypothesis: $H\,0$: $\mu = 5.7$

Alternative hypothesis: $H\,a$: $\mu \neq 5.7$

Now calculate the mean and standard deviation:

$$\begin{array}{r} 5 \\ 9 \\ 4 \\ 11 \\ +\ 8 \\ \hline 37 \end{array} \qquad \bar{x} = \frac{37}{5} = 7.4$$

$$s = 2.88$$

The *t*-value will be:

$$t = \frac{7.4 - 5.7}{\frac{2.88}{\sqrt{5}}} = \frac{1.7}{1.29} = 1.32$$

Now, check the critical value from the *t*-table. The tabled value for $t\,.025,4$is 2.776. The calculated *t* of 1.32 is smaller, so we cannot reject the null hypothesis that is the mean of this team is equal to the population mean. The coach cannot deduce that his team is different from the national distribution of runs scored.

Problem:

The water diet requires you to drink 2 cups of water every half hour from when you get up until you go to bed but eat anything you want. Four adult volunteers agreed to test this diet. They are weighed prior to beginning the diet and 6 weeks after. Their weights in pounds are

Person	1	2	3	4	mean	S.d.
Weight before	180	125	240	150	173.75	49.56
Weight after	170	130	215	152	166.75	36.09
Difference	10	-5	25	-2	7	13.64

Conduct a one-sample t-test using the difference with the following hypotheses: H0: Diff = 0 Ha: Diff ≠ 0 Report the test statistic with the P-value, then summarize your conclusion.

Ans. Hypotheses:

H0: Diff = 0 (no difference -- there is no difference at all in the mean of the weight difference)

Ha: Diff ≠ 0 (difference – diet made difference in the means of the weight differences)

Test statistic:

From the data, we know $\overline{Diff} = 7$ and $s_{Diff} = 13.64$. Then we get

$$t = \frac{\overline{Diff} - \mu_0}{\frac{s_{Diff}}{\sqrt{n}}} = \frac{7 - 0}{\frac{13.64}{\sqrt{4}}} = \frac{7}{6.82} = 1.026.$$

P-value:

Because n is equal to 4, we use the t distribution with df equals 3 to obtain the probability. According to given table, for t = 1.026 for df = 3, the probability is between 0.15 and 0.20. Since this is a two-sided test, P-value should be between 0.30 and 0.40.

Conclusion:

Because the values between 0.30 and 0.40 are less than 0.05, we fail to discard the null hypothesis at the 0.05 significance level. We do not have enough confirmation to deduce that the water diet has an impact on the weight.

What is Bivariate Analysis?

Bivariate analysis is an analysis of two variables to determine the relationships between them. They are often reported in quality of life research. It is one of the simplest forms of quantitative (statistical) analysis. It involves the analysis of two variables (it is often denoted as X, Y), for the purpose of determining the empirical relationship between them.

Bivariate analysis is extremely helpful in testing simple hypotheses of association. It is very helpful in determining to what extent it becomes easier to know and predicts a value for one variable (possibly a dependent variable) if the value of the other variable (possibly the independent variable) is known (see also correlation and simple linear regression). There can be a contrast between bivariate analysis and univariate analysis in which only one variable is analyzed. Both univariate analysis and bivariate analysis can be descriptive or inferential. We can say, it is the analysis of the relationship between the two variables. Bivariate analysis is a simple (two-variable) and special case of multivariate analysis (where simultaneously multiple relations between multiple variables are examined).

Bivariate analysis can be defined as the analysis of bivariate data. It is one of the simplest forms of statistical analysis, which is used to find out if there is a relationship between two sets of values. Usually, it involves the variables X and Y.

The univariate analysis involves an analysis of one (“uni”) variable.

The bivariate analysis involves the analysis of exactly two variables.

The multivariate analysis involves the analysis of more than two variables.

The results we get from the bivariate analysis can be stored in a two-column data table. For example, you might be eager to find out the relationship between caloric intake and weight (of course, the two are related very strongly). Caloric intake will be your independent variable, X, and weight will be your dependent variable, Y.Caloric Intake XWeight Y3500250 lbs2000225 lbs1500110 lbs2250145 lbs4500380 lbs

Bivariate analysis and two sample data analyses are not the same. With two sample data analysis (like a two-sample is a test in Excel), X and Y are not directly related and there will also be a different number of data values in each sample. With bivariate analysis, there is a Y value for each X. For example, suppose you had a caloric intake of 3,000 calories per day and a weight of 300lbs. You will have to write that with the x-variable followed by the y-variable: (3000,300).

Types of bivariate analysis

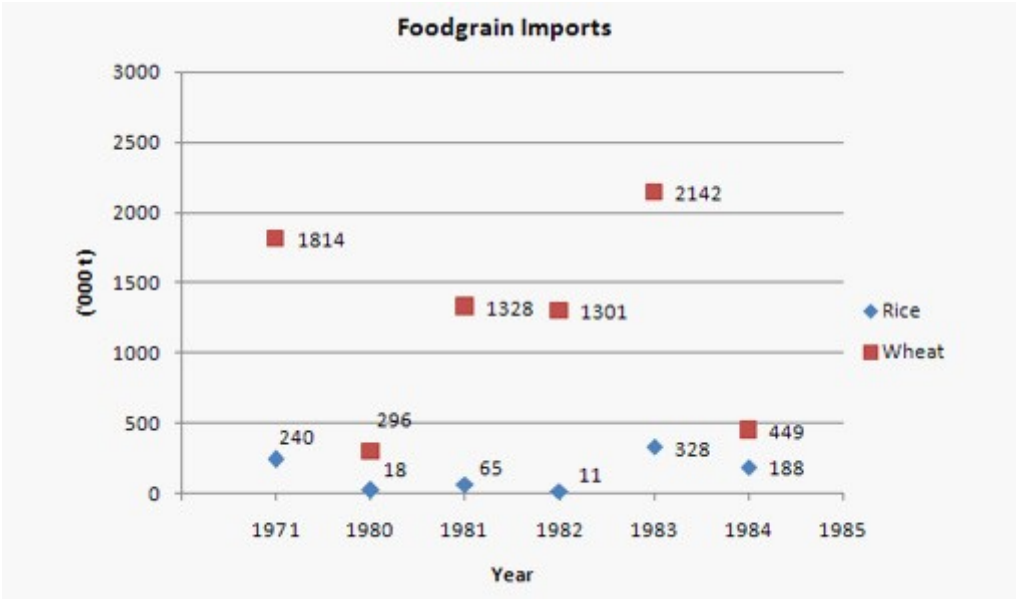
Many kinds of bivariate analysis can be used to determine how two variables are related. Here are some of the most common types.

1. Scatterplots

A scatterplot is a graph that shows how two variables are related to each other. It shows the values of one variable on the x-axis and the values of the other variable on the y-axis.

Eg:

The pattern shows what kind of relationship there is between the two variables and how strong it is.



The total values of food grains (rice and wheat) imported during these years are given below:

1971	:	Rs.	123	crore
1980	:	Rs.	80	crore
1981	:	Rs.	314	crore
1982	:	Rs.	295	crore
1983	:	Rs.	587	crore

1984 : Rs. 158 crore

1. Wheat formed what percent of the volume of total imports of food grains from 1980-84?

1. 75 %

2. 66 %

3. 90 %

4. 95 %

Solution: Option 3
Adding up the approximate quantity of wheat and dividing it by the total quantity of food grains (i.e. wheat + rice) will give us ≈ 5516 / 6126 » 90 %.

2. If the import price of wheat was Rs. 2,400 per tonne in 1983, then what was the import price of rice per tonne during that year?

1. 3,200

2. 2,225

3. 2,850

4. 1,800

Solution: Option 2
In the year 1983, 2,142,000 tonnes of wheat @ Rs 2,400/tonne will mean an expenditure of Rs 514 crore.
So the remaining is 587 – 514 = 73 crore, which was spent on importing 328,000 tonnes of rice. So the price of rice = 730,000 / 328 ≈ Rs 2,225 /tonne

3. In which year was the ratio of rice to wheat imports the highest?

1. 1971

2. 1980

3. 1983

4. 1984

Solution: Option 4
The ratio is maximum when the numerator is maximum and the denominator is minimum. In the year 1984 the imports of rice are highest and the imports of wheat is the least which gives the maximum ratio.

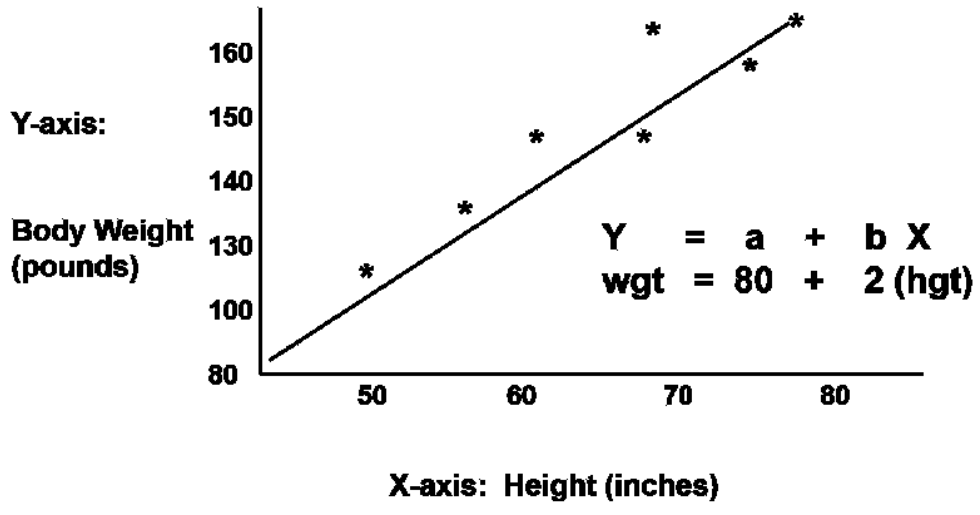
2. Simple Correlation

Simple correlation is a measure used to determine the strength and the direction of the relationship between two variables, X and Y. A simple correlation coefficient can range from –1 to 1. However, maximum (or minimum) values of some simple correlations cannot reach unity (i.e., 1 or –1).

Yield of paddy and the use of fertilizers is an example of simple correlation as yield of paddy depends on the use of fertilizers i.e. presence of one variable affects another variable.

3. Simple Regression:

Regression analysis makes use of mathematical models to describe relationships. For example, suppose that height was the only determinant of body weight. If we were to plot height (the independent or 'predictor' variable) as a function of body weight (the dependent or 'outcome' variable), we might see a very linear relationship, as illustrated below.



We could also describe this relationship with the equation for a line, Y = a + b(x), where 'a' is the Y-intercept and 'b' is the slope of the line. We could use the equation to predict weight if we knew an individual's height. In this example, if an individual was 70 inches tall, we would predict his weight to be:

Weight = 80 + 2 x (70) = 220 lbs.

In this simple linear regression, we are examining the impact of one independent variable on the outcome. If height were the only determinant of body weight, we would expect that the points for individual subjects would lie close to the line. However, if there were other factors (independent variables) that influenced body weight besides height (e.g., age, calorie intake, and exercise level), we might expect that the points for individual subjects would be more loosely scattered around the line, since we are only taking height into account.

4. Chi-square test

The chi-square test is a statistical method for identifying disparities in one or more categories between what was expected and what was observed. The test's primary premise is to assess the actual data values to see what would be expected if the null hypothesis was valid.

Chi-Square Test

Formula

The chi-squared test is done in order to check the difference that is present between the observed value and expected value. The chi-square formula can be written as;

$$\chi_c{}^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here the subscript "c" represents the degree of freedom. "O" represents the observed value and "E" represents the expected value. To find a critical value of chi-square by hand this formula is used rarely. The summation symbol in the formula represents performing a calculation for each and every single data sample in the data set. As one can probably imagine, the calculations can be very lengthy and tedious. Probably instead of this, we want to use technology:

- Chi-Square Test in SPSS.
- Chi-Square P-Value in Excel.

A chi-square statistic is one of the ways to represent the relationship between two categorical variables. A chi-squared statistic is found to be a single number that provides the difference that exists between observed counts and the expected count.

There are a few variations that are found in the chi-square statistic. Which one to use depends upon the method of the data collected and the hypothesis that is being tested. However, all of these variations use the same idea of comparing the expected values with the values that are collected actually. One of the most common forms that can be used for contingency tables:

$$c^2 = \sum_{i=1}^{k} [\frac{(O_i - E_i)^2}{E_i}]$$

Where “O” is represented as the observed value, “E” is represented as the expected value, and “i” is the “ith” position in the contingency table.

If a chi-square value is less then there will be a high correlation found between two sets of data. In theory, if there is no difference between observed and expected values then the chi-square value would be zero, this means it is an event that does not happen in real life. It is not that easy to decide whether a chi-square test statistic is enough to indicate a statistically significant difference. Take the calculated chi-square value and compare it with the critical value found in a chi-square table. If the calculated chi-square value is more when compared to the critical value, then there is a significant difference.

P-value	Description	Hypothesis Interpretation
P−value≤0.05P−value≤0.05	It suggests that the null hypothesis is highly improbable.	Rejected
P−value>0.05P−value>0.05	It strongly suggests that the null hypothesis is highly probable.	Accepted or it “fails to reject”.
P−value>0.05P−value>0.05	The P-value is close to the threshold and is deemed as borderline.	The hypothesis needs more attention.

Properties of chi-square test

The following are the important properties of the chi-square test:

- Variance is equal to twice the number of degrees of freedom.

- The degree of freedom is equal to the mean distribution.
- The chi-square distribution curve is found to approach the normal distribution curve when there is an increase in the degree of freedom.

