

BIOSTATISTICS AND RESEARCH METHODOLOGY

For Eighth Semester B.Pharm



VIPIN XAVIER

MSc., MCA

Research

Research is an ORGANIZED and SYSTEMATIC way of FINDING ANSWERS to QUESTIONS. Research is a process of systematic inquiry that entails collection of data; documentation of critical information; and analysis and interpretation of that data/information, in accordance with suitable methodologies set by specific professional fields and academic disciplines. It is the careful consideration of study regarding a particular concern or problem using scientific methods. According to the American sociologist Earl Robert Babbie, “research is a systematic inquiry to describe, explain, predict, and control the observed phenomenon. It involves inductive and deductive methods.”

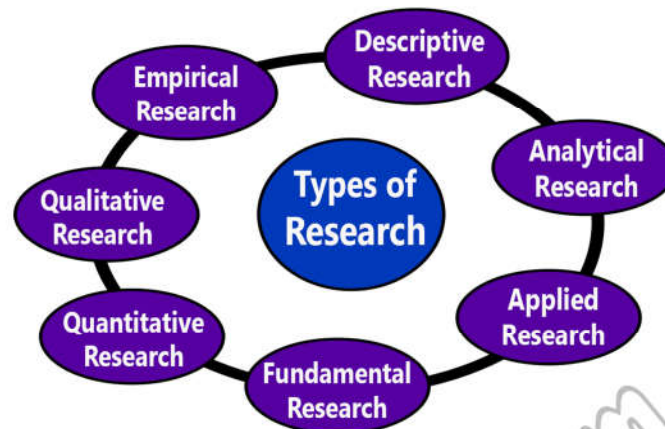
Inductive research methods *analyze an observed event*, while **deductive** methods *verify the observed event*. Inductive approaches are associated with qualitative research, and deductive methods are more commonly associated with quantitative analysis.

Characteristics of research

Good research follows a systematic approach to capture accurate data. Researchers need to practice ethics and a code of conduct while making observations or drawing conclusions.

- The analysis is based on logical reasoning and involves both inductive and deductive methods.
- Real-time data and knowledge is derived from actual observations in natural settings.
- There is an in-depth analysis of all data collected so that there are no anomalies associated with it.
- It creates a path for generating new questions. Existing data helps create more research opportunities.
- It is analytical and uses all the available data so that there is no ambiguity in inference.
- Accuracy is one of the most critical aspects of research. The information must be accurate and correct. For example, laboratories provide a controlled environment to collect data. Accuracy is measured in the instruments used, the calibrations of instruments or tools, and the experiment’s final result.

Types of research



1. Descriptive Research

Descriptive research refers to the methods that describe the characteristics of the variables under study. This methodology focuses on answering questions relating to “what” than the “why” of the research subject. The primary focus of descriptive research is to simply describe the nature of the demographics under study instead of focusing on the “why”. Descriptive research is called an observational research method as none of the variables in the study are influenced during the process of the research.

Example:

Descriptive Study of Drug Compliance in Uncontrolled Hypertensive Patients (THERMO-HTA). According to the World Health Organization (WHO), “insufficient adherence is the main reason why patients do not get all the benefits they could expect from their medicines. It causes medical and psychosocial complications, diminishes the quality of life of patients, increases the likelihood of drug resistance, and waste of resources.

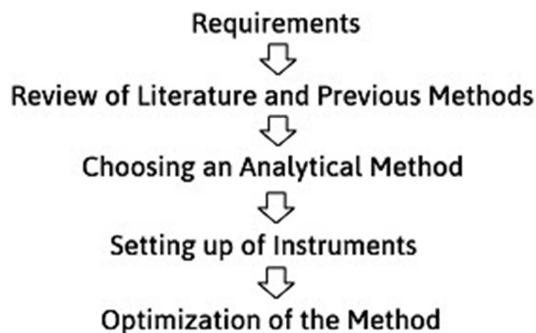
2. Analytical Research

Analytical research is a specific type of research that involves critical thinking skills and the evaluation of facts and information relative to the research being conducted. Students, doctors and psychologists use analytical research during studies to find the most relevant information. From analytical research, a person finds out critical details to add new ideas to the material being produced.

Some researchers conduct analytical research to find supporting evidence to current research being done in order to make the work more reliable. Other researchers conduct analytical research to form new ideas about the topic being studied. Analytical research is conducted in a variety of ways including literary research, public opinion, scientific trials and Meta-analysis.

Example:

Pharmaceutical analysis to determine the quality of drug products via analytical chemistry.



3. Applied Research

Applied research refers to scientific study and research that seeks to solve practical problems. It is defined as a research which is used to answer a specific question, determine why something failed or succeeded, problem related to product development or gain better understanding.

It examines a specific set of circumstances and its ultimate goal is relating the results to a particular situation.

Example:

- Improve agricultural crop production
- Treat or cure a particular disease
- Improve energy efficiency of homes, transportation (solar power, electric vehicles)
- Diagnose the low use of a particular drug

4. Fundamental Research

Fundamental research, also called pure research or basic research, is a type of scientific research with the aim of improving scientific theories for better understanding and prediction of natural or other phenomena. In contrast, applied research uses scientific theories to develop technology or techniques which can be used to intervene and alter natural or other phenomena.

It focuses on creating and supporting theories that explain observed phenomena. Pure research is the source of most new scientific ideas and ways of thinking about the world. Basic research generates new ideas, principles, and theories, which may not be immediately utilized but form the basis of progress and development in different fields.

Example:

Today's computers, for example, could not exist without research in pure mathematics conducted over a century ago, for which there was no known practical

application at the time. Basic research rarely helps practitioners directly with their everyday concerns; nevertheless, it stimulates new ways of thinking that have the potential to revolutionize and dramatically improve how practitioners deal with a problem in the future.

5. Quantitative Research

Quantitative research is referred to as the process of collecting as well as analyzing numerical data. It is generally used to find patterns, averages, predictions, as well as cause-effect relationships between the variables being studied. It is also used to generalise the results of a particular study to the population in consideration. Quantitative market research is widely used in science; both natural and social sciences. It is formed from a deductive approach. Quantitative data is any data that is in numerical form such as statistics, percentages. The researcher analyses the data with the help of statistics and hopes the numbers will yield an unbiased result that can be generalized to some larger population

Examples:

- How much has the average temperature changed globally over the last 5 years?
- Does environmental pollution affect newborn children?
- Has working from home during the pandemic improved the productivity of employees? Does this increase in productivity have anything to do with the cut down in travelling time of the employees?