PROBLEM:

The dealership expected an equal distribution of colour preferences (50 for each colour). Calculate the Chi-Square value to determine if colour preference significantly deviates from the expected distribution.

Solution.

Using the Chi-Square Formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\Rightarrow \chi^2 = \frac{(50-50)^2}{50} + \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(50-50)^2}{50}$$

$$\Rightarrow \chi^2 = \frac{0}{50} + 1.2 + 2 + \frac{0}{50}$$

$$\Rightarrow \chi^2 = 0 + 1.2 + 2 + 0$$

$$\Rightarrow \chi^2 = 3.2$$

Therefore, the chi square value is 3.2.

Problem:

As per the survey on cars owned by each family in the locality, the data has been arranged in the following table:

Number of cars	Oi	Ei
One car	30	25.6
Two cars	14	15
Three cars	6	5.2
Total	50	

Solution.

Below is the presented table:

Number of cars	Oi	Ei	(Oi−Ei)2	$\frac{(O_i - E_i)^2}{E_i}$
One car	30	25.6	19.36	0.645
Two cars	14	15	1.21	0.086
Three cars	6	5.2	0.64	0.106
Total	50			0.837

Therefore, $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 0.837$.

Problem:

What conclusion should be made with respect to an experiment when the significance level is 0.05 (p=0.05)?

Solution.

Since the pp-value of 0.0680.068 exceeds 0.050.05, the null hypothesis cannot be rejected.

Since the p-value is greater than 0.050.05 (p>0.05p>0.05), the null hypothesis fails to be rejected.

Problem:

Calculate the Chi-square value for the following data of incidences of water-borne diseases in three tropical regions.

	India	Equador	South America	Total
Typhoid	31	14	45	90
Cholera	2	5	53	60
Diarrhoea	53	45	2	100
	86	64	100	250

Solution.

Below is the presented table:

Observed	Expected	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i-E_i)^2}{E_i}$
31	30.96	0.04	0.0016	0.0000516
14	23.04	9.04	81.72	3.546
45	36.00	9.00	81.00	2.25
2	20.64	18.64	347.45	16.83
5	15.36	10.36	107.33	6.99
53	24.00	29.00	841.00	35.04
53	34.40	18.60	345.96	10.06
45	25.60	19.40	376.36	14.70
2	40.00	38.00	1444.00	36.10
				$\sum \frac{(O_i-E_i)^2}{E_i} = 125.516$

Therefore, $\chi^2 = \sum \frac{(O_i-E_i)^2}{E_i} = 125.516.$

Paired T-Test

A paired t-test (also known as a dependent or correlated t-test) is a statistical test that compares the averages/means and standard deviations of two related groups to determine if there is a significant difference between the two groups.

A significant difference occurs when the differences between groups are unlikely to be due to sampling error or chance.

The groups can be related by being the same group of people, the same item, or being subjected to the same conditions.

For example, let us assume two paired sets, such as Xi and Yi for i=1,2,…,n such that their paired difference is independent which is identically and normally distributed. Then the paired t-test concludes whether they notably vary from each other.

Hypotheses of a Paired T-Test:

There are two possible hypotheses in a paired t-test.

- The **null hypothesis** (H0) states that there is no significant difference between the means of the two groups.
- The **alternative hypothesis**(Ha) states that there is a significant difference between the two population means, and that this difference is unlikely to be caused by sampling error or chance.

Assumptions of a Paired T-Test:
The assumptions for a paired t-test are given below:

- The dependent variable is normally distributed.
- The observations are sampled independently.
- The dependent variable is measured on an incremental level, such as ratios or intervals.
- The independent variables must consist of two related groups or matched pairs.

Paired T-Test Formula

Paired t-test is a test which is based on the differences between the values of a single pair, i.e., one deducted from the other. In the formula for a paired t-test, this difference is denoted by 'd'. The formula of the paired t-test is defined as the sum of the differences of each pair divided by the square root of n times the sum of the differences squared minus the sum of the squared differences, overall n−1

The formula for the paired t-test is given by

$$t = \frac{\sum d}{\sqrt{\frac{n(\sum d^2) - (\sum d)^2}{n-1}}}.$$

Here, $\sum d$ is the sum of the differences.

.

Paired T-Test Table

Paired T-test table enables the t-value from a t-test to be converted to a statement about significance. The table is given below:

Two Tailed Significance						
Degree of freedom (n-1)	**α=0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**
1	3.078	6.314	12.706	31.821	63.657	318.300
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.305	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733

1.What conclusion should be made with respect to an experiment when the significance level is 0.068?

Solution: Since the p-value of 0.068 is greater than $\alpha = 0.05$, it would fail to reject the null hypothesis.

As the value of $p < 0.05$, the null hypothesis is rejected.

2.In which of the following cases would you use a paired-sample t-test?
(a). When comparing the same participant's performance before and after training.
(b). When comparing two separate groups of people.

Solution: A T-test can be used in making observations on the same sample before and after an event.

In option (b) the data does not involve observations before and after an event for the same set of people.

Thus, the correct answer is (a): when comparing the same participant's performance before and after training.

ANOVA Test

ANOVA Test is used to analyze the differences among the means of various groups using certain estimation procedures. ANOVA means analysis of variance. ANOVA test is a statistical significance test that is used to check whether the null hypothesis can be rejected or not during hypothesis testing.

An ANOVA test can be either one-way or two-way depending upon the number of independent variables. In this article, we will learn more about an ANOVA test, the one-way ANOVA and two-way ANOVA, its formulas and see certain associated examples.

What is ANOVA Test?

ANOVA test, in its simplest form, is used to check whether the means of three or more populations are equal or not. The ANOVA test applies when there are more than two independent groups. The goal of the ANOVA test is to check for variability within the groups as well as the variability among the groups. The ANOVA test statistic is given by the f test.

ANOVA Test Definition

ANOVA test can be defined as a type of test used in hypothesis testing to compare whether the means of two or more groups are equal or not. This test is used to check if the null hypothesis can be rejected or not depending upon the statistical significance exhibited by the parameters. The decision is made by comparing the ANOVA test statistic with the critical value.

ANOVA Test Example

Suppose it needs to be determined if consumption of a certain type of tea will result in a mean weight loss. Let there be three groups using three types of tea - green tea, earl grey tea, and jasmine tea. Thus, to compare if there was any mean weight loss exhibited by a certain group, the ANOVA test (one way) will be used.

Suppose a survey was conducted to check if there is an interaction between income and gender with anxiety level at job interviews. To conduct such a test a two-way ANOVA will be used.

ANNOVA FORMULA:

Source of Variation	**Sum of Squares**	**Degree of Freedom**	**Mean Squares**	**F Value**
Between Groups	SSB = Σnj(X̄j– X̄)2	df1 = k – 1	MSB = SSB / (k – 1)	f = MSB / MSE or, F = MST/MSE
Error	SSE = Σnj(X̄- X̄j)2	df2 = N – k	MSE = SSE / (N – k)	

Source of Variation	Sum of Squares	Degree of Freedom	Mean Squares	F Value
Total	SST = SSB + SSE	df3 = N – 1		

where,

- F = ANOVA Coefficient
- MSB = Mean of the total of squares between groupings
- MSW = Mean total of squares within groupings
- MSE = Mean sum of squares due to error
- SST = total Sum of squares
- p = Total number of populations
- n = The total number of samples in a population
- SSW = Sum of squares within the groups
- SSB = Sum of squares between the groups
- SSE = Sum of squares due to error
- s = Standard deviation of the samples
- N = Total number of observations

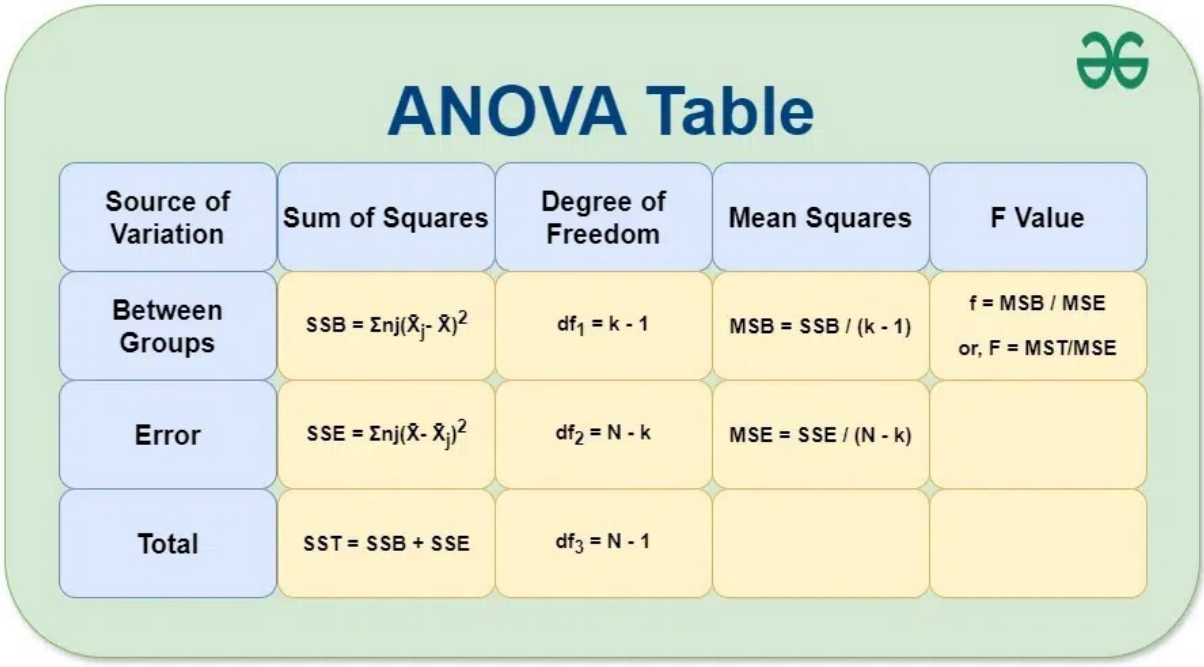
Examples of the use of ANOVA Formula

Assume it is necessary to assess whether consuming a specific type of tea will result in a mean weight decrease. Allow three groups to use three different varieties of tea: green tea, Earl Grey tea, and Jasmine tea. Thus, the ANOVA test (one way) will be utilized to examine if there was any mean weight decrease displayed by a certain group.

Assume a poll was undertaken to see if there is a relationship between salary and gender and stress levels during job interviews. A two-way ANOVA will be utilized to carry out such a test.

ANOVA Table

An ANOVA (Analysis of Variance) test table is used to summarize the results of an ANOVA test, which is used to determine if there are any statistically significant differences between the means of three or more independent groups. Here's a general structure of an ANOVA table:



Types of ANOVA Formula

One-Way ANOVA

This test is used to see if there is a variation in the mean values of three or more groups. Such a test is used where the data set has only one independent variable. If the test statistic exceeds the critical value, the null hypothesis is rejected, and the averages of at least two different groups are significant statistically.

Two-Way ANOVA

Two independent variables are used in the two-way ANOVA. As a result, it can be viewed as an extension of a one-way ANOVA in which only one variable influences the dependent variable. A two-way ANOVA test is used to determine the main effect of each independent variable and whether there is an interaction effect. Each factor is examined independently to determine the main effect, as in a one-way ANOVA. Furthermore, all components are analyzed at the same time to test the interaction impact.

Solved Examples on ANOVA Formula

Example 1: Three different kinds of food are tested on three groups of rats for 5 weeks. The objective is to check the difference in mean weight(in grams) of the rats per week. Apply one-way ANOVA using a 0.05 significance level to the following data:

Food I	Food II	Food III
8	4	11
12	5	8

Food I	Food II	Food III
19	4	7
8	6	13
6	9	7
11	7	9

Solution:

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1: The means are not equal

Since, $\bar{X}_1 = 5, \bar{X}_2 = 9, \bar{X}_3 = 10$

Total mean $= \bar{X} = 8$

$SSB = 6(5-8)^2 + 6(9-8)^2 + 6(10-8)^2 = 84$

$SSE = 68$

$MSB = SSB/df_1 = 42$

$MSE = SSE/df_2 = 4.53$

$f = MSB/MSE = 42/4.53 = 9.33$

Since f > F, the null hypothesis stands rejected.

Problem:

Three types of fertilizers are used on three groups of plants for 5 weeks. We want to check if there is a difference in the mean growth of each group. Using the data given below apply a one way ANOVA test at 0.05 significant level.

Fertilizer 1	Fertilizer 2	Fertilizer 3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Solution:

$H_0: \mu_1 = \mu_2 = \mu_3$

H_1: The means are not equal

Fertilizer 1	Fertilizer 2	Fertilizer 3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12
$\bar{X}_1 = 5$	$\bar{X}_1 = 9$	$\bar{X}_1 = 10$

Total mean, $\overline{\overline{X}} = 8$

$n_1 = n_2 = n_3 = 6, k = 3$

$SSB = 6(5 - 8)^2 + 6(9 - 8)^2 + 6(10 - 8)^2$

$= 84$

$df_1 = k - 1 = 2$

Fertilizer 1	$(X - 5)^2$	Fertilizer 2	$(X - 9)^2$	Fertilizer 3	$(X - 10)^2$
6	1	8	1	13	9
8	9	12	9	9	1
4	1	9	0	11	1
5	0	11	4	8	4
3	4	6	9	7	9
4	1	8	1	12	4
$\overline{X}_1 = 5$	Total = 16	$\overline{X}_1 = 9$	Total = 24	$\overline{X}_1 = 10$	Total = 28

SSE = 16 + 24 + 28 = 68

N = 18

$df_2 = N - k = 18 - 3 = 15$

$MSB = SSB / df_1 = 84 / 2 = 42$

$MSE = SSE / df_2 = 68 / 15 = 4.53$

ANOVA test statistic, f = MSB / MSE = 42 / 4.53 = 9.33

Using the f table at $\alpha\alpha = 0.05$ the critical value is given as F(0.05, 2, 15) = 3.68

As f > F, thus, the null hypothesis is rejected and it can be concluded that there is a difference in the mean growth of the plants.