6. Qualitative Research

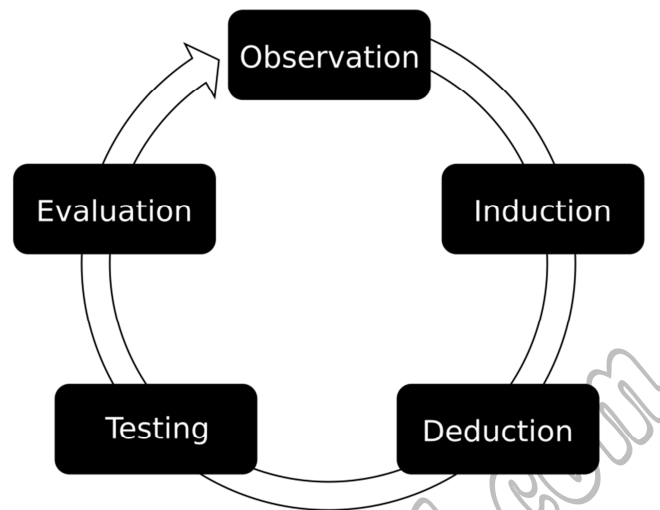
Qualitative research involves collecting and analyzing non-numerical data (e.g., text, video, or audio) to understand concepts, opinions, or experiences. It can be used to gather in-depth insights into a problem or generate new ideas for research. It deals with phenomena that are difficult or impossible to quantify mathematically, such as beliefs, meanings, attributes, and symbols. Qualitative researchers aim to gather an in-depth understanding of human behaviour and the reasons that govern such behaviour. The qualitative method investigates the why and how of decision making, not just what, where, when.

Qualitative research is the opposite of quantitative research, which involves collecting and analyzing numerical data for statistical analysis.

Examples:

- How does social media shape body image in teenagers?
- How do children and adults interpret healthy eating in the UK?
- How is anxiety experienced around the world?

7. Empirical Research



Empirical research is a type of research methodology that makes use of verifiable evidence in order to arrive at research outcomes. In other words, this type of research relies solely on evidence obtained through observation or scientific data collection methods. Empirical research is research that is based on observation and measurement of phenomena, as directly experienced by the researcher. The data thus gathered may be compared against a theory or hypothesis, but the results are still based on real life experience.

The Criteria of a good research

1. The purpose of research or the problem involved should be clearly defined. The statement of research problem should have analysis into its simplest element, its scope and limitations. If the researcher failed to do this adequately, he will raise the doubts in readers' minds.
2. It is important to write the research procedure in sufficient detail in order to let another researcher repeat the research.
3. The design of procedure should be in order to gain objective results. Direct experiments should have satisfactory controls. Direct observations should be recorded in writing as soon as possible after the event.
4. The researcher should report with complete explication, demerits in the procedural design and estimate their effect upon the findings. Some demerits effect on data and make them unreliable or lack validity.
5. An analysis of data should be completely enough to reveal its significance and the

method of analysis used should be appropriate. The validity and reliability of data should be checked carefully. The data should be classified in a way that the research reaches good conclusions. When the statistical methods are used the probability of errors should be evaluated and the criteria of statistical significance applied.

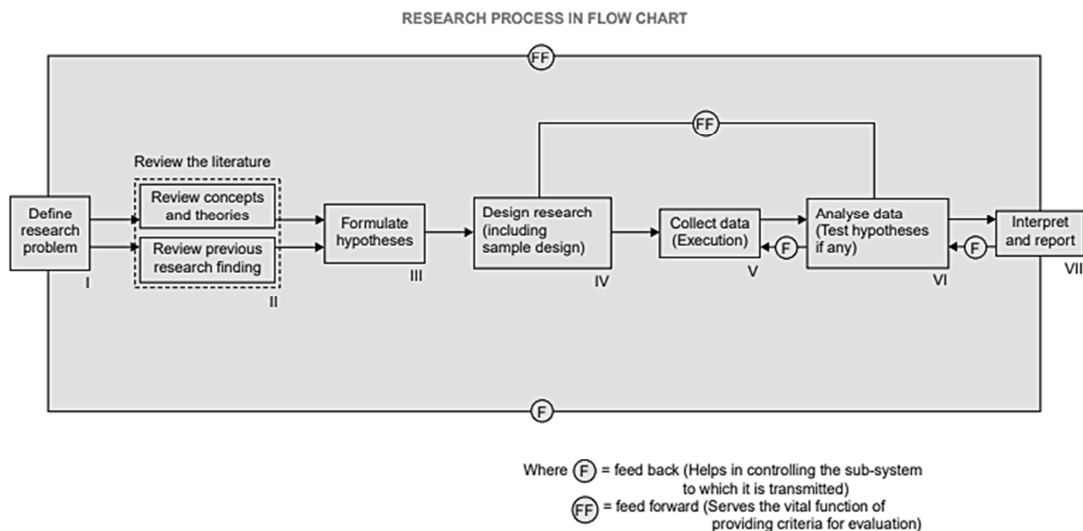
6. "Conclusions should be confined to those justified by the data of the research and limited to those for which the data provide an adequate basis." Researchers are often tempted to expand the bases of exhorting by including personal experiences not subject to the controls under which the research data were gathered. This tends to decrease the objectivity of the research and weaken confidence in findings.
7. If the researcher is honest, a greater confidence in the research is warranted. Were it is possible for the readers of a research report to get an enough information about the researcher, this criterion would be a good bases for judging the degree of confidence a piece of research warrants. For this reason, the research should accompanied by more information about the researcher.

Research Process

The Research Process is a process of multiple scientific steps in conducting the research work. Each step is interlinked with other steps. The process starts with the research problem at first. Then it advances in the next steps sequentially. Generally, a researcher conducts research work within seven steps. In research work, primarily, you require a Research Proposal. It is because the proposal approves the research project whether you achieve the ability to conduct research or not. So when you write a research proposal, present the detailed plans and specific objectives of your research correctly.



Steps of the Research Process



1. Identifying the Research Problem

The first step in the process is to identify a problem or develop a research question. The research problem may be something the agency identifies as a problem, some knowledge or information that is needed by the agency or the desire to identify a recreation trend nationally.

A research problem is a statement about an area of concern, a condition to be improved, a difficulty to be eliminated, or a troubling question that exists in scholarly literature, in theory, or in practice that points to the need for meaningful understanding and deliberate investigation.

2. Review the Literature

Once the research problem is identified and defined, the next step is to review the existing research. The researcher must learn more about the topic under investigation. To do this, the researcher must review the literature related to the research problem. This step provides foundational knowledge about the problem area. The review of literature also educates the researcher about what studies have been conducted in the past, how these studies were conducted, and the conclusions in the problem area.

Get Background Information:

- Read about your topic using websites or encyclopedias.
- It introduces you to the topic, helps you to focus on its key elements and can help you decide to broaden or narrow your focus.
- These sources often include bibliographies that you can refer to find more sources on your topic.

For example, in the obesity study, the review of literature enables the programmer to

discover statistics related to the long-term effects of childhood obesity in terms of health issues, death rates, and projected medical costs.

The information discovered during this step helps the programmer fully understand the magnitude of the problem, recognize the future consequences of obesity, and identify a strategy to combat obesity.

3. Formulating Objectives & Hypothesis

In this step, the researcher makes the problem precise.

- The research work is topic focused and refined.
- Then the researcher steps forward to how the problem would be approached? The nature of the research problem can decide to formulate a definite hypothesis.
- A hypothesis is tested. Effective research work formulates a hypothesis in such a way that collected factual data will provide evidence that either supports or disproves them. Formulation of Hypothesis in Research will make you more expert.
- In the end, the hypothesis turns into a practical theory.

In order to develop working hypotheses researcher should adopt the following approach

- a. Discussions with colleagues and experts about the problem, its origin and the objectives in seeking a solution;
- b. Examination of data and records, if available, concerning the problem for possible trends, peculiarities and other clues;
- c. Review of similar studies in the area or of the studies on similar problems; and
- d. Exploratory personal investigation which involves original field interviews on a limited scale with interested parties and individuals with a view to secure greater insight into the practical aspects of the problem.

4. Research Design

Research design decides how the research materials will be collected. One or more research methods, for example, experiment, survey, interview, etc. are chosen depending on the research objectives. Research Design actually provides insights into “how” to conduct research using a particular Research Methodology. Basically, every researcher has a list of research questions that need to be assessed that can be done with research design.

Function of research design is to provide for the collection of relevant evidence with minimal expenditure of effort, time and money. This depends mainly on the research purpose. Research purposes may be grouped into four categories,

- a. Exploration,
- b. Description,
- c. Diagnosis, and
- d. Experimentation.

For example, Experimental and hypothesis testing

Experimental designs can be either informal designs (such as before-and after without control, after-only with control, before-and-after with control)

Formal designs (such as completely randomized design, randomized block design, Latin square design, simple and complex factorial designs), out of which the researcher must select one for his own project.

5. Carry out Research Process

While the research design is decided, then the researcher collects data, records information. The researcher proceeds with the research. Practical difficulties may arise in this stage. For example, the research method may not suit properly. The interviewer might be unwilling to let carry out the research as planned. Moreover, a false interpretation could potentially bias the result of the study. So, when you collect data, you need to know the effective techniques of data collection in order to gather necessary and relevant information with regard to research.

6. Preparing Research Results

Interpret your research results in order to report the findings. No matter what kind of research you are doing, there comes a moment when your head is full of ideas that originated from your analysis. Ideally, you'll write them down as they come to you. Now you need to convert the mass of those elements and ideas into a written text that makes sense to the reader and can do justice to your research question.

7. Reporting Research Findings

The final step of the research process outline is to report the research findings. Describe the significance of the research study. Work out how they relate to the previous research findings. Usually, the research report is published as a journal article or book. This is the last stage in terms of the individual research project. Mostly, a research report discusses questions that remained unanswered & suggest further research in the future in general.

Research Gap

A *research gap* is a question or a problem that has not been answered by any of the existing studies or research within your field. Sometimes, a research gap exists when there is a concept or new idea that hasn't been studied at all. Sometimes you'll find a research gap if all the existing research is outdated and in need of new/updated research (studies on Internet use in 2001, for example). Or, perhaps a specific population has not been well studied (perhaps there are plenty of studies on teenagers and video games, but not enough studies on toddlers and video games, for example). These are just a few examples, but any research gap you find is an area where more studies and more research need to be conducted.

If you are a young researcher, or even still finishing your studies, you'll probably notice that your academic environment revolves around certain research topics, probably linked to your department or to the interest of your mentor and direct colleagues. For example, if your department is currently doing research in nanotechnology applied to medicine, it is only natural that you feel compelled to follow this line of research. Hopefully, it's something you feel familiar with and interested in – although you might take your own twists and turns along your career.

Many scientists end up continuing their academic legacy during their professional careers, writing about their own practical experiences in the field and adapting classic methodologies to a present context. However, each and every researcher dreams about being a pioneer in a subject one day, by discovering a topic that hasn't been approached before by any other scientist. This is a *research gap*. Research gaps are particularly useful for the advance of science, in general. Finding a research gap and having the means to develop a complete and sustained study on it can be very rewarding for the scientist.

Formulating a Research Problem

A *research problem* is a statement about an area of concern, a condition to be improved, a difficulty to be eliminated, or a troubling question that exists in scholarly literature, in theory, or in practice that points to the need for meaningful understanding and deliberate investigation. In some social science disciplines the research problem is typically posed in the form of a question. A research problem does not state how to do something.

Statement of a research problem: An adequate statement of your research problem plays an important role in the success of your academic paper and study. It's possible to generate a number of researchable issues from the same subject because there are many issues that may arise out of it. Your study should pursue only one in detail.

Characteristics of research problem:

- Reflecting on important issues or needs;
- Basing on factual evidence (it's non-hypothetical)
- Being manageable and relevant
- Suggesting a testable and meaningful hypothesis.

Formulating the research problem: Formulating the research problem enables you to make a purpose of your study clear to yourself and target readers. Focus your paper on providing relevant data to address it. A problem statement is an effective and essential tool to keep you on track with research and evaluate it. We can consider 5 ways to formulate a powerful research problem:

1. Specify your research objectives
2. Review its context or environment
3. Explore its nature
4. Determine variable relationships
5. Possible consequences of alternative approaches

1. Specific research objectives

A clear statement that defines all objectives can help you conduct and develop effective and meaningful research. They should be manageable to bring you success. A few goals will help you keep your study relevant. This statement also helps guides to evaluate the questions your research project answers and different methods that you use to address them.

2. Review context or environment

It's necessary to work hard to define and test all kinds of environmental variables to make your project successful. This step helps to define if the important findings of the study will deliver enough data to be worth considering. Identify specific environmental variables that may potentially affect your research and start formulating effective methods to control all of them.

3. Explore the nature of research problem

Research problems may range from simple to complex, and everything depends on a range of variables and their relationships. Some of them can be directly relevant to specific research questions, while others are completely unimportant for your project.

4. Determine variable relationships

Scientific, social, and other studies often focus on creating a certain sequence of repeating behaviors over time. Identify the variables that affect possible solutions to your research problem. Decide on the degree to which we can use and control all of them for study purposes. Choose the most critical variables for a solution of your research problem.

During the formulation stage, it's necessary to consider and generate as many potential approaches and variable relationships as you can.

5. Consequences of alternative approaches

There are different consequences that each course of action or approach can bring and we need to anticipate them. Referring scientific papers to notice their research questions are crucial for determining the quality of answers, methods, and findings.

Defining a Research Problem

Defining a research problem is the fuel that drives the scientific process, and is the foundation of any research method and experimental design, from true experiment to case study. It is one of the first statements made in any research paper and, as well as defining the research area, should include a quick synopsis of how the hypothesis was arrived at. Defining a research problem is crucial in defining the quality of the answers, and determines the exact research method used.

A scientist may even review a successful experiment, disagree with the results, the tests used, or the methodology, and decide to refine the research process, retesting the hypothesis.

This is called the conceptual definition, and is an overall view of the problem. A science report will generally begin with an overview of the previous research and real-world observations. The researcher will then state how this led to defining a research problem.

The Operational Definitions: The operational definition is the determining the scalar properties of the variables.

For example, temperature, weight and time are usually well known and defined, with only the exact scale used needing definition. If a researcher is measuring abstract concepts, such as intelligence, emotions, and subjective responses, then a system of measuring numerically needs to be established, allowing statistical analysis and replication. Intelligence may be measured with IQ and human responses could be measured with a questionnaire from '1- strongly disagree', to '5 - strongly agree'.

Behavioral biologists and social scientists might design an ordinal scale for measuring and rating behavior. These measurements are always subjective, but allow statistics and replication of the whole research method. This is all an essential part of defining a research problem.

Research methods v/s methodology

Researchers implement systematic methods to conduct a research. They define the research topic to establish a deeper and clearer understanding in the methods section. Furthermore, methods consist of all techniques, strategies, and tools employed by a researcher to complete the experiment and find solution to the research problem.