Answer: Reject the null hypothesis

Problem:

A trial was run to check the effects of different diets. Positive numbers indicate weight loss and negative numbers indicate weight gain. Check if there is an average difference in the weight of people following different diets using an ANOVA Table.

Low Fat	Low Calorie	Low Protein	Low Carbohydrate
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

Solution:

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: The means are not equal

Low Fat	$(X - 6.6)^2$	Low Calorie	$(X - 3)^2$	Low Protein	$(X - 3.4)^2$	Low Carbohydrate	$(X - 1.2)^2$
8	2	2	1	3	0.2	2	0.6
9	5.8	4	1	5	2.6	2	0.6
6	0.4	3	0	4	0.4	-1	4.8
7	0.2	5	4	2	2	0	1.4

$\bar{\bar{X}}_1$ = 6.6	Tota l = 21.4	$\bar{\bar{X}}_2$ = 3	Tota l = 10	$\bar{\bar{X}}_3$ = 3.4	Tota l = 5.4	$\bar{\bar{X}}_4$ = 1.2	Tota l = 10.6

Total mean, $\bar{\bar{X}} = 3.6$

$n_1 = n_2 = n_3 = n_4 = 5, k = 4$

$SSB = n_1(\bar{\bar{X}}_1 - \bar{\bar{X}})^2 + n_2(\bar{\bar{X}}_2 - \bar{\bar{X}})^2 +\& n_3(\bar{\bar{X}}_3 - \bar{\bar{X}})^2 + n_4(\bar{\bar{X}}_4 - \bar{\bar{X}})^2$

$= 75.8$

$SSE = 21.4 + 10 + 5.4 + 10.6 = 47.4$

The ANOVA Table can be constructed as follows:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma n_j(\bar{\bar{X}}_j - 3.6)^2 = 75.8$	$df_1 = k - 1 = 4 - 1 = 3$	$MSB = SSB / (k - 1) = 25.3$	$f = MSB / MSE = 8.43$
Error	$SSE = \Sigma\Sigma(X - \bar{\bar{X}}_j)^2 = 47.4$	$df_2 = N - k = 20 - 4 = 16$	$MSE = SSE / (N - k) = 3$	

Total	SST = SSB + SSE = 123.2	df_3 = N - 1 = 19		

As no significance level is specified, $\alpha = 0.05$ is chosen.

$F(0.05, 3, 16) = 3.24$

As $8.43 > 3.24$, thus, the null hypothesis is rejected and it can be concluded that there is a mean weight loss in the diets.

Answer: Reject the null hypothesis

Mann-Whitney U Test

The Mann-Whitney U Test, also known as the Wilcoxon Rank Sum Test, is a non-parametric statistical test used to compare two samples or groups.

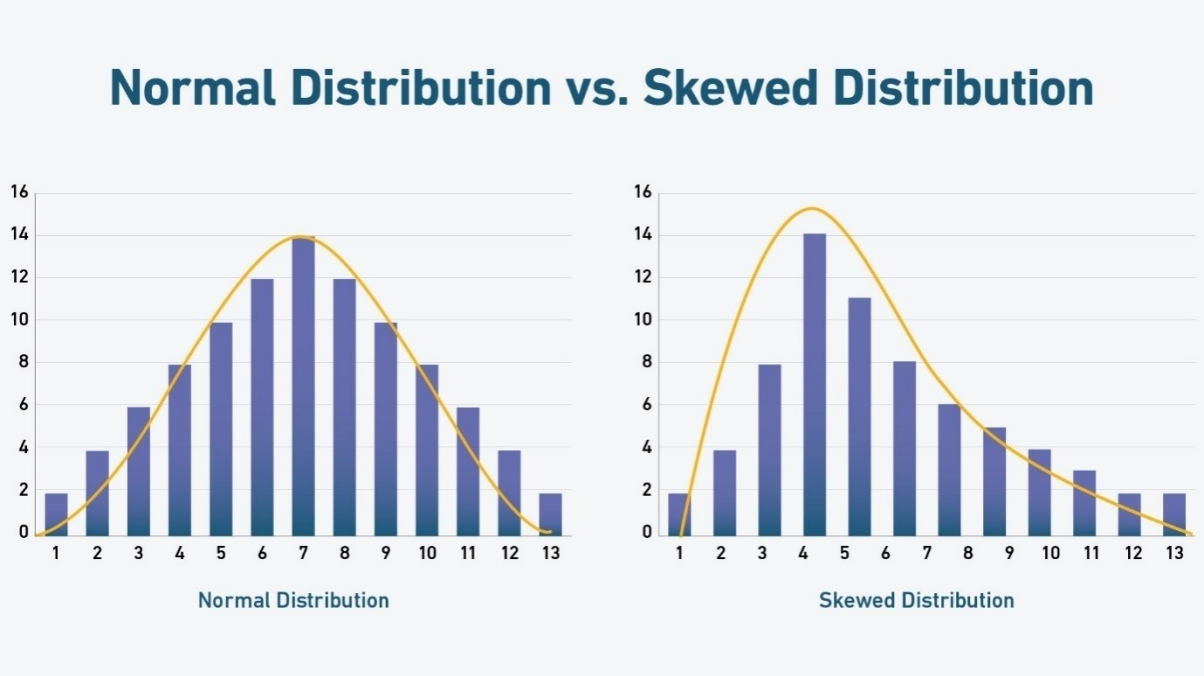
The Mann-Whitney U Test assesses whether two sampled groups are likely to derive from the same population, and essentially asks; do these two populations have the same shape with regards to their data? In other words, we want evidence as to whether the groups are drawn from populations with different levels of a variable of interest. It follows that the hypotheses in a Mann-Whitney U Test are:

- The null hypothesis (H0) is that the two populations are equal.
- The alternative hypothesis (H1) is that the two populations are not equal.

Some researchers interpret this as comparing the medians between the two populations (in contrast, parametric tests compare the means between two independent groups). In certain situations, where the data are similarly shaped (see assumptions), this is valid – but it should be noted that the medians are not actually involved in calculation of the Mann-Whitney U test statistic. Two groups could have the same median and be significantly different according to the Mann-Whitney U test.

When to use the Mann-Whitney U Test

Non-parametric tests (sometimes referred to as 'distribution-free tests') are used when you assume the data in your populations of interest do not have a Normal distribution. You can think of the Mann Whitney U-test as analogous to the unpaired Student's t-test, which you would use when assuming your two populations are normally distributed, as defined by their means and standard deviation (the parameters of the distributions).



The Mann-Whitney U Test is a common statistical test that is used in many fields including economics, biological sciences and epidemiology. It is particularly useful when you are assessing the difference between two independent groups with low numbers of individuals in each group (usually less than 30), which are not normally distributed, and where the data are continuous. If you are interested in comparing more than two groups which have skewed data, a Kruskal-Wallis One-Way analysis of variance (ANOVA) should be used.

Mann-Whitney U Test Assumptions

Some key assumptions for Mann-Whitney U Test are detailed below:

- The variable being compared between the two groups must be continuous (able to take any number in a range – for example age, weight, height or heart rate). This is because the test is based on ranking the observations in each group.
- The data are assumed to take a non-Normal, or skewed, distribution. If your data are normally distributed, the unpaired Student's t-test should be used to compare the two groups instead.
- While the data in both groups are not assumed to be Normal, the data are assumed to be similar in shape across the two groups.
- The data should be two randomly selected independent samples, meaning the groups have no relationship to each other. If samples are paired (for example, two measurements from the same group of participants), then a paired samples t-test should be used instead.
- Sufficient sample size is needed for a valid test, usually more than 5 observations in each group.

Mann-Whitney U Test Example

Consider a randomized controlled trial evaluating a new anti-retroviral therapy for HIV. A pilot trial randomly assigned participants to either the treated or untreated groups

(N=14). We want to assess the viral load (quantity of virus per milliliter of blood) in the treated versus the untreated groups. In practice, a Mann-Whitney U Test would be easily and quickly calculated using statistical software such as SPSS or Stata, but the steps are laid out below.

The data are shown below:

Treated	540	670	1000	960	1200	4650	4200
Untreated	5000	4200	1300	900	7400	4500	7500

These data are both skewed with a sample size of n=7 in each treatment arm, and so a non-parametric test is appropriate. Before we calculate the test, we choose a significance level (usually α=0.05). The first step is to assign ranks to the values from the full sample (both treatment groups pooled together) in order from smallest to largest. We can then generate a test statistic based on the ranks.

The table below shows the viral load values in the treated and untreated groups ranked smallest to largest, along with the summed ranks of each group:

Viral load (Treated)	Viral load (Untreated)	Rank (Treated)	Rank (Untreated)
540		1	
670		2	
	900		3
960		4	
1000		5	
1200		6	
	1300		7
4200		8	
	4500		9
4650		10	
	5000		11
	6100		12
	7400		13
	7500		14
		R_1=36	**R_2=69**

After summing the ranks for each group, the Mann-Whitney U test statistic is selected as the smallest of the two following calculated U values:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_1 = 7 \text{ x } 7 + \frac{7 \text{ x } 8}{2} - 36$$

$$= 41$$

$$U_2 = 7 \text{ x } 7 + \frac{7 \text{ x } 8}{2} - 69$$

$$= 8$$

Where we let 1 denote the treated group and 2 denote the untreated group (denotation of groups is arbitrary), where n1 and n2 are the number of participants and where R1 and R2 are the sums of the ranks in the treated and untreated groups, respectively. In this example, U1=41 and U2=8. We therefore select U=8 as the test statistic.

Normal approximation

There are situations where the sample size may be too large for the reference table to be used to calculate the exact probability distribution – in which case we can use a Normal approximation instead. Since U is found by adding together independent, similarly distributed random samples, the central limit theorem applies when the sample is large (usually >20 in each group). The standard deviation of the sum of the ranks can be used to generate a z-statistic and a significance value generated this way. If the null hypothesis is true, the distribution of U approximates to a Normal distribution.

Next we determine a 'critical value' of U with which to compare our calculated test statistic, which we can do using a reference table of critical values and using our sample sizes (n=7 in both groups) and two-sided level of significance (α=0.05).

In our current example, the critical value can be determined from the reference table as 8. Finally, we can use this to accept or reject the null hypothesis using the following decision rule: Reject H0 if $U \leq 8$.

Given that our U statistic is equal to the critical value, we can reject the null hypothesis that the two groups are equal and accept the alternative hypothesis that there is evidence of a difference in viral load between the groups treated with the new therapy versus untreated.

Wilcoxon Signed Rank Test

Wilcoxon Signed Rank Test: When we cannot be sure that the data is normally distributed we use a non-parametric statistical test for paired data called the Wilcoxon Signed Rank Test. Similar to the student's t-test when we know the data follows a normal distribution.

Examples of How to Conduct a Wilcoxon Signed Rank Test

Example 1

A high school track and field coach is interested in determining if his new training program will improve his athletes 400 meter sprint time and wishes to conduct an analysis. He obtains the permission of 7 athletes to record their sprint times both before and after his new training program which are shown below. According to a Wilcoxon signed-ranks test, at a 95% confidence level, do these scores provide evidence of an increase in median sprint time?

Athlete	Sprint time before (seconds)	Sprint time after (seconds)
1	63	58
2	61	57
3	62	59
4	58	57
5	59	58
6	56	55
7	61	55

Step 1 : Calculate the differences between the measurements. (before - after)

From the table we take the athlete times **before** the coaches new training program and subtract off the **after**.

Athlete	Sprint time before (seconds)	Sprint time after (seconds)	Differences (before - after)
1	63	58	5
2	61	57	4
3	62	59	3
4	58	57	1
5	59	58	1
6	56	55	1
7	61	55	6

Step 2: Order the differences and assign a rank to each difference. If the difference is negative, multiply the rank by -1. In the case of tied ranks, calculate the average rank of the tied.

Note that for this example all of our differences are positive. And because there are 3 similar differences of 1. We calculate the average rank of the three. i.e., $\frac{(1+2+3)}{3} = 2$

Ordered Differences	Ranks

Ordered Differences	Ranks
1	2
1	2
1	2
3	4
4	5
5	6
6	7

Step 3 : Sum up the rank of both the negative and positive differences.

Differences (before - after)	Ranks	Postive	Negative
1	1	1	
1	1	1	
1	1	1	
3	4	4	
4	5	5	
5	6	6	
6	7	7	
		Tot = 25	Tot = 0

We find that the sum of the positive ranks is 2 + 2 + 2 + 4 + 5 + 6 + 7 = 28 and the sum of negative ranks is 0.

Step 4: Find the critical value by using a Table of critical values for W and use the Wilcoxon (W) test statistic to make a conclusion. We find the W-statistic by using the absolute value of the smaller of the two sums.

The given level of significance is alpha = 0.05 and the sample size is n=7. Using a standard W-table of critical values, we find the value to be 4.

Our W-score is the smallest of the two sums, 28 and 0. Because 0 is smaller than 28, we use this as our test-statistic.

Comparing this score to our critical value, it is clear that 0 <4. Hence, we can reject our null hypothesis and conclude that the median difference of the coaches new program is positive.

Kruskal Wallis Test:

It is a nonparametric test. It is sometimes referred to as One-Way ANOVA on ranks. It is a nonparametric alternative to One-Way ANOVA. It is an extension of the Man-Whitney Test to situations where more than two levels/populations are involved. This test falls under the family of Rank Sum tests. It depends on the ranks of the sample observations.