However, methodology is a systematic and theoretical approach to collect and evaluate data throughout the research process. It allows researchers to validate a study's rigor to acquire new information. The purpose of research methodology is to prove the credibility, validity, and reliability of a chosen research method.

Research Methods	Research Methodology
The objective of methods is to find solution to the research problem.	The objective of methodology is to determine appropriateness of the methods applied with a view to ascertain solution.
Methods are just behaviour or tools used to select a research technique.	Methodology is analysis of all the methods and procedures of the investigation.
Methods are applied during the later stage of the research study.	Methodologies are applied during the initial stage of the research process.
It comprises different investigation techniques of the study.	It is a systematic strategy to find solution to the research problem.
Methods encompasses of carrying out experiments, conducting surveys, tests, etc.	Methodology encompasses several techniques used while conducting these experiments, surveys, tests, etc.

Research Protocol

A research protocol outlines the plan for how a study is run. The study plan is developed to answer research questions. It provides evidence for feasibility of a study, detailed objectives, design, methodology, statistical considerations and how the study will be conducted and evaluated. A well-written and complete protocol is essential for a high quality study, ensures clarity as to what has been ethically approved and will make publishing the results easier.

The research protocol need not be lengthy, but should include the following minimum information:

- Background information
- Aim(s) and hypothesis
- Study objective
- Study plan and procedures
- Statistical analysis

During the process of the development of the protocol, investigators can and should try to benefit from the advice of colleagues and experts in refining their plans. But once a protocol for the study has been developed and approved, and the study has started and progressed, it should be adhered to strictly and should not be changed. This is particularly important in multi-centre studies. Violations of the protocol can discredit the whole study.

Format for the protocol

The research protocol is generally written according to the following format.

- Project title
- Project summary
- Project description:
 - Rationale
 - Objectives
 - Methodology
 - Data management and analysis
- Ethical considerations
- Gender issues
- References

Project title: The title should be descriptive and concise. It may need to be revised after completion of the writing of the protocol to reflect more closely the sense of the study.

Project summary: The summary should be concise, and should summarize all the elements of the protocol. It should stand on its own, and not refer the reader to points in the

project description.

Project description

- **Rationale:** This is equivalent to the introduction in a research paper. It puts the proposal in context. It should answer the question of why and what: why the research needs to be done and what will be its relevance. A brief description of the most relevant studies published on the subject should be provided to support the rationale for the study.
- **Objective(s):** Specific objectives are statements of the research question(s). Objectives should be simple (not complex), specific (not vague), and stated in advance (not after the research is done). After statement of the primary objective, secondary objectives may be mentioned.
- **Methodology:** The methodology section has to be thought out carefully and written in full detail. It is the most important part of the protocol. It should include information on the research design, the research subjects, interventions introduced, observations to be made and sample size.
 - **Research design:** The choice of the design should be explained in relation to the study objectives.
 - **Research subjects or participants:** Depending on the type of the study, the following questions should be answered:
 - What are the criteria for inclusion or selection?
 - What are the criteria for exclusion?
 - In intervention studies, how will subjects be allocated to index and comparison groups?
 - What are the criteria for discontinuation?
 - **Interventions:** If an intervention is introduced, a description must be given of the drugs or devices to be used, and whether they are already commercially available, or in phases of experimentation. For drugs and devices that are commercially available, the protocol must state their proprietary names, manufacturer, chemical composition, dose and frequency of administration.
 - **Observations:** Information should be provided on the observations to be made, how they will be made, and how frequently will they be made. If the observation is made by a questionnaire, this should be appended to the protocol. Laboratory or other diagnostic and investigative procedures should be described. For established procedures, reference to appropriate published work is enough. For new or modified procedures, an adequate description is needed, with a justification for their use.
 - **Sample size:** The protocol should provide information and justification about sample size. A larger sample size than needed to test the research hypothesis increases the cost and duration of the study and will be unethical if it exposes human subjects to

any potential unnecessary risk without additional benefit. A smaller sample size than needed can also be unethical if it exposes human subjects to risk with no benefit to scientific knowledge.

- **Data management and analysis:** The protocol should provide information on how the data will be managed, including data coding for computer analysis, monitoring and verification. Information should also be provided on the available computer facility. The statistical methods used for the analysis of data should be clearly outlined.

Ethical considerations: Ethical considerations apply to all types of health research. These include research involving human experimentation, whether the research is of therapeutic or diagnostic nature that is carried out on patients who may expect a potential benefit from their participation, or is of a purely scientific nature for which human subjects volunteer to advance medical science but will not draw any therapeutic or diagnostic benefit. There are also ethical considerations for research involving human subjects but not experimentation. Epidemiological, field and qualitative studies fall under this category. Although no experimentation is involved, such studies can be as intrusive on the individual's privacy and even on communities. The ethics of research involving experimentation on animals has been receiving proper and increasing attention recently.

Gender issues: The Commission on the Status of Women made the statement: "Ensure, where indicated, that clinical trials of pharmaceuticals, medical devices and other medical products include women with their full knowledge and consent and ensure that the resulting data is analysed for sex and gender differences."

Women were often excluded from clinical trials on disease conditions that affect both men and women, on the basis of biological variability, and/or vulnerability. But women were given the same drugs, which had not been tested on them, as men if the drugs proved safe and effective for men.

Drugs and devices intended for use by women only were sometimes tested on them without their proper informed consent, particularly in poor resource settings.

When women were included with men as research subjects, gender was not always taken into consideration when results were analysed.

References: The protocol should end with relevant references on the subject.

Research Ethics

Research ethics involves the application of fundamental ethical principles to research activities which include the design and implementation of research, respect towards society and others, the use of resources and research outputs, scientific misconduct and the regulation of research. Research ethics is not a mere “formality,” which is required by academic journal editors, but it is a significant part of research, which is influenced by both the general trust in scientists, data protection, anonymity, and confidentiality, and the ability to build trust-based relationship with the respondents and retain it.

- **Honesty:** Strive for honesty in all scientific communications. Honestly report data, results, methods and procedures, and publication status. Do not fabricate, falsify, or misrepresent data. Do not deceive colleagues, research sponsors, or the public.
- **Objectivity:** Strive to avoid bias in experimental design, data analysis, data interpretation, peer review, personnel decisions, grant writing, expert testimony, and other aspects of research where objectivity is expected or required. Avoid or minimize bias or self-deception. Disclose personal or financial interests that may affect research.
- **Integrity:** Keep your promises and agreements; act with sincerity; strive for consistency of thought and action
- **Carefulness:** Avoid careless errors and negligence; carefully and critically examine your own work and the work of your peers. Keep good records of research activities, such as data collection, research design, and correspondence with agencies or journals.
- **Openness:** Share data, results, ideas, tools, resources. Be open to criticism and new ideas.
- **Respect for Intellectual Property:** Honor patents, copyrights, and other forms of intellectual property. Do not use unpublished data, methods, or results without permission. Give proper acknowledgement or credit for all contributions to research. Never plagiarize.
- **Confidentiality:** Protect confidential communications, such as papers or grants submitted for publication, personnel records, trade or military secrets, and patient records.
- **Responsible Publication:** Publish in order to advance research and scholarship, not to advance just your own career. Avoid wasteful and duplicative publication.
- **Responsible Mentoring:** Help to educate, mentor, and advise students. Promote their welfare and allow them to make their own decisions.
- **Respect for colleagues:** Respect your colleagues and treat them fairly.
- **Social Responsibility:** Strive to promote social good and prevent social harms through research, public education, and advocacy.