Non-Parametric Test: It is a test which does not follow normal distribution.

Elements of a Kruskal Wallis Test

- One independent variable with two or more levels. This independent variable is Categorical.
- One dependent variable which can be in Ordinal, Interval or Ratio level of measurement.

Assumptions of Kruskal Wallis Test

- Independence of Observations – Each observation can belong to only one level.
- No assumption of normality.
- Additional Assumption – The distributions of the dependent variable for all levels of the independent variable must have similar shapes. We can male use of Histograms or Boxplots to determine if the distributions have similar shapes. If this assumption is met it allows you to interpret the results of the Kruskal Wallis Test in terms of medians and not just mean ranks.

Null Hypothesis of Kruskal Wallis Test

The Kruskal Wallis Test has one Null Hypothesis i.e. – The distributions are Equal.

H Statistics of Kruskal Wallis Test

$$H = \left[\frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i}\right] - 3(n+1)$$

n_i = number of items in sample i

R_i = sum of ranks of all items in sample i

K = total number of samples

$n = n_1 + n_2 + \ldots\ldots + n_K$; Total number of observations in all samples.

Steps to perform Kruskal Wallis Test

Let us take an example to understand how to perform this test.

Example:- The score of a sample of 20 students in their university examination are arranged according to the method used in their training : 1) Video Lectures 2) Books and Articles 3) Class Room Training. Evaluate the Effectiveness of these training methods at 0.10 level of significance.

Video Lecture	Books and Articles	Class Room Training
76	80	70
90	80	85

Video Lecture	Books and Articles	Class Room Training
84	67	52
95	59	93
57	91	86
72	94	79
	68	80

Step 1: Identify Independent and Dependent variables
Here,
Independent variable – method of training. It has three levels.
Dependent variable – examination scores.

Step 2: State the Hypothesis
H_0 = The mean examination scores of students trained by each of the three methods are equal. $u_1=u_2=u_3$.
H_1 = At least one of the mean examination scores is not equal.
Step 3: Sort the data for all groups in ascending order and allot them ranks. If more than one entry has the same score then take the average of the ranks and allot the same rank to each of those entries.

Rank	Score	Training Method	Rank	Score	Training Method
1	52	CR	11	80	BA
2	57	VL	11	80	CR
3	59	BA	13	84	VL
4	67	BA	14	85	CR
5	68	BA	15	86	CR
6	70	CR	16	90	VL
7	72	VL	17	91	BA
8	76	VL	18	93	CR

Rank	Score	Training Method	Rank	Score	Training Method
9	79	CR	19	94	BA
11	80	BA	20	95	VL

In this the score 80 had three ranks 10, 11 and 12. So we took the average of these ranks which was 11.

Step 4: Arrange back according to the levels an calculate the sum of ranks for each level.

Video Lecture	Rank	Books and Articles	Rank	Class Room Training	Rank
57	2	59	3	52	1
72	7	67	4	70	6
76	8	68	5	79	9
84	13	80	11	80	11
90	16	80	11	85	14
95	20	91	17	86	15
		94	19	93	18
	∑=66		**∑=70**		**∑=74**

Step 5: Calculate H Statistics

H = 0.0938

Step 6: Find the critical chi-square value

- The chi-square distribution can be used when all the sample sizes are at least 5.

Degree of freedom = K-1 => 3-1=2

Alpha = 0.10

Use this chi-square table to find the value.

$X^2 = 4.605$

Step 7: Compare H value and Critical Chi- Square value

- If $H_{calc} < X^2$; Accept the Null Hypothesis
- If $H_{calc} > X^2$; Reject the Null Hypothesis

Here, 0.0938 < 4.605.

Since, $H_{calc} < X^2$. We **accept the Null Hypothesis**. We can say that there is no difference in the result obtained by using the three training methods.

Multivariate analysis

In data analytics, we look at different variables (or factors) and how they might impact certain situations or outcomes.

For example, in marketing, you might look at how the variable "money spent on advertising" impacts the variable "number of sales." In the healthcare sector, you might want to explore whether there's a correlation between "weekly hours of exercise" and "cholesterol level." This helps us to understand why certain outcomes occur, which in turn allows us to make informed predictions and decisions for the future.

There are three categories of analysis to be aware of:

Univariate analysis, which looks at just one variable

Bivariate analysis, which analyzes two variables

Multivariate analysis, which looks at more than two variables

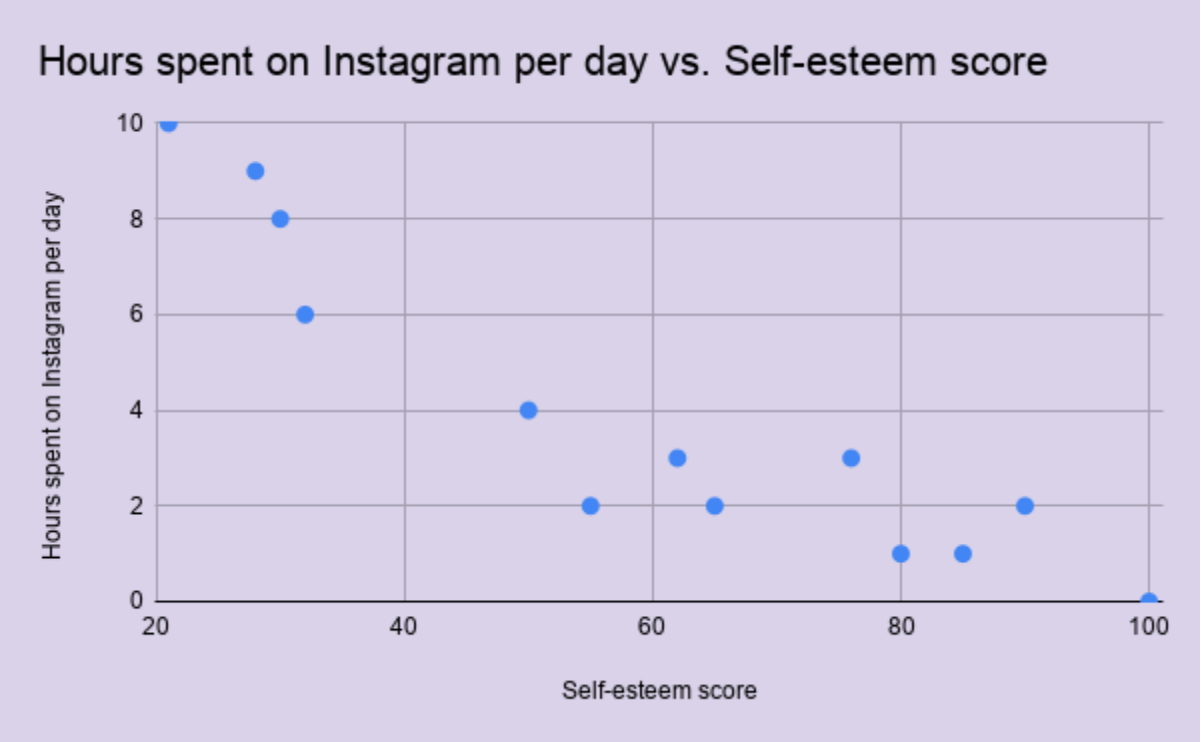
As you can see, multivariate analysis encompasses all statistical techniques that are used to analyze more than two variables at once. The aim is to find patterns and correlations between several variables simultaneously—allowing for a much deeper, more complex understanding of a given scenario than you'll get with bivariate analysis.

An example of multivariate analysis

Let's imagine you're interested in the relationship between a person's social media habits and their self-esteem. You could carry out a bivariate analysis, comparing the following two variables:

How many hours a day a person spends on Instagram

Their self-esteem score (measured using a self-esteem scale)



You may or may not find a relationship between the two variables; however, you know that, in reality, self-esteem is a complex concept. It's likely impacted by many different factors—not just how many hours a person spends on Instagram. You might also want to consider factors such as age, employment status, how often a person exercises, and relationship status (for example). In order to deduce the extent to which each of these variables correlates with self-esteem, and with each other, you'd need to run a multivariate analysis.

Multivariate data analysis techniques and examples

There are many different techniques for multivariate analysis, and they can be divided into two categories:

- Dependence techniques
- Interdependence techniques

Dependence methods

Dependence methods are used when one or some of the variables are dependent on others. Dependence looks at cause and effect; in other words, can the values of two or more independent variables be used to explain, describe, or predict the value of another, dependent variable? To give a simple example, the dependent variable of "weight" might be predicted by independent variables such as "height" and "age."

In machine learning, dependence techniques are used to build predictive models. The analyst enters input data into the model, specifying which variables are independent and which ones are dependent—in other words, which variables they want the model to predict, and which variables they want the model to use to make those predictions.

Interdependence methods

Interdependence methods are used to understand the structural makeup and underlying patterns within a dataset. In this case, no variables are dependent on others, so you're not looking for causal relationships. Rather, interdependence methods seek to give meaning to a set of variables or to group them together in meaningful ways.

So: One is about the effect of certain variables on others, while the other is all about the structure of the dataset.

MULTIPLE CORRELATION

When the value of a variable is influenced by another variable, the relationship between them is a simple correlation. In a real life situation, a variable may be influenced by many other variables. For example, the sales achieved for a product may depend on the income of the consumers, the price, the quality of the product, sales promotion techniques, the channels of distribution, etc. In this case, we have to consider the joint influence of several independent variables on the dependent variable. Multiple correlations arise in this context.

Suppose Y is a dependent variable, which is influenced by n other variables X1, X2, …,Xn. The multiple correlation is a measure of the relationship between Y and X1, X2,…,Xn considered together.

The multiple correlation coefficients are denoted by the letter R. The dependent variable is denoted by X1. The independent variables are denoted by X2, X3, X4,…, etc.

Meaning of notations:

R1.23 denotes the multiple correlation of the dependent variable X1 with two independent variables X 2 and X3 . It is a measure of the relationship that X1 has with X2 and X3 .

R2.13 is the multiple correlation of the dependent variable X2 with two independent variables X1 and X3.

R3.12 is the multiple correlation of the dependent variable X3 with two independent variables X1 and X2.

R1.234 is the multiple correlation of the dependent variable X1 with three independent variables X2 , X3 and X4.

Coefficient Of Multiple Linear Correlations

The coefficient of multiple linear correlation is given in terms of the partial correlation

coefficients as follows:

$$R_{1.23} = \frac{\sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}}{\sqrt{1 - r^2_{23}}}$$

$$R_{2.13} = \frac{\sqrt{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}}{\sqrt{1 - r^2_{13}}}$$

$$R_{3.12} = \frac{\sqrt{r^2_{31} + r^2_{32} - 2\, r_{31}\, r_{32}\, r_{12}}}{\sqrt{1 - r^2_{12}}}$$

Properties Of The Coefficient Of Multiple Linear Correlations:

1. The coefficient of multiple linear correlations R is a non-negative quantity. It varies between 0 and

2. $R_{1.23} = R_{1.32}$

   $R_{2.13} = R_{2.31}$

   $R_{3.12} = R_{3.21}$, etc.

3. $R_{1.23} \geq |r_{12}|$,

   $R_{1.32} \geq |r_{13}|$, etc.

Problem 3

If the simple correlation coefficients have the values $r_{12} = 0.6$, $r_{13} = 0.65$, $r_{23} = 0.8$, find the multiple correlation coefficient $R_{1.23}$

Solution:

We have

$$R_{1.23} = \frac{\sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}}{\sqrt{1 - r^2_{23}}}$$

$$= \frac{\sqrt{(0.6)^2 + (0.65)^2 - 2x0.6x0.65x0.8}}{\sqrt{1 - (0.8)^2}}$$

$$= \frac{\sqrt{0.36+ 0.4225- 0.624}}{\sqrt{1 - 0.64}}$$

$$= \frac{\sqrt{0.7825- 0.624}}{\sqrt{0.36}}$$

$$= \frac{\sqrt{0.1585}}{\sqrt{0.36}}$$

$$= \sqrt{0.4403}$$

$$= 0.6636$$

Problem 4

Given that $r_{21} = 0.7$, $r_{23} = 0.85$ and $r_{13} = 0.75$, determine $R_{2.13}$

Solution:

We have $R_{2.13} = \frac{\sqrt{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}}{\sqrt{1 - r^2_{13}}}$

$$= \frac{\sqrt{(0.7)^2+ (0.85)^2- 2\ x0.7x0.85x0.75}}{\sqrt{1 - (0.75)^2}}$$

$$= \frac{\sqrt{0.49+ 0.7225- 0.8925}}{\sqrt{1 - 0.5625}}$$

$$= \frac{\sqrt{1.2125- 0.8925}}{\sqrt{0.4375}}$$

$$= \frac{\sqrt{0.32}}{\sqrt{0.4375}}$$

$$= \sqrt{0.7314}$$

$$= 0.8552$$

Multiple Regression

Multiple Regression is a set of techniques that describes-line relationships between two or more independent variables or predictor variables and one dependent or criterion variable. A dependent variable is modeled as a function of various independent variables with corresponding coefficients along with the constant terms. Multiple regression requires multiple independent variables and, due to this it is known as multiple regression. In multiple regression, the aim is to introduce a model that describes a dependent variable y to multiple independent variables.In this article, we will study what is multiple regression, multiple regression equation, assumptions of multiple regression and difference between linear regression and multiple regression.

Multiple Regression Equation

There is only one dependent variable and one independent variable is included in linear regression whereas in multiple regression, there are multiple independent variables that enable us to estimate the dependent variable y.Multiple regression equation is derived by:Y = a + b1*1 + b2*2 + b3*3……. bk*kHere, y is an independent variables whereas b1, b2 and bk

Multiple Regression Assumptions

- There should be systematic specification of the model in multiple regression. It implies that only relevant variables should be included in the model and the model should be accurate.
- Assumption of linearity is necessary.
- The multiple regression model should be linear in nature.
- Assumption of normality is necessary in multiple regression. It implies that in multiple regression, variables must have normal distribution.
- Assumption of Homoscedasticity is necessary in multiple regression
- The variance is constant across all levels of the independent variable.
- The independent variables are not highly correlated with each other.

There are various terminologies that help us to understand multiple regression in a better way. These terminologies are as follows:

- The beta value is used in measuring how effectively the independent variable influences the dependent variable. It is measured in terms of standard deviation.
- R, is the measure of linkage between the observed value and the predicted value of the dependent variable. R Square, or R^2, is the square of the measure of association which represents the percentage of overlap between the independent variables and the dependent variable. Adjusted R^2 is an estimate of the R^2 if you make use of multiple regression models with a new data set.

Factor Analysis

Factor analysis is used to uncover the latent structure of a set of variables. It reduces attribute space from a large no. of variables to a smaller no. of factors and as such is a non-dependent procedure. Factor analysis could be used for any of the following purpose-

1. To reduce a large no. of variable to a smaller no. of factors for modeling purposes, where the large number of variables precludes modeling all the measures, individually. As such factor analysis is integrated in structural equation modeling, helping create the latent variables modeled by SEM (structure equation model).
2. To select a subset of variables from a large set based on which original variable have the highest correlations with the principal component factors.
3. To create a set of factors to be treated as uncorrelated variable as one approach to handling multicollinearity regression.

Assumptions:

Factor analysis is a part of the Multiple General Linear Hypothesis (MGLH), family of procedures and makes many of the same assumptions as multiple regressions. Linear relationship interval or near-interval data, untruncated data, proper specification (relevant variable included extraneous are excluded), lack of high multicollinearity and multivariable normality for purpose of significant testing. Factor analysis generates a table on which the rows and the observed row indicator variables and the columns are the factor or latent variables which explain as much of the variable in those variables as possible. The cells in this table are factor loadings and the meaning of the factors must be induced from seeing which variables are most heavily loaded on which factors this inferential process can be fraught with difficulty as diverse researchers impute different tables.

Methods

There are several different types of factor analysis-

1. Principal component method

2. Principal axes method

3. Summation method

4. Centroid method

Assumptions of factor analysis model

1) Measurement error has constant variance and is on average zero, i.e.,
   $var(e_i) = \sigma_i^2$
   $E[e_i] = 0$
2) No association between the factor and measurement error, $\text{cov}(F, e_j) = 0$
3) No association between errors, $\text{cov}(e_j, e_k) = 0$

4) **Local (i.e., conditional independence):** Given factor, observed variables are independent of one another, $\text{cov}(X_j, X_k \mid F) = 0$

Steps in Exploratory Factor Analysis:

1. Collect data: choose relevant variables.

2. Extract initial factors (via principal component).

3. Choose number of factors to retain.

4. Choose estimation method, estimate model.

5. Rotate and interpret.

6. (a) Decide on changes need to be made (e.g. drop items include items)

(b) Repeat (4), (5).

7. Construct scales and use on further analysis.

Methods for finding number of factors to be extracted

1) Thumb Rule:

All the interrelated factors must explain at least as much as variances as an average variable. Check, if a variable is under a factor then the percentage of variable explaining variance should be less than the percentage of factor explaining.

2) Eigen Value Index:

When the eigen value of a factor is less than 1, it explains less variance then the variables included in the factor itself such a factor should not be considered.

3) Fruckter Formula:

$$\text{Number of factors} = \frac{(2n-1)-\sqrt{8n+1}}{2}$$

Where n is the number of variables included in the study.

4) Residual correlation matrix method:

In this method, the residual correlation is observed and if it is soon that most of the correlation coefficient in this matrix are zero, and then the extraction of the factor can be determinate.

5) Scree Plot test:

This method is to decide about the number of factors to be retained from the extracted factors. The test determines which of the extracted factors are actually contributing variance and does not measure random errors. The number of factors is plotted against the proportion of variance. It extracts in the order of the extracted factors.

Standardization of Responses:

$$\hat{X}_i = \frac{X_i - \overline{X}}{\sigma}$$

where X_i is a value corresponding to a response and σ is the variance.

Factor Loading: It is the correlation between a factor and a variable. It helps interpret the meaning of a factor by indicating how well the factor fits the standardized responses to a variable. The greater the value of factor loading the better is the fit of the factor to the data from the concerned statement. All variables load on all factors but they load highly on some specific factors. Range is 1 ±

There may be some variables which may by loading highly to more than two factors, decide in which factors which variables are to be kept.

Eigen Values:

It is the measurement of the amount of variants explained by a factor. A factor eigen value is the sum of the square of its factor loading. Communality: It indicates the proportion of variance in the responses to the statement which is explained by the identified factors.

Communality:

It indicates the proportion of variance in the responses to the statement which is explained by the identified factors.

Percentage of Variance:

$$= \frac{\text{eigen value of the factor}}{\text{sum of all eigen values}} \times 100$$

If any variable is not combined in any of the groups, then it can be left or can be considered as another factor. To remove variable which is totally different use rotation i.e., we are basically changing its direction.

Rotation of Factors:

For the purpose of simplifying the interpretation of obtained factors and to increase the number of high and low positive loadings in the columns of a factor, factor rotation is used. There are two methods for this:

1) Orthogonal Rotation/ Variance Rotation: Here factors are rotated such that the original factors as well as rotated factors are orthogonal. The line between the factors axis remains 90•

2) Promax Rotation: The factors are rotated such that the line between original and rotated factors is more than or less than 90•

Advantages of Factor Analysis:

1. Both objective and subjective attributes can be used.

2. It can be used to identify the hidden dimensions or constraints which may or may not be apparent from direct analysis.

3. It is not extremely difficult to do and at the same time its inexpensive and gives accurate results. 4. There is flexibility in naming and using dimensions.

Disadvantages of Factor Analysis:

1. The usefulness depends on the researcher's ability to develop a complete and accurate set of product attributes. If important attributes are missed the value of procedure is reduced accordingly. 2. Naming of the factors can be difficult multiple attributes can be highly correlated with no apparent reasons.

3. If the observed variables are completely unrelated the factor analysis is unable to produce meaningful pattern.

4. It is not possible to know factors actually represents, only theory can help inform the researcher's on this.

Friedman Test:

It is a non-parametric test alternative to the one way ANOVA with repeated measures. It tries to determine if subjects changed significantly across occasions/conditions. For example:- Problem-solving ability of a set of people is the same or different in Morning, Afternoon, Evening. It is used to test for differences between groups when the dependent variable is ordinal. This test is particularly useful when the sample size is very small.

Elements of Friedman Test

- One group that is measured on three or more blocks of measures overtime/experimental conditions.
- One dependent variable which can be Ordinal, Interval or Ratio.

Assumptions of Friedman Test

- The group is a random sample from the population.
- Samples are not normally distributed.

Friedman Test Formula:

The Friedman test statistic (χ_F^2) is calculated as:

$$\chi_F^2 = \frac{12}{nk(k+1)} \left[\sum R_j^2\right] - 3n(k+1)$$

Where:

- n is the number of subjects.
- k is the number of treatments (or conditions).
- R_j is the sum of ranks for treatment j.

Steps for Conducting the Friedman Test:

1. **Rank the Data**: Rank the scores for each individual across treatments. The smallest value gets the smallest rank (1), and so on.
2. **Compute Rank Sums**: Sum the ranks for each treatment across all individuals.
3. **Calculate the Friedman Statistic**: Use the formula to calculate the test statistic based on the rank sums.
4. **Compare with Critical Value**: The test statistic follows a chi-square distribution with k−1k - 1k−1 degrees of freedom. Compare the test statistic to the critical chi-square value from the table for the given degrees of freedom and significance level (α=0.05\alpha = 0.05α=0.05 is common).
5. **Make a Decision**: If the test statistic is larger than the critical value, reject the null hypothesis, indicating that there are significant differences between the treatments.

Example:

Let's take a simplified example from an experiment where 5 individuals tried three different types of diets, and their weight loss was recorded. We want to check if there is a significant difference in the effectiveness of the diets using the Friedman test.

Data:

Individual	Diet A	Diet B	Diet C
1	5	6	8
2	7	7	5
3	6	5	9
4	4	5	6
5	8	7	7

1. **Rank the Data:**

Individual	Diet A (Rank)	Diet B (Rank)	Diet C (Rank)
1	1	2	3
2	3	3	1
3	2	1	3
4	1	2	3
5	3	2	2

2. Compute Rank Sums:

- Rank sum for Diet A: 1+3+2+1+3=101 + 3 + 2 + 1 + 3 = 101+3+2+1+3=10
- Rank sum for Diet B: 2+3+1+2+2=102 + 3 + 1 + 2 + 2 = 102+3+1+2+2=10
- Rank sum for Diet C: 3+1+3+3+2=123 + 1 + 3 + 3 + 2 = 123+1+3+3+2=12

3. **Friedman Test Statistic Calculation:**

Using the formula:

$$\chi_F^2 = \frac{12}{nk(k+1)} \left[\sum R_j^2\right] - 3n(k+1)$$

Where:

- $n = 5$ (the number of individuals),
- $k = 3$ (the number of diets).

Substitute the values:

$$\chi_F^2 = \frac{12}{5 \times 3 \times (3+1)} \left[10^2 + 10^2 + 12^2\right] - 3 \times 5 \times (3+1)$$

$$\chi_F^2 = \frac{12}{60} \times (100 + 100 + 144) - 60$$

$$\chi_F^2 = \frac{12}{60} \times 344 - 60$$

$$\chi_F^2 = 68.8 - 60$$

$$\chi_F^2 = 8.8$$

4. **Compare with Critical Value:** The degrees of freedom (df) for the test is k−1=3−1=2k - 1 = 3 - 1 = 2k−1=3−1=2.

Using a chi-square distribution table, the critical value for α=0.05\alpha = 0.05α=0.05 and 2 degrees of freedom is **5.99**.

Since the calculated test statistic (8.88.88.8) is **greater** than the critical value (5.99), we **reject the null hypothesis**.

Conclusion:

There is a significant difference in weight loss across the three diets at the 0.05 significance level.

CLUSTER ANALYSIS

Cluster analysis is a statistical technique used to group similar objects or observations into clusters or categories based on their characteristics or features. The goal of cluster analysis is to identify natural groupings within the data, where objects in the same group are more similar to each other than to objects in other groups.

Applications of Cluster Analysis:

Cluster analysis is widely used across various fields such as:

- **Marketing**: Segmenting customers based on purchasing behavior.
- **Biology**: Grouping species based on genetic or phenotypic traits.
- **Social Sciences**: Identifying patterns in human behavior or attitudes.
- **Healthcare**: Classifying patients based on symptoms for disease diagnosis.

Types of Cluster Analysis Methods

1. **Hierarchical Clustering**:
    - **Agglomerative** (bottom-up): Starts with each observation as its own cluster and then progressively merges the closest clusters.
    - **Divisive** (top-down): Starts with all observations in one cluster and then divides them into smaller clusters.
    - **Dendrogram**: A tree-like diagram used to visualize hierarchical clustering.
2. **Partitioning Clustering**:
    - Divides the data into a predetermined number of clusters.
    - **K-Means Clustering**: One of the most common partitioning methods. It minimizes the variance within each cluster by repeatedly adjusting the cluster centers (centroids) and reassigning points to the closest centroid.
    - **K-Medoids** (or PAM): Similar to K-Means, but instead of centroids, it uses actual data points (medoids) as cluster centers.
3. **Density-Based Clustering**:
    - Clusters are defined as dense regions of points separated by areas of low density.
    - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Groups together closely packed points and marks points that are isolated as noise.

4. **Model-Based Clustering**:
    - Assumes that the data is generated from a mixture of underlying probability distributions (e.g., Gaussian distributions).
    - **Expectation-Maximization (EM)**: A popular model-based approach that estimates the parameters of the probability distributions to find the clusters.
5. **Fuzzy Clustering**:
    - Allows each data point to belong to more than one cluster, assigning a membership value to each cluster.
    - **Fuzzy C-Means (FCM)**: Similar to K-Means but allows for soft clustering by giving partial membership to data points.

Steps in Cluster Analysis:

Data Pre-processing:

Standardize the data if the variables are on different scales to ensure that no variable dominates the distance calculation.

Handle missing values and decide whether to remove outliers or treat them in a special way.

Choosing a Clustering Method:

Based on the type of data and the desired outcome, select a clustering algorithm (e.g., K-Means, hierarchical, DBSCAN).

Define Similarity Measure:

Use a similarity or distance measure to determine how close or far apart the data points are from each other.

Common distance measures:

Euclidean distance: Measures the straight-line distance between two points.

Manhattan distance: Measures the sum of the absolute differences between two points.

Cosine similarity: Measures the angle between two vectors, often used for text or high-dimensional data.

Determine the Number of Clusters:

For methods like K-Means, the number of clusters □ needs to be specified in advance. Common techniques for determining □ include:

Elbow Method: Plots the explained variance as a function of the number of clusters and looks for the "elbow point."

Silhouette Analysis: Measures how similar an object is to its own cluster compared to other clusters.

Perform Clustering:

Apply the chosen clustering algorithm and generate the clusters.

Evaluate Clustering Results:

Internal evaluation: Measures like cohesion (within-cluster similarity) and separation (between-cluster difference) can help assess the quality of clustering.

External evaluation: If true labels are available, use measures such as purity, Rand index, or F-measure to compare the clustering with the known groups.

Interpret the Clusters:

Analyze the clusters to identify patterns, relationships, or common characteristics within each group.