- **Non-discrimination:** Avoid discrimination against colleagues or students on the basis of sex, race, ethnicity, or other factors not related to scientific competence and integrity.
- **Competence:** Maintain and improve your own professional competence and expertise through lifelong education and learning; take steps to promote competence in science as a whole.
- **Legality:** Know and obey relevant laws and institutional and governmental policies.
- **Animal Care:** Show proper respect and care for animals when using them in research. Do not conduct unnecessary or poorly designed animal experiments.
- **Human subject's protection:** When conducting research on human subjects, minimize harms and risks and maximize benefits; respect human dignity, privacy, and autonomy; take special precautions with vulnerable populations; and strive to distribute the benefits and burdens of research fairly.

Institutional Review Boards (IRB)

Under FDA regulations, an Institutional Review Board is a group that has been formally designated to review and monitor biomedical research involving human subjects. In accordance with FDA regulations, an IRB has the authority to approve, require modifications in (to secure approval), or disapprove research. This group review serves an important role in the protection of the rights and welfare of human research subjects.

The purpose of IRB review is to assure, both in advance and by periodic review, that appropriate steps are taken to protect the rights and welfare of humans participating as subjects in the research. To accomplish this purpose, IRBs use a group process to review research protocols and related materials (e.g., informed consent documents and investigator brochures) to ensure protection of the rights and welfare of human subjects of research.

Regulations: Good Clinical Practice and Clinical Trials: Comprehensive list of regulations governing human subject protection and the conduct of clinical trials.

Guidance for Institutional Review Boards and Clinical Investigators: A series of Information Sheets providing the Agency's current guidance on the protection of people who are subjects of research. Topics include Institutional Review Boards and Sponsor-Investigator-IRB interrelationships, FDA clinical investigator inspections and sanctions, clinical trials protocols, informed consent, and documents necessary for the conduct of clinical trials.

Information for Health Professionals: Additional links to information on subject protection from FDA and other government agencies.

Clinical Safety Data Management: Provides the definitions and terminology associated with clinical safety experience and the standards for expedited reporting of adverse drug reactions that occur during clinical trials.

Significance of research in Pharmaceutical Sciences

Different branches of science are expanding very fast and there are vast volumes of new information which are progressively discovered and added to the present human knowledge. With no doubt, pharmaceutical sciences is amongst the most dynamic disciplines of science that its content is attributed to different basic and applied researches including studies in the field of physical and organic chemistry, engineering, biochemistry, biology, pharmacology and pharmacotherapy. These researches are aimed to understand how to develop new drugs, optimize their delivery to the body and translate these integrated understanding into new therapies against human disease as well as improved community health.

Clinical observations about the nature, pathogenesis and development of diseases as well as responses to and complications caused by therapies can effectively drive basic science investigations. Knowledge obtained by basic science then provides clinical practitioners with new treatment strategies. This continuous translational research cycle can hopefully result in more rational drug design, improved efficacy of therapeutic agents, and accelerated optimization of investigational compounds for clinical use, which can finally be used in the public health sphere. Besides, this will lead to a more cost-effective drug discovery process.

→ Pharmacy Practice

Patient oriented research conducted with human subjects; it includes mechanisms of human disease, therapeutic interventions, clinical trials, development of new technologies.

→ Health Services

Improving the way health care services are organized, regulated, managed, financed, paid for, used and delivered, in the interest of improving the health and quality of life.

→ Pharmacoepidemiology

Provides an estimate of the probability of beneficial effects of a drug in a population and the probability of adverse effects

→ Computational Drug Discovery

Research in this area uses tools to discover new antiviral and immune checkpoints' small molecule drugs and specializes in understanding the nature and biophysical processes underlying protein-drug interaction, protein-protein interactions, protein-DNA interactions, drug off-target interactions and predicting drug-mediated toxicity.

→ Drug Delivery

A field that concentrates on formulation methods and technologies that are employed to transport pharmaceutical compounds to various sites in the body. It may include synthetic drug carrier systems capable of improving the bioavailability and/or tissue-specific targeting of an active pharmaceutical ingredient (API) to the site of desired drug action, such as bone drug delivery, or drug targeting of cancer cells.

→ Pharmacodynamics

Studies what a drug does to the body. Drug effect involves receptor binding, post-receptor effects, and biochemical interactions. It is a drug's pharmacokinetics that determines the onset, duration, and intensity of a drug's effect.

→ Pharmacokinetics

Studies what the body does to a drug. This field deals with the movement of a drug over time into, through, and out of the body via the processes of absorption, distribution, metabolism, excretion, and transport.

vipinsicp@gmail.com

Statistical data

The data collected for the different types of studies are analysed to assess changes in health or disease situations in the community or population by standard parameters (descriptive and inferential statistics). The statistical data obtained from different sources can be divided into two broad categories:

1. Qualitative
2. Quantitative

1. Qualitative (Discrete) data

There is no magnitude or size of the characteristic or attribute as the same cannot be measured. They are classified by frequency of samples having the same characteristic. Samples with same characteristic are classified into groups such as died, cured, relieved, vaccinated, treated, not treated, on drug, on placebo etc.

In medical studies such data are mostly collected in pharmacology to find the action of a drug, in clinical practice to test or compare the efficacy of a drug, vaccine or line of treatment and in demography to find vital statistics.

2. Quantitative (Continuous) data

Quantitative data is data that can be counted or measured in numerical values. Quantitative data is numerical in nature and can be mathematically computed. Height in feet, age in years, and weight in pounds are examples of quantitative data. Quantitative approaches have the advantage that they are cheaper to implement, are standardized so comparisons can be easily made and the size of the effect can usually be measured.

Primary data

It is the first hand information. Data that has been collected from first-hand-experience is known as primary data. Primary data has not been published yet and is more reliable, authentic and objective. Primary data has not been changed or altered by human beings; therefore its validity is greater than secondary data.

In statistical surveys it is necessary to get information from primary sources and work on primary data. For example, the statistical records of female population in a country cannot be based on newspaper, magazine and other printed sources. A research can be conducted without secondary data but a research based on only secondary data is least reliable and may have biases because secondary data has already been manipulated by human beings. One of such sources is old and secondly they contain limited information as well as they can be misleading and biased.

Sources: Experiments, Survey, Questionnaire, Interview, Observations.

Secondary data

Data collected from a source that has already been published in any form is called as secondary data. The review of literature in any research is based on secondary data. It is collected by someone else for some other purpose (but being utilized by the investigator for another purpose). For examples, Census data being used to analyze the impact of education on career choice and earning.

Sources: Books, Records, Biographies, Newspapers, Published censuses or other statistical data, Data archives, Internet articles, Research articles by other researchers (journals), Databases, etc.

Methods of data collection

There are many methods of collecting data. The main methods include –

1. Questionnaire method
2. Interviews
3. Focus Groups
4. Observation
5. Survey
6. Case-studies
7. Experimental Method

1. Questionnaire method

A questionnaire is a research instrument consisting of a series of questions and other prompts for the purpose of gathering information from respondents. Although they are often designed for statistical analysis of the responses, this is not always the case. It has standardized answers that make it simple to compile data. Informants are expected to read and understand the questions and reply in the space provided in the questionnaire itself. The questionnaire should deal with an important or significant topic to create interest among respondents. It should be designed to collect information which can be used subsequently as data for analysis.

2. Interviews

In social science, interviews are a method of data collection that involves two or more people exchanging information through a series of questions and answers. The questions are designed by a researcher to elicit information from interview participants on a specific topic or set of topics. These topics are informed by the author's research questions. Interviews typically involve an in-person meeting between two people (an interviewer and an interviewee), but interviews need not be limited to two people, nor must they occur in-

person. Interviews are also useful when your topic is rather complex, requires lengthy explanation, or needs a dialogue between two people to thoroughly investigate. Additionally, interviews may be the best method to utilize if your study involves describing the process by which a phenomenon occurs, like how a person makes a decision. For example, you could use interviews to gather data about how people reach the decision not to have children and how others in their lives have responded to that decision.

3. Focus groups

A focus group is a research technique used to collect data through group interaction. The group comprises a small number of carefully selected people who discuss a given topic. Focus groups are used to identify and explore how people think and behave, and they throw light on why, what and how questions. Group discussions are especially useful for researching new products, testing new concepts.

4. Observation

The observation method of data collection involves seeing people in a certain setting or place at a specific time and day. Essentially, researchers study the behavior of the individuals or surroundings in which they are analyzing. This can be controlled, spontaneous, or participant-based research. This data collection method does not require researchers' technical skills when it comes to data gathering. When a researcher utilizes a defined procedure for observing individuals or the environment, this is known as structured observation. When individuals are observed in their natural environment, this is known as naturalistic observation.

5. Survey

A survey is a data collection tool that lists a set of structured questions to which respondents provide answers based on their knowledge and experiences. It is a standard data gathering process that allows you to access information from a predefined group of respondents during research.

In a survey, you would find different types of questions based on the research context and the type of information you want to have access to. Many surveys combine open-ended and closed-ended questions including rating scales and semantic scales. This means you can use them for qualitative and quantitative research.

6. Case studies

A case study is a method of obtaining in-depth information on a person, group or phenomenon to provide descriptions of specific or rare cases. Case studies allow for the development of novel hypotheses for later testing, provide detailed descriptions of rare events, and can explore the intricacies of existing theories of causation. Case studies cannot

directly indicate cause and effect relationships or test hypotheses. In addition, findings from case studies cannot be generalized to a wider population.

7. Experimental method

An experiment is a data collection method where you as a researcher change some variables and observe their effect on other variables. The variables that you manipulate are referred to as independent while the variables that change as a result of manipulation are dependent variables. Imagine a manufacturer is testing the effect of drug strength on number of bacteria in the body. The company decides to test drug strength at 10mg, 20mg and 40mg. In this example, drug strength is the independent variable while number of bacteria is the dependent variable. The drug administered is the treatment, while 10mg, 20mg and 40mg are the levels of the treatment.

Experimental studies

Experimental studies are ones where researchers introduce an intervention and study the effects. Experimental studies are usually randomized, meaning the subjects are grouped by chance. In an experiment is a study, a treatment, procedure, or program is intentionally introduced and a result or outcome is observed. The American Heritage Dictionary of the English Language defines an experiment as “A test under controlled conditions that is made to demonstrate a known truth, to examine the validity of a hypothesis, or to determine the efficacy of something previously untried.”

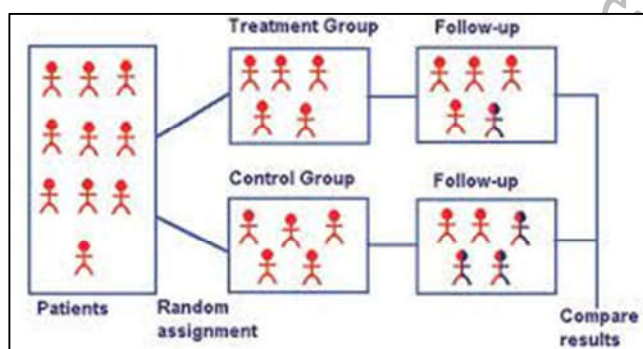
True experiments have four elements: manipulation, control, random assignment, and random selection. The most important of these elements are manipulation and control. Manipulation means that something is purposefully changed by the researcher in the environment. Control is used to prevent outside factors from influencing the study outcome. When something is manipulated and controlled and then the outcome happens, it makes us more confident that the manipulation “caused” the outcome. In addition, experiments involve highly controlled and systematic procedures in an effort to minimize error and bias which also increases our confidence that the manipulation “caused” the outcome.

Types of experimental studies:

1. Randomized Controlled Clinical Trial

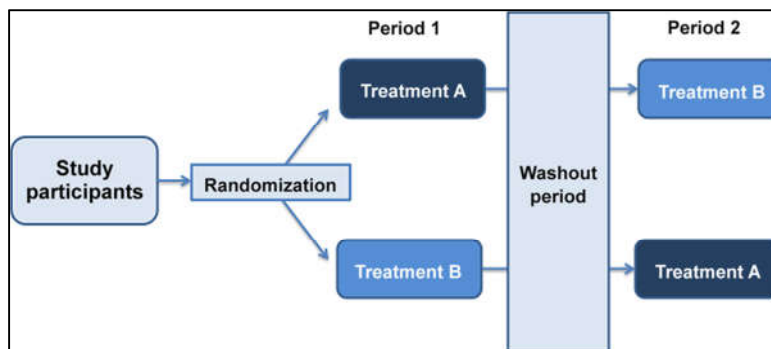
Randomized controlled trials (RCTs) are intervention studies in which a group of subjects with similar characteristics are randomized to receive one of several defined interventions. RCTs are powerful tools to test a hypothesis. These studies are often the basis for evidence-based medicine practice, and are considered the gold-standard in clinical research.

At the outset of a randomised controlled trial the criteria for entry to the study sample must be specified (for example, in terms of age, sex, diagnosis, etc). As in other epidemiological investigations, the subjects studied should be representative of the target population in whom it is hoped to apply the results. In comparison of two treatments for rheumatoid arthritis in a series of hospital patients, subjects who satisfy the entry criteria are asked to consent to participation. Those subjects who agree to participate are then randomised to the treatments under comparison. Thus in a study comparing two treatments, A and B, patients might be randomised in blocks of six. Of the first six patients entering the trial, three would be allocated to treatment A and three to treatment B – which patient received which treatment being determined randomly.



2. Crossover studies

Another modification of the randomised controlled trial is the crossover design. This is particularly useful when outcome is measured by reports of subjective symptoms, but it can only be applied when the effects of treatment are short lived (for example, pain relief from an analgesic). In a crossover study, eligible patients who have consented to participate receive each treatment sequentially, often with a “wash out” period between treatments to eliminate any carry over effects. However, the order in which treatments are given is randomised so that different patients receive them in different sequence. Outcome is monitored during each period of treatment, and in this way each patient can serve as his own control.



3. Pre-experimental Research

In pre-experimental research design, either a group or various dependent groups are observed for the effect of the application of an independent variable which is presumed to cause change. It is the simplest form of experimental research design and is treated with no control group.