Example: K-Means Clustering

Let's go through an example of K-Means Clustering using simple data.

Problem:

Suppose we have data on the height and weight of individuals, and we want to group them into clusters based on these two features.

Individual	Height (cm)	Weight (kg)
1	170	60
2	165	58
3	180	75
4	155	54
5	160	57
6	175	72
7	167	62
8	182	78

Step-by-Step K-Means Clustering Process:

1. **Choose kkk (Number of Clusters)**: Suppose we decide to divide the individuals into 2 clusters (i.e., k=2k = 2k=2).
2. **Initial Centroid Assignment**: Randomly assign two initial centroids (means) from the data points. Let's assume we pick Individual 2 (Height: 165, Weight: 58) and Individual 6 (Height: 175, Weight: 72) as initial centroids.
3. **Assign Data Points to Nearest Centroid**: Calculate the Euclidean distance of each individual from both centroids and assign them to the nearest centroid.

Euclidean distance formula:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

After calculation, individuals will be assigned to one of the two clusters based on which centroid they are closer to.

4.**Update the Centroids**: After assigning individuals to clusters, compute the new centroid by averaging the heights and weights of the individuals in each cluster.

5.**Repeat Steps 3 and 4**: Reassign individuals to the new centroids and recalculate the centroids. Continue the process until the centroids stabilize, meaning no further changes in cluster assignment.

6. **Final Clusters**: After the iterations, the final clusters represent the groupings of individuals based on similarity in height and weight.

CONFIRMATORY FACTOR ANALYSIS (CFA):

Confirmatory Factor Analysis (CFA) is a statistical technique used to test whether a set of observed variables represent a certain number of underlying latent factors. CFA is part of a larger family of techniques known as **structural equation modelling (SEM)** and is used to verify or confirm a hypothesized factor structure or model.

While **Exploratory Factor Analysis (EFA)** explores the possible underlying structure without preconceived hypotheses, CFA is used when you have a specific theory or model about how the variables relate to the latent factors and you want to test the fit of this model.

Applications of CFA:

CFA is used widely in social sciences, psychology, education, and other fields to:

- Validate the structure of psychological tests or surveys.
- Confirm the relationships between observed variables and their underlying constructs (latent variables).
- Test a pre-specified hypothesis regarding the factor structure of a dataset.

Advantages of CFA:

1. **Theory-Driven**: CFA is a hypothesis-driven approach that allows researchers to test specific models.
2. **Measurement Validation**: It is widely used to validate the structure of psychometric instruments (e.g., tests and surveys).
3. **Complex Models**: CFA allows the testing of complex models involving multiple factors and covariance between them.

Challenges in CFA:

- **Model Fit**: Achieving good model fit can be difficult, and poor fit may require theoretical justifications for model modifications.
- **Sample Size**: CFA requires a relatively large sample size for stable and reliable results.
- **Model Identification**: Some models may not be identified, meaning there is insufficient information to estimate the parameters.

Example of CFA:

Let's assume we are testing a survey that measures two latent factors: Job Satisfaction and Work-Life Balance. The survey consists of 6 items (observed variables):

Item	Question
Q1	I am satisfied with my job overall.
Q2	My job provides me with opportunities for growth.
Q3	I find my work meaningful.
Q4	I can balance my work with my personal life.
Q5	I have enough time for personal activities after work.
Q6	My job allows me to maintain a healthy work-life balance.

We hypothesize that:

- **Job Satisfaction** is measured by Q1, Q2, and Q3.
- **Work-Life Balance** is measured by Q4, Q5, and Q6.

Steps:

1. **Model Specification**:
    - We define two latent factors: **Job Satisfaction** and **Work-Life Balance**.
    - Factor 1 (Job Satisfaction) is expected to explain Q1, Q2, and Q3.
    - Factor 2 (Work-Life Balance) is expected to explain Q4, Q5, and Q6.
2. **Model Estimation**:
    - Using CFA software, we estimate the factor loadings for each question on its corresponding latent factor.
3. **Fit Evaluation**:
    - After estimating the model, we obtain the following fit indices:
        - **CFI = 0.95** (indicating a good fit).
        - **RMSEA = 0.04** (below 0.05, indicating a close fit).
        - **SRMR = 0.03** (below 0.08, indicating a good fit).
4. **Model Interpretation**:
    - Suppose we find the following factor loadings:
        - Q1: 0.80, Q2: 0.85, Q3: 0.70 (Job Satisfaction).
        - Q4: 0.78, Q5: 0.82, Q6: 0.77 (Work-Life Balance).
    - These loadings suggest that all items are good indicators of their respective latent factors.
5. **Model Modification**:
    - If the initial fit were poor, we might modify the model by allowing the error terms of Q5 and Q6 to correlate (since they are closely related in content) if theory justifies it.

Interpretation:

The CFA results confirm that the survey's items are valid measures of two distinct but related constructs: **Job Satisfaction** and **Work-Life Balance**. The strong factor loadings and good fit indices provide evidence that the hypothesized model fits the data well.

STRUCTURAL EQUATION MODELING (SEM)

Structural Equation Modelling (SEM) is a comprehensive statistical technique that allows for the examination of complex relationships among observed and latent (unobserved) variables. It combines elements of **factor analysis** and **multiple regression** to estimate interrelated dependencies and test theoretical models.

SEM is used to:

- Model complex relationships between variables.
- Test theories or hypotheses about how variables are related.
- Confirm models (e.g., confirmatory factor analysis) or explore relationships (e.g., path analysis).

Advantages of SEM:

Simultaneous Estimation: SEM allows for the simultaneous estimation of multiple relationships between variables, unlike traditional regression models.

Latent Variables: SEM can handle both observed and latent variables, offering a more comprehensive view of constructs that are not directly measurable.

Mediation and Moderation: SEM is excellent for testing complex relationships, including mediation and moderation effects.

Theory Testing: SEM is primarily a confirmatory technique, allowing researchers to test theoretical models with empirical data.

Concepts in SEM:

Observed Variables:

These are directly measured variables, often referred to as indicators. In surveys or tests, they are the responses or data points that researchers directly collect.

Latent Variables:

Unobserved variables that are inferred from observed variables. These represent underlying theoretical constructs (e.g., intelligence, satisfaction) that cannot be measured directly.

Measurement Model:

This portion of the SEM specifies how latent variables are measured by observed variables. It is equivalent to a confirmatory factor analysis (CFA).

Structural Model:

The structural model specifies the relationships between latent variables, often through regression-like equations. This is the core of SEM, showing how latent variables influence each other.

Path Diagram:

SEM models are often represented graphically using path diagrams. Variables are represented by circles (for latent variables) or squares (for observed variables), and arrows indicate the direction of influence.

Factor Loadings:

The coefficients that describe the relationships between latent variables and their observed indicators in the measurement model. High factor loadings mean the observed variable is a strong indicator of the latent variable.

Direct and Indirect Effects:

SEM allows for the estimation of both direct effects (relationships between two variables with a single arrow) and indirect effects (the effect of one variable on another through one or more mediators).

Model Fit:

SEM provides several fit indices to evaluate how well the proposed model fits the observed data. Key fit indices include the Chi-Square, RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index), and TLI (Tucker-Lewis Index).

Example of SEM

Imagine we want to test the relationship between **Job Satisfaction** (latent variable) and **Job Performance** (latent variable), with **Work-Life Balance** (latent variable) acting as a mediator.

- **Observed Variables** for **Job Satisfaction**: Satisfaction with pay, opportunities for growth, relationships with colleagues.
- **Observed Variables** for **Work-Life Balance**: Time for personal activities, flexible work schedule.
- **Observed Variables** for **Job Performance**: Quality of work, punctuality, meeting deadlines.

Hypothesized Model:

- **Job Satisfaction** influences **Work-Life Balance**.
- **Work-Life Balance** mediates the effect of **Job Satisfaction** on **Job Performance**.
- **Job Satisfaction** also has a direct effect on **Job Performance**.

In this model:

- **Direct effect**: From Job Satisfaction to Job Performance.
- **Indirect effect**: From Job Satisfaction to Job Performance through Work-Life Balance.

Path Diagram:

- Latent variables (Job Satisfaction, Work-Life Balance, and Job Performance) are represented by circles.
- Observed variables (e.g., Satisfaction with pay) are represented by squares.
- Arrows indicate hypothesized causal relationships.

Software for SEM:

Several software packages can perform SEM:

1. **AMOS (Analysis of Moment Structures)**: User-friendly with a drag-and-drop interface for building path diagrams.
2. **LISREL**: One of the oldest SEM programs, favored for advanced users.
3. **Mplus**: Offers a range of SEM features and is highly versatile.
4. **lavaan package in R**: A popular SEM tool for R users, offering flexibility and open-source accessibility.

MULTIPLE DISCRIMINANT ANALYSIS (MDA):

Multiple Discriminant Analysis (MDA) is a classification and dimensionality reduction technique used to distinguish between two or more groups of objects or individuals based on several predictor variables. It is a generalization of linear discriminant analysis (LDA) and is primarily used in cases where the dependent variable is categorical and the independent variables are continuous.

MDA is used when:

- You have multiple groups (typically 2 or more) to classify.
- You want to understand how different predictor variables discriminate between these groups.

Concepts in MDA:

Dependent (Categorical) Variable:

The variable that indicates group membership. For example, if you are studying customers' buying behavior, the dependent variable could be whether they purchase a product (Yes/No), or which product category they belong to.

Independent (Continuous) Variables:

The set of continuous variables that are used to predict group membership. For instance, income, age, or spending habits could serve as predictor variables.

Discriminant Function:

The core of MDA is the discriminant function, which is a linear combination of the predictor variables. It takes the form:

$$D = b_1X_1 + b_2X_2 + \cdots + b_nX_n + c$$

Where:

- D is the discriminant score.
- $X_1, X_2, \ldots, X_n$ are the independent variables.
- $b_1, b_2, \ldots, b_n$ are the coefficients.
- c is a constant.

Discriminant Coefficients:

These coefficients (or weights) represent the contribution of each predictor variable to the discriminant function. Larger absolute values suggest that the variable contributes more to distinguishing between groups.

Centroids:

The mean values of the discriminant scores for each group. These help determine which group an observation belongs to by comparing its discriminant score to the group centroids.

Classification:

Once the discriminant function is created, new cases can be classified into one of the predefined groups based on their discriminant scores.

Advantages of MDA:

Multivariate: MDA allows for the simultaneous consideration of multiple independent variables to classify individuals or objects into groups.

Interpretability: MDA provides a linear discriminant function that is easy to interpret, showing the contribution of each variable to the discrimination between groups.

Predictive Power: MDA can be used to predict group membership for new observations based on the discriminant function.

Handling of Multiple Groups: MDA can handle more than two groups, unlike simple LDA, which is limited to binary classification.

Limitations of MDA:

Strict Assumptions: The assumption of multivariate normality and homogeneity of variance-covariance matrices can be difficult to meet in real-world data.

Sensitive to Outliers: Outliers can distort the results of MDA by significantly affecting the discriminant function.

Linear Relationships: MDA assumes linear relationships between the independent variables and group membership, which may not always hold.

Multicollinearity: High correlations between independent variables can lead to unreliable discriminant functions.

UNIT-V

PREPARATION OF RESEARCH REPORT

INTERPRETATION:

It refers to the task of drawing inferences from the collected facts after an analytical and / or experimental study.

➢It is a search for broader meaning of research findings

➢It has two important aspects:

i. The effort to establish continuity in research through linking the results of a given study with those of another.

ii.The establishment of some explanatory concepts.

➢In one sense, it is concerned with relationships within the collected data, partially overlapping analysis.

➢It also extends beyond the data of the study to include the results of other research, theory and etc

➢Thus, interpretation is the device through which the factors that seem to explain what has been observed by researcher in the course of the study can be better understood and it also provides a theoretical conception which can serve as a guide for further research.

Need for interpretation:

It is through interpretation that the researcher can understand the abstract principle that works beneath his findings .Through this he can link up his findings with those of other studies having the same abstract principle and thereby can predict about the concrete world of events. Fresh enquiries can test these predictions later on. This way the continuity in research can be maintained.

Interpretation leads to the establishment of explanatory concepts than can serve as a guide for further research studies; it opens new avenues of intellectual adventure and stimulate the quest for more knowledge.

Researcher can better appreciate only through interpretation why his findings are what they are and can make others understand the real significance of his research findings.

The interpretation of the findings of exploratory research study often results into hypothesis for experimental research and as such interpretation is involved in the transition from exploratory to experimental research.

Techniques of interpretation:

Interpretation requires great skill and dexterity. It is an art that one learns through practice and experience.

Steps involved in interpretation:

➢ Researcher must give reasonable explanation of the relation and he must interpret relationship in terms of the underlying processes. This is the technique of how generalization should be done and concept be formulated.

➢Extraneous information, if collected during the study, must be considered while interpreting the final results.

➢It is advisable to get frank and honest opinion of experts.

➢All relevant factors must be considered before generalization.

Precautions in interpretation

➢The researcher must invariably satisfy himself that (a) the data are appropriate, trustworthy and adequate (b) the data reflect good homogeneity and (c) proper statistical analysis has been applied.

➢ He must remain cautious about the errors that can possibly arise in the process of interpretation. He should be well equipped with the knowledge of correct use of statistical measures of drawing inferences concerning the study.

➢As the task of interpretation is very much intertwined with analysis and cannot be distinctly separated, it must be taken as a special aspect of analysis.

➢ His task is not only to make sensitive observations but also to identify the factors which were not known initially. Broad generalization should be avoided because the coverage is restricted to a particular time, a particular area or particular condition.

➢There should be constant interaction between initial hypothesis, empirical observation and theoretical conceptions. It is here opportunities for originality and creativity lie.

REPORT WRITING

The importance of report writing in research needs no emphasis. A research is said to be incomplete unless it is presented in a report format. Any research will be appreciated only when it is made known to others through research report. The exotic dishes in a dinner are appreciated by the guests when the host (home maker) lays the table neatly, explains the dishes and serves in a meticulous way. Similarly the efforts of the researcher and the fruits of the research will be appreciated only when it is presented as a report in a logical sequence incorporating all the relevant details.

Meaning

A research report is a formal statement of the details of the research process and its results. It gives an account of the problem(s) studied, objectives, methodology, findings and conclusions of the research study.

Purpose or functions of a research report

➢ To communicate the methodology and results of the study to the targeted audience.

➢ To enable the person(s) concerned determine the validity of the results/conclusion and judge the quality of the research project as well and as the ability and competence of the researcher to do research.

➢ To provide as a base for formulating policies and strategies in the relevant areas.

➢ To provide additional knowledge to tackle certain problems / issues.

➢ To serve as a basic reference for future study.

Characteristics of a good research report

Not only is the report narrative, it must be an authoritative document on the outcome.

➢ It must be specific and accurate and there is no question of beating around the bush.

➢ It must be written with the targeted audience in mind.

➢ It must be non-persuasive. That is, extra caution is needed while advocating a particular course of action based on the finding.

➢ It must be simple, logical and understandable.

PREPARING A RESEARCH REPORT:

The research report is considered as a major component of the research work, because through this report the research problem, the research design, the analysis and findings are brought to the knowledge of the world. The research report converts the research work into a public asset from its earlier state of private asset.

The research report shows the readers the progress in knowledge made in the specific area or discipline. The report by bringing to light the new frontiers of knowledge enhances the society's intellectual well-being. The report by highlighting the design and methodology, runs as a fore-runner for future researchers in this or related area. The analyses and interpretations may give a boost to knowledge. The findings and suggestions take the readers into enlightenment from ignorance. Every research must endeavor to achieve this.

Research report is a record of the whole of every bit of the research work. This document is a reservoir of knowledge for current and future references and use to solve societal problems. Research report is the means through which communication of the entire work to the society is made. For other researchers, a documented research is a source of information and that a research report generates more research interests. Research report propagates knowledge throughout the humanity or the globe.

The role of a research report is best known in the absence of the same – Assume for a while, that no researcher gives out his research work in the form of a report. Then the research work is just like a lamp in the pot. When, it takes the form of report it is like a lamp on the hillock illuminating the surroundings. If a research report is not made, even the researcher may not be able tell his work at a future date. Thanks to human's potentials to forget. Such waste of efforts should never occur. If only a research report was made out, re-inventing the wheel would not take place otherwise, same problem may be analyzed by different people at

different places or in the same place at different times or at the same time. This is a greater waste of human energy. Thus a research report conserves energy that would otherwise would have been spent uselessly.

Content of Research Report:

A research report generally contains three aspects:

- Preliminary Section,
- Main Body and
- Reference Section.

1. Preliminary Section

The preliminary section deals with title, acknowledgement, etc.

Title Page:

The title of the research report usually bears the investigator's name, a statement as to the course for which the study has been required, the date of submission, and the name of the institution making that requirement. In reports of studies not undertaken for any course, the investigator's name, the institution he belongs to and the date of completion of the work is indicated. In a published thesis the latter information is substituted or supplemented by the name of the publishers and the date and place of publication.

Acknowledgement Page:

The acknowledgement page is largely one of courtesy in which the investigator acknowledges the guidance and assistance he has received in the development of the study. Acknowledgement may not refer to the guide so much as to others who may have aided in a special way. It is rightly said that good taste calls for acknowledgements to be expressed simply and tactfully.

Preface or Foreword:

Sometimes a preface or foreword of one or two pages long, follows the acknowledgement page, bearing some initial remarks and perhaps a brief statement of the scope, aim and general character of the research.

Table of Contents:

A well-developed table of contents renders a good deal of assistance to a reader in choosing rapidly and judiciously what he should, subsequently, read carefully. It is usually desirable to include in it not only the chapter headings, but also the headings of the major subdivisions of the chapters. Sometimes the topics within the subdivisions are also included and are found enlightening by the readers.

Lists of Tables and Figures:

Another device used to supplement the table of contents for throwing more light on the subject of the thesis is that of giving lists of tables and figures which occur in the report.

2. Main Body of the Report

The main body of the research report contains all the material aspect of the research work.

Introduction:

The first part of the main body of the report, the Introduction, usually includes a statement of the factors leading up to the choice of the problem, the purposes of the study, the value and significance attached to the problem by the investigator as a contribution to knowledge and any other information to express the sincerity of the investigator in his selection. A statement and elucidation of the problem sometimes forms a part of the introduction; but more often/it is set up as a separate unit. If this is stated in a clear-cut and logical manner, the reader is able to get a sufficiently clear insight into the study from the very beginning. The problem should be defined in detail. The exact area the investigation is supposed to cover must be well demarcated. The sources of information selected and their nature and delimitation's should be mentioned and justified. All terms of a technical nature or those which may seem vague to the lay reader need to be defined carefully. The objectives, limitations, hypotheses, etc. are given. The methodology and design of the study are also given in introduction. To explain the developmental process used for the sturdy the investigator has to describe the techniques and tools he has used for collecting, organizing, analyzing and interpreting his data. The sources of data tapped, the channels prepared or adapted and utilized, the nature of data collected, their validity and reliability – all these should be given in a clear and adequate manner. Data collected, but rejected and the methods tried but not pursued – these should also find their place in the report and should not be just left out of the picture.

Survey of Related Literature:

Any research worker has to be up-to-date in his information about studies, related to his own problem, already made by others. References are made to such similar or related studies and their evaluation too is made for the benefit of the reader either in the introductory chapter, or else in a separate chapter. Herein the author finds another opportunity to justify his own endeavor and to emphasize the worthwhile elements in the treatment, selected by him, of the problem. Read More: The Literature Review in Research

Analysis and Interpretation:

The analysis and interpretation section deals with the main works undertaken. Each objective of the research work, each hypothesis, each research question posed and such other major constituents of the research work are thoroughly probed, analysed using the statistical data collected applying appropriate tools of analysis and interpretations are made in the light of the analysis made. Unusual or complex techniques of collection, organization, analysis and interpretation are explained in full. Whether the original data themselves should be included in the text or given in the appendix depends on the nature of the data. If they are not too extensive and are necessary to clarify the discussion, they should certainly find a place in the text proper, or in the footnotes. If they are extensive and cumbersome, they should be placed in the appendix. Of the various aids used to make the presentation of data more effective, tables and figures are most common. When statistical data are assembled according to certain common factors in the form of tables, significant relationships show up clearly. Depending on the type of material at one's disposal, many kinds of figures are found useful, e.g., statistical diagrams, photographs and maps, etc. All the information described above is sometimes confined to one chapter with separate subdivisions arranged stage-wise. Otherwise, separate chapters are devoted each major functional area or objective studied. The

arrangement depends on the quantity of information one has to convey to the reader regarding the different stages in the process of the development of the study.

Conclusion:

The final unit of the report usually contains the findings of the study, the conclusions the investigator has arrived at, and the generalization he has formulated on the basis of the study. In stating the conclusions, the investigator must indicate what his contribution has been to his field of study. He should indicate on what data his various conclusions are based. He should clearly demarcate between the inevitable conclusions and his own interpretation of certain data. The range of applicability of the conclusions should be indicated on the basis of the limitations of the sources, the sample, the tools of collection and analysis, etc. Negative as well as positive results should find a place in the conclusions. Any recommendations, as to the application of the findings, the investigator wishes to make, can find a place in this chapter. Recommendations or suggestions for further study in the field touched by the present research are also found useful and are usually included in the concluding chapter.

3. Referencing Section of the Report

Referencing section of any research report has three elements namely, bibliography, appendix and index.

Bibliography:

The 'works cited' form of bibliography is preferable over the 'sources consulted'. Every book, thesis, article, documents which has been cited should be included in the list of 'works cited'. The bibliography should follow a logical arrangement in alphabetical order. In report of current practice is to have one comprehensive listing-not to divide into books, journals, newspapers, official papers, documents and manuscripts. The author(s) name, the title of the work, date of publication, name of the publisher and the place of publication be mentioned. For articles, the volume number and inclusive pages be also given, the author's initials or surname should follow the name. When there are three or more authors of a particular work, the co-authors may be referred alphabetically If there by more than one work by the same author, the author's name should be listed only once; subsequently a line will substitute his name. This bibliographical listing should not be numbered. It should be given only at the end of the thesis,

Appendix: The appendix section gives a copy of the tools of research used, certain sample statistical workings, articles published by the researcher, etc. Each class of material given may be numbered as Appendix I, Appendix II and so on. It is saner to give the appendices in the same order in which the relevant items are used.

Index:

Index is a very important component which facilitates easy location of a concept or entity mentioned in the main body of the work. Here alphabetical order is followed. Page number is given to easy location. Author Index, Subject Index and Sponsor Index are certain indices used. All the three may be separately given and merged into one single class of 'index'.

STYLE OF WRITING RESEARCH REPORT(APA, MLA, Anderson, Harvard)

APA Citation Style:

The APA citation style is widely used in the social sciences and is known for its detailed and systematic approach to citations. In APA style, in-text citations include the author's name and the year of publication, similar to Harvard style but with slight variations.

If the author's name is mentioned within the text, the year of publication is placed immediately after the name in parentheses. However, if the author's name is not mentioned, both the name and the year are included in parentheses.

For instance, if you were citing a journal article by Emily Johnson published in 2019, your citation would look like this: "Johnson (2019) found that…". On the other hand, if the author's name was not mentioned in the text, the citation would be: "(Johnson, 2019)".

The reference list in APA style provides comprehensive information about the sources cited, including the author's name, publication year, title, and publication details. Additionally, the APA style incorporates specific formatting guidelines for headings, abstracts, and overall paper structure.

MLA Citation Style:

The MLA citation style is predominantly used in the humanities disciplines and focuses on providing concise information within the text and the Works Cited page. In MLA style, in-text citations typically include the author's last name and the page number without the need for a comma between them. For example, if you were quoting a passage from a book by Jane Anderson found on page 45, your citation would look like this: "(Anderson 45)".

This citation style also encourages the use of signal phrases to introduce sources, which can enhance the flow and clarity of the text. The Works Cited page in MLA style lists the full details of the sources cited, including the author's name, title, publication information, and medium of publication (e.g., print, web).

MLA style places significant emphasis on the uniformity and legibility of formatting, including guidelines for font type, line spacing, and indentation.

Harvard Citation Style:

The Harvard citation style, also known as the author-date system, emphasizes the inclusion of the author's name and publication date within the text. In Harvard style, in-text citations typically take the form of (Author's Last Name, Year of Publication), enabling readers to easily locate the corresponding entry in the reference list. For example, if you were citing a book by John Smith published in 2020, your citation would look like this: "(Smith, 2020)".

The reference list in Harvard style is arranged alphabetically by the author's last name and includes comprehensive details about the sources cited, such as the author's name, publication year, title, and publication information. The Harvard style is commonly used in the social sciences, natural sciences, and humanities disciplines.

Anderson Style:

The Anderson style of research report writing typically refers to guidelines that may be set by a specific institution, professor, or course. Unlike APA, MLA, or Harvard, Anderson style is not a globally recognized citation system, but here's a general guide to its possible structure:

Characteristics of Anderson Style:

Title Page:

Contains the title of the report, the author's name, course details, and the date.

Abstract:

Brief summary (around 150-250 words) of the research, including objectives, methods, findings, and conclusions.

Table of Contents:

Lists all sections and subsections, along with page numbers.

Introduction:

Outlines the background, problem statement, objectives, and the scope of the research.

Methodology:

Describes the research methods used (qualitative, quantitative, etc.), sample size, and data collection techniques.

Results and Discussion:

Presents research findings and interprets them in the context of the research question or hypothesis.

Conclusion:

Summarizes the findings and provides final thoughts or recommendations.

References/Bibliography:

Citations may follow a style similar to APA or Harvard, depending on the instructor's preference.

Author-date format for in-text citations (e.g., Smith, 2020), followed by a full list of references.

Appendices (if required):

Includes additional materials such as raw data, charts, or questionnaires.

Formatting:

Font: Times New Roman or Arial, 12 pt.

Spacing: Double-spaced text, with clear section headings.

Margins: 1-inch margins on all sides.

Page Numbers: Centered or aligned to the right at the top or bottom of each page.

Note: Since Anderson style is not a standardized global style, it's essential to follow specific instructions from the relevant professor or institution. It usually blends elements from recognized styles like APA, Harvard, or MLA.

MECHANICS OF WRITING A RESEARCH REPORT:

The mechanics of writing a research report refer to the basic guidelines and structure necessary to create a clear, coherent, and professional report. Here's a breakdown of the main components and steps involved:

1. Title Page

Title: Clear and concise, reflecting the research topic.

Author's Name: Your name and any co-authors.

Institution: The institution where the research was conducted.

Date: Submission date.

2. Abstract

Length: 150–250 words.

Content: A brief summary of the report, including the purpose, methods, key findings, and conclusions. It provides a snapshot of the entire report.

3. Table of Contents

Content: Lists all sections and subsections of the report with page numbers, allowing readers to navigate the document easily.

4. Introduction

Background Information: Provides context for the research and introduces the problem or question being addressed.

Purpose or Objectives: Clearly states the aim of the research.

Research Hypothesis or Questions: Defines what the report seeks to prove or explore.

Significance: Explains why the research is important and its potential contributions to the field.

5. Literature Review (if applicable)

Summary of Previous Research: Discusses the existing work in the area to provide a foundation for the current research.