4. Quasi-experimental Research Design

The word "quasi" means partial or half. In quasi-experiments, the participants are not randomly assigned, and as such, they are used in settings where randomization is difficult or impossible. This is very common in educational research, where administrators are unwilling to allow the random selection of students for experimental samples.

5. True Experimental Research Design

The true experimental research design relies on statistical analysis to approve or disprove a hypothesis. It is the most accurate type of experimental design and may be carried out with or without a pretest on at least 2 randomly assigned dependent subjects. The true experimental research design must contain a control group, a variable that can be manipulated by the researcher, and the distribution must be random.

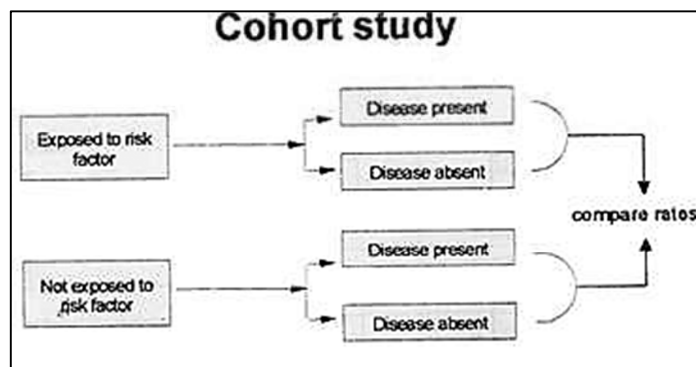


Observational studies

Observational studies are ones where researchers observe the effect of a risk factor, diagnostic test, treatment or other intervention without trying to change who is or isn't exposed to it. Cohort studies and case control studies are two types of observational studies.

1. Cohort study

For research purposes, a cohort is any group of people who are linked in some way. For instance, a birth cohort includes all people born within a given time frame. Researchers compare what happens to members of the cohort that have been exposed to a particular variable to what happens to the other members who have not been exposed.



a. Retrospective Cohort Study:

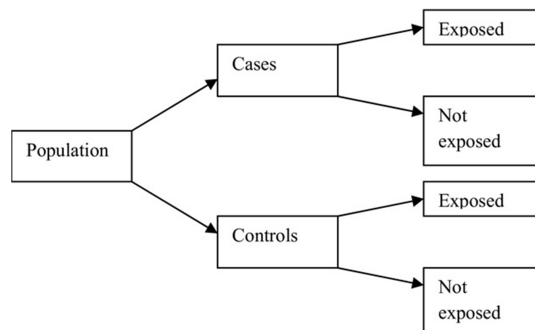
Retrospective cohort studies look at data that already exists and tries to identify risk factor for particular condition. Here the outcomes have occurred before starting the study. The study looks back into the past to try to determine why the participants have the disease or outcome.

b. Prospective Cohort Study:

The study is usually forward looking or planned in advance and carried out over a future period of time. A defined study cohort is followed forward in time to determine how factors/exposures affect a given outcome. Prospective studies can help identify risk factors for disease because data is collected at set time intervals. Draw back to these studies is loss to follow-up and length of time needed to determine associations.

2. Case control study

Observational study in which likelihood of exposure is compared between representative groups with (case) and without (control) a disease. Here researchers identify people with an existing health problem (“cases”) and a similar group without the problem (“controls”) and then compare them with respect to an exposure or exposures. Data can be conducted either retrospectively or prospectively. These studies can identify associations between a disease and risk factors or disease outcomes. Case-control studies generally require fewer subjects and are less expensive than cohort studies, and they can be particularly useful in epidemiologic investigations to identify risk factors for disease.



Questionnaire

A questionnaire is a research instrument consisting of a series of questions (or other types of prompts) for the purpose of gathering information from respondents through survey or statistical study. The questionnaire was invented by the Statistical Society of London in 1838. Questionnaires have advantages over some other types of surveys in that they are cheap, do not require as much effort from the questioner as verbal or telephone surveys, and often have standardized answers that make it simple to compile data. Questionnaires can be an effective means of measuring the behavior, attitudes, preferences, opinions and intentions of relatively large numbers of subjects more cheaply and quickly than other methods.

Rating Scales

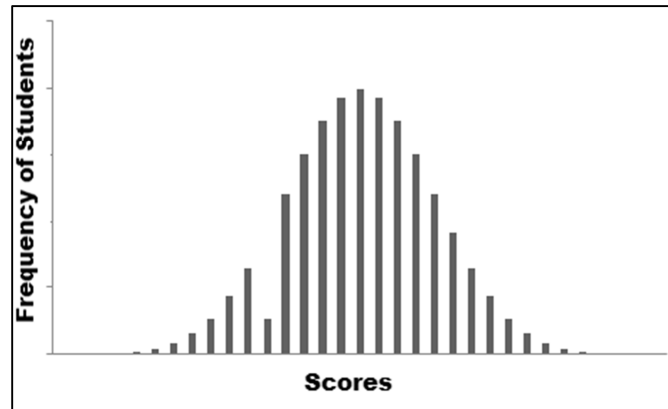
Rating scale is defined as a closed-ended survey question used to represent respondent feedback in a comparative form for specific particular features/products/services. It is one of the most established question types for online and offline surveys where survey respondents are expected to rate an attribute or feature. Rating scale is a variant of the popular multiple-choice question which is widely used to gather information that provides relative information about a specific topic.

Researchers use a rating scale in research when they intend to associate a qualitative measure with the various aspects of a product or feature. Generally, this scale is used to evaluate the performance of a product or service, employee skills, customer service performances, processes followed for a particular goal etc. Rating scale survey question can be compared to a checkbox question but rating scale provides more information than merely Yes/No.



Data distributions

There are two types of data distribution based on two different kinds of data: **Discrete** and **Continuous**. Discrete data distributions include binomial distributions, Poisson distributions, and geometric distributions. Continuous data distributions include normal distributions and the Student's t-distribution.



Normal Distribution:

When a **large** set of observation, of any variable characteristic such as height, weight, BP, Blood sugar etc. are arranged in ascending or descending order and make a frequency distribution keeping the group interval small, it can be seen that;

1. Majority of the observations will be seen in the middle around the measures of central tendency (mean, median and mode) and a fewer at the extremes.
2. The distribution is symmetric on both sides of the mean.
3. The mean, median and mode coincides
4. The frequency curve will be bell-shaped

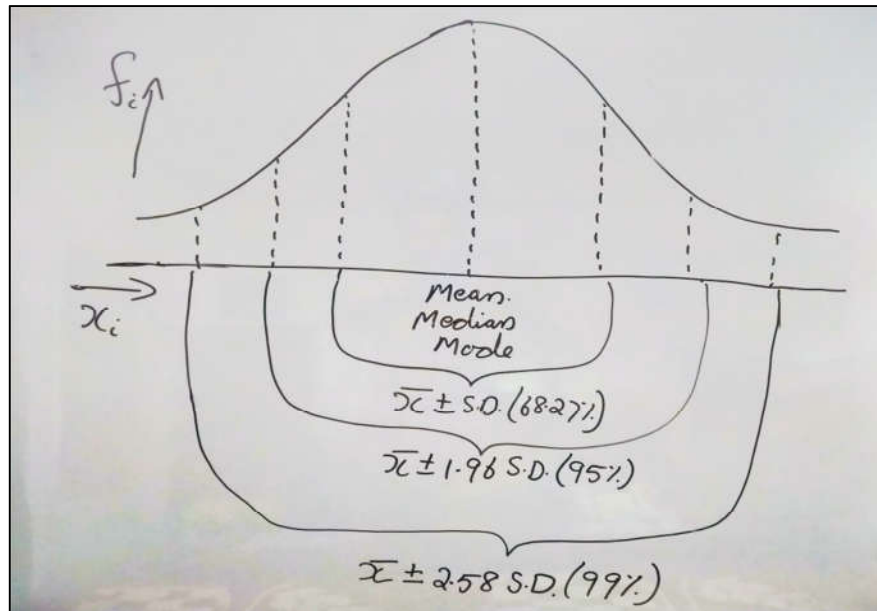
A distribution of this nature is called **Normal Distribution** or **Gaussian distribution**.

A curve of this shape is called **Normal Curve**. It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

The **Limits of Normality** can be arithmetically expressed in terms of **mean** and **standard deviation** as follows:

1. Mean \pm SD covers approximately 68.27 % of total observation.
2. Mean \pm **1.96** SD covers approximately 95 % of total observation.
3. Mean \pm **2.58** SD covers approximately 99 % of total observation.

Normal Curve



Properties of Normal Curve:

1. It is a bell shaped curve
2. It is a symmetric curve
3. It is a continuous curve. No part of the curve lies below the x-axis
4. The mean, median and mode coincides
5. The first and third quartiles are equidistant from the second.
6. Probability on either side are equal and is 0.5.

Processing of data: Coding and Tabulation

After the data have been collected, the researcher turns to the task of analyzing them. The data, after collection, has to be processed and analyzed in accordance with the outline laid down for the purpose at the time of developing the research plan. This is essential for a scientific study and for ensuring that we have all relevant data for making contemplated comparisons and analysis.

Coding: Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness (i.e., there must be a class for every data item) and also that of mutual exclusivity which means that a specific answer can be placed in one and only one cell in a given category set. Coding is necessary for efficient analysis and through it the several replies may be reduced to a small number of classes which contain the critical

information required for analysis.

Tabulation: It is a systematic & logical presentation of numeric data in rows and columns to facilitate comparison and statistical analysis. It facilitates comparison by bringing related information close to each other and helps in further statistical analysis and interpretation. In other words, the method of placing organised data into a tabular form is called as tabulation. It may be complex, double or simple depending upon the nature of categorisation.
(Give examples of tables: Frequency tables, Cross tables)

(Graphical Presentation of data: Refer assignment)

vipinsjcp@gmail.com

Epidemiology studies

Epidemiology is defined as the branch of medical science that deals with the incidence, distribution, determinant/characteristics and control of a disease in a population. Epidemiologic studies are the foundation for disease control and prevention through tracking the prevalence of the disease, characterizing the natural history, and identifying determinants or causes of the disease. . It defines risk factors for a disease and targets for preventive medicine.

Epidemiologic studies fall into two categories: experimental and observational.

Experimental studies

In an experimental study, the investigator determines through a controlled process the exposure for each individual (clinical trial) or community (community trial), and then tracks the individuals or communities over time to detect the effects of the exposure. For example, in a clinical trial of a new vaccine, the investigator may randomly assign some of the participants to receive the new vaccine, while others receive a placebo shot. The investigator then tracks all participants, observes who gets the disease that the new vaccine is intended to prevent, and compares the two groups (new vaccine vs. placebo) to see whether the vaccine group has a lower rate of disease. Similarly, in a trial to prevent onset of diabetes among high-risk individuals, investigators randomly assigned enrollees to one of three groups - placebo, an anti-diabetes drug, or lifestyle intervention. At the end of the follow-up period, investigators found the lowest incidence of diabetes in the lifestyle intervention group, the next lowest in the anti-diabetic drug group, and the highest in the placebo group.

Observational studies

In an observational study, the epidemiologist simply observes the exposure and disease status of each study participant. John Snow's studies of cholera in London were observational studies. The two most common types of observational studies are cohort studies and case-control studies; a third type is cross-sectional studies.

Cohort study: A cohort study is similar in concept to the experimental study. In a cohort study the epidemiologist records whether each study participant is exposed or not, and then tracks the participants to see if they develop the disease of interest. Note that this differs from an experimental study because, in a cohort study, the investigator observes rather than determines the participants' exposure status. After a period of time, the investigator compares the disease rate in the exposed group with the disease rate in the unexposed group. The unexposed group serves as the comparison group, providing an estimate of the baseline or expected amount of disease occurrence in the community. If the disease rate is substantively different in the exposed group compared to the unexposed group, the

exposure is said to be associated with illness.

The length of follow-up varies considerably. In an attempt to respond quickly to a public health concern such as an outbreak, public health departments tend to conduct relatively brief studies. On the other hand, research and academic organizations are more likely to conduct studies of cancer, cardiovascular disease, and other chronic diseases which may last for years and even decades. These studies are sometimes called follow-up or **prospective cohort studies**, because participants are enrolled as the study begins and are then followed prospectively over time to identify occurrence of the outcomes of interest.

An alternative type of cohort study is a **retrospective cohort study**. In this type of study both the exposure and the outcomes have already occurred. Just as in a prospective cohort study, the investigator calculates and compares rates of disease in the exposed and unexposed groups. Retrospective cohort studies are commonly used in investigations of disease in groups of easily identified people such as workers at a particular factory or attendees at a wedding.

Case-control study: In a case-control study, investigators start by enrolling a group of people with disease (at CDC such persons are called case-patients rather than cases, because case refers to occurrence of disease, not a person). As a comparison group, the investigator then enrolls a group of people without disease (controls). Investigators then compare previous exposures between the two groups. The control group provides an estimate of the baseline or expected amount of exposure in that population. If the amount of exposure among the case group is substantially higher than the amount you would expect based on the control group, then illness is said to be associated with that exposure. The study of hepatitis A traced to green onions, described above, is an example of a case-control study. The key in a case-control study is to identify an appropriate control group, comparable to the case group in most respects, in order to provide a reasonable estimate of the baseline or expected exposure.

Cross-sectional study: In this third type of observational study, a sample of persons from a population is enrolled and their exposures and health outcomes are measured simultaneously. The cross-sectional study tends to assess the presence (prevalence) of the health outcome at that point of time without regard to duration. For example, in a cross-sectional study of diabetes, some of the enrollees with diabetes may have lived with their diabetes for many years, while others may have been recently diagnosed. A cross-sectional study is a perfectly fine tool for descriptive epidemiology purposes. Cross-sectional studies are used routinely to document the prevalence in a community of health behaviors (prevalence of smoking), health states (prevalence of vaccination against measles), and health outcomes, particularly chronic conditions (hypertension, diabetes).

In summary, the purpose of an analytic study in epidemiology is to identify and quantify the relationship between an exposure and a health outcome. The hallmark of such a study is the presence of at least two groups, one of which serves as a comparison group. In an experimental study, the investigator determines the exposure for the study subjects; in an observational study, the subjects are exposed under more natural conditions. In an observational cohort study, subjects are enrolled