Gaps in Knowledge: Identifies what has not been addressed by prior research.

6. Methodology

Research Design: Describes the type of research (e.g., qualitative, quantitative).

Sample and Participants: Explains who or what was studied.

Data Collection Methods: Discusses how data was gathered (e.g., surveys, experiments, interviews).

Data Analysis: Describes how the data was analyzed (e.g., statistical analysis, thematic analysis).

7. Results

Data Presentation: Presents findings clearly and concisely, often using tables, charts, or graphs.

Analysis: Discusses the results without interpreting their meaning. Only factual findings are presented in this section.

8. Discussion

Interpretation of Results: Explains what the results mean in the context of the research questions or hypothesis.

Comparison to Previous Research: Relates the findings to the existing literature and discusses any similarities or differences.

Limitations: Acknowledges any limitations in the research design or execution that might affect the validity or generalizability of the results.

Implications: Discusses the impact of the findings and possible areas for further research.

9. Conclusion

Summary: Summarizes the main findings and their significance.

Recommendations: Provides suggestions for practical applications or future research.

Final Thoughts: May include personal reflections on the research process.

10. References/Bibliography

Citations: Lists all sources cited in the report, following the appropriate citation style (e.g., APA, MLA, Harvard). Use a consistent format for all entries.

11. Appendices (if applicable)

Additional Materials: Contains supplementary information such as raw data, questionnaires, or detailed explanations of methods that were too lengthy to include in the main body.

Writing Mechanics

Clarity and Precision: Write clearly and concisely, avoiding unnecessary jargon.

Active vs. Passive Voice: Use active voice where possible for stronger, more direct writing.

Tense: Use past tense when describing completed research and present tense when discussing established knowledge or general facts.

Consistency: Maintain a consistent format in headings, font, and citation style.

Grammar and Spelling: Proofread carefully to ensure there are no errors in grammar, spelling, or punctuation.

ETHICS IN RESEARCH:

Research ethics is a discipline that ensures research is conducted responsibly and fairly, and with respect for participants and the researcher's integrity. Some key principles of research ethics include:

Honesty: Researchers should honestly report data, methods, procedures, and results. They should not fabricate, falsify, or misrepresent data.

Respect for participants: Researchers should protect the rights and well-being of participants, and treat them with respect and dignity.

Informed consent: Participants should be given informed consent.

Confidentiality: Researchers should protect the confidentiality of participants' data.

Conflict of interest: Researchers should be aware of potential conflicts of interest.

Avoiding plagiarism: Researchers should avoid plagiarism by properly citing and referencing previously published content.

What is Plagiarism?

According to the American Heritage Dictionary of the English Language the Latin “plagium” means "kidnapping". It is intellectual theft. It is a serious scientific misconduct. According to the Merriam-Webster Online Dictionary, to “plagiarize” means

1) “to steal and pass off (the ideas or words of another) as one's own”

2) “to use (another's production) without crediting the source”

3) “to commit literary theft”

4) “to present as new and original an idea or product derived from an existing source”

Concisely, we can define plagiarism as an act of fraud.

Plagiarism is an issue of great worry amongst the scholars. Plagiarism is a moral, ethical, and legal issue. In this 4G era internet supplies users with easy access to various kinds of data and information. The large quantity of information available makes it easier and increases the temptation to steal other’s ideas, therefore now the disease of plagiarism has taken the shape of an epidemic. Plagiarism is taking someone else’s work and passing it off as one’s own, is sometimes committed deliberately and other times accidentally. Often, copyrights are violated, which is considered to be unethical act by society, and thus is punishable offense.

AVOIDING PLAGIARISM IN A RESEARCH REPORT:

Avoiding plagiarism in a research report is essential to maintaining academic integrity. Plagiarism occurs when you present someone else’s work, ideas, or words as your own without proper acknowledgment. Here are key strategies to help you avoid plagiarism when writing a research report:

1. Understand What Constitutes Plagiarism

Direct Plagiarism: Copying someone else's words verbatim without quotation marks or citation.

Paraphrasing Plagiarism: Rewriting someone's ideas or sentences without properly citing the original source.

Self-Plagiarism: Reusing your own previous work without permission or citation.

Mosaic Plagiarism: Piecing together phrases or ideas from various sources without proper citation.

2. Cite Your Sources

Use Proper Citation Styles: Whether it's APA, MLA, Harvard, or another style, ensure that you follow the correct format for in-text citations and reference lists.

Example (APA): "Research has shown that…" (Smith, 2020).

Example (MLA): "Research has shown that…" (Smith 123).

Cite All Sources of Information: Any ideas, data, theories, or direct quotations from books, journal articles, websites, or any other sources need to be cited.

Use Primary and Secondary Sources: When citing ideas or data from another source, make sure you are aware of whether you're citing the primary source or another author citing it (secondary source).

3. Paraphrase Effectively

Understand the Original Text: Before attempting to paraphrase, fully understand the meaning of the text.

Use Your Own Words: When paraphrasing, rewrite the idea completely in your own words, but still cite the original source.

Avoid Copying the Structure: Don't just change a few words or rearrange sentences; the structure and wording should be entirely your own.

Example:

Original: "Global warming has had significant effects on the Arctic region."

Paraphrase: "The Arctic has experienced notable changes due to rising global temperatures" (Smith, 2020).

4. Use Quotation Marks for Direct Quotes

Direct Quotes: If you are using the exact words of another author, put the text in quotation marks and cite the source.

Limit Quotations: Use direct quotes sparingly. Whenever possible, paraphrase instead, and only quote when the original wording is crucial for the point you are making.

Example: According to Smith (2020), "Global warming has severely affected Arctic wildlife" (p. 45).

5. Keep Track of Your Sources

Organize Your Research: As you gather information, keep a detailed record of every source you consult, including books, journal articles, websites, and more.

Use Citation Management Tools: Tools like Zotero, EndNote, or Mendeley can help you track and organize your sources, ensuring you don't forget to cite them properly.

Highlight or Annotate: While taking notes, clearly distinguish between your own thoughts and ideas that come from your research sources.

6. Include a Reference List

Full Citations: At the end of your research report, include a reference list (or bibliography) that contains full citations for every source cited in your text.

Alphabetical Order: Ensure that your references are organized alphabetically by the author's last name.

Complete Information: Provide all necessary details like author, title, publication date, and publisher for books or articles, and full URLs for websites.

7. Use Plagiarism Checkers

Online Tools: Run your work through plagiarism detection tools like Turnitin, Grammarly, or others to identify unintentional plagiarism.

Manual Check: Review your work carefully to ensure that all external ideas are properly credited.

8. Don't Over-Rely on One Source

Diversify Your Research: Use multiple sources to gather information on your topic. Relying too much on one source can make it more likely that you'll unintentionally mimic its wording or structure.

Synthesize Ideas: Combine ideas from different sources to form a more well-rounded argument, always giving credit to the original authors.

9. Seek Permission When Necessary

Reuse of Your Own Work: If you are reusing portions of previous research or writing (self-plagiarism), make sure you get approval from your instructor or supervisor and properly cite your earlier work.

Images, Charts, and Data: When using figures, charts, or images from other sources, ensure you have permission or use open-access material, and always give credit.

10. Be Aware of Common Knowledge

No Need to Cite Common Knowledge: Facts that are widely known (e.g., "The Earth orbits the sun") do not need to be cited.

Field-Specific Common Knowledge: Information that is considered common within a specific field of study might not need citation either, but if in doubt, cite the source.

By following these steps, you can effectively avoid plagiarism and produce an original, ethically sound research report. Remember, proper attribution not only gives credit to the original authors but also strengthens the credibility of your own research.

Anti-Plagiarism Tools:

There are many anti plagiarism tools are available nowadays.

1. Plagium:

Plagium is a very good plagiarism detection tool. This is a service of Septet Systems Inc. which is a New York-based company that expertises in advanced search solutions for industry, the public sector, and government. It provides an easy to use service that applies to a broad base of users.

2. Turnitin:

Turnitin is one of the leading anti-plagiarism tool across the world. The company's cloud-based service for originality checking, online grading and peer review saves instructors time and provides rich feedback to users.

3.Duplichecker:

Duplichecker analyzes each sentence entered in the text box. It provides free online service (https://www.duplichecker.com/) to the users.

4. Plagiarism detector:

It works like Duplichecker. It also provides free plagiarism detection service.

5. Urkund:

It is a fully automated plagiarism detecting system. Urkund become very popular plagiarism detection tool in higher education institute all over the globe. It verifies all documents against three central sources: the Internet, published materials, and materials previously submitted by students such as projects or assignments etc.a

6. Quetext:

Quetext is free intelligent plagiarism detection software. Simply input text, then it will be analyzed based on lexical frequencies, phrase patterns, and many other factors. Then the text is mapped into an internal network where it is compared against the entire internet and other databases. After the text is finished being scanned, results will appear with an indication of an exact match, or a similar match with a percentage of similarity along with the similar text.

7. Copyleaks:

Copyleaks plagiarism checker fights plagiarism and copyright infringement online. It has advanced technology that works in any language. Check for plagiarism to detect if the content is being used by others.

8. PlagiarismChecker.com:

It helps to find out whether a work of a student has been copied from the Internet or not.

9. Plagiarism.org:

This was created by the students and alumni of University of California, Berkeley. This software doesn‟t differentiate between quoted materials and original writing.

FUNDING AGENCIES FOR BUSINESS RESEARCH:

Funding agencies for business research are organizations, both governmental and non-governmental, that provide financial support to individuals, institutions, or businesses to conduct research that advances knowledge, innovation, and economic development. Below are some notable funding agencies that support business research at national and international levels:

1. Governmental Funding Agencies

These agencies are often national bodies that fund research projects with economic and societal benefits:

India

- Department of Science and Technology (DST): Provides grants for research and innovation in business and technology.
- Schemes: Science for Equity Empowerment and Development (SEED), National Initiative for Developing and Harnessing Innovations (NIDHI).
- Small Industries Development Bank of India (SIDBI): Provides financial assistance and grants to small and medium enterprises (SMEs) involved in business research and innovation.
- Ministry of Micro, Small & Medium Enterprises (MSME): Offers funding for research projects related to SME development, especially in improving business operations, manufacturing, and technology adoption.
- National Science & Technology Entrepreneurship Development Board (NSTEDB): Encourages entrepreneurship through financial support to innovative business models.

United States

- Small Business Innovation Research (SBIR) Program: Offers competitive grants to small businesses for research and development that have the potential for commercialization.
- Small Business Technology Transfer (STTR) Program: Similar to SBIR, but requires collaboration with research institutions.
- National Science Foundation (NSF): Provides grants for business research, particularly those related to innovation, technology, and economic impact.

United Kingdom

- Innovate UK: Part of UK Research and Innovation, it provides funding for innovative business ideas and research with potential commercial applications.

- Economic and Social Research Council (ESRC): Supports research in business, economics, and management.

European Union

- Horizon Europe: The EU's flagship funding program for research and innovation, which supports projects in areas like digital transformation, sustainability, and business innovation.
- European Innovation Council (EIC): Offers funding to small and medium enterprises (SMEs) and start-ups for cutting-edge business research.

2. International Funding Agencies

These organizations provide cross-border grants for businesses, universities, and researchers to advance global business innovation and research:

World Bank: Provides funding through various initiatives aimed at supporting business and economic research in developing countries.

United Nations Development Programme (UNDP): Supports research projects that align with sustainable business practices and development goals.

International Finance Corporation (IFC): Provides financial support for private sector development, especially for research in emerging markets.

Asian Development Bank (ADB): Funds business and economic research that promotes development and innovation in Asia.

3. Private Foundations and Corporations

Several private organizations and foundations fund research projects that have business, entrepreneurial, or economic potential:

Bill & Melinda Gates Foundation: Provides funding for research that focuses on solving global challenges, including in the areas of business development and entrepreneurship.

Ford Foundation: Supports business research that focuses on social change, economic development, and sustainable business practices.

Kauffman Foundation: Focuses on entrepreneurship and offers grants for research that drives business innovation and economic growth.

Rockefeller Foundation: Funds projects that promote innovative business solutions to social and environmental challenges.

4. University and Academic Funding

Several academic institutions offer research grants to support faculty and student business research projects:

Harvard Business School (HBS) Research Funding: Provides support for faculty and doctoral students engaged in cutting-edge business research.

Stanford Graduate School of Business (GSB): Offers grants and funding for business research projects that drive innovation and impact business practices.

London Business School (LBS): Supports research initiatives in areas like entrepreneurship, finance, and management.

5. Industry-Specific Funding Agencies

Certain sectors offer research grants for business innovation and development:

Agriculture Business Research: The United States Department of Agriculture (USDA) funds research into agri-business and rural economic development.

Tech Startups: Agencies like Y Combinator, Techstars, and Google for Startups offer grants and support for early-stage tech business research.

Energy and Clean Technology: The Clean Energy Finance Corporation (CEFC) and Energy Research Accelerator (ERA) provide funding for research in renewable energy and sustainable business practices.

6. Crowdfunding Platforms

Crowdfunding is an alternative way to raise funds for business research:

Kickstarter: A platform where innovators can raise money from the public to fund business ideas and research.

Indiegogo: Similar to Kickstarter, it allows businesses and researchers to seek funding for creative and research-based projects.

How to Secure Funding

Prepare a Strong Proposal: Research agencies usually require a detailed research proposal, outlining objectives, methodology, expected outcomes, and budget.

Align with Funding Priorities: Make sure your research aligns with the agency's priorities, such as economic impact, innovation, or sustainability.

Collaborate with Academia: Many funding bodies prefer projects with academic partnerships or backing from research institutions.

These agencies and organizations provide a wide range of funding opportunities for business research, fostering innovation, entrepreneurship, and economic development across different sectors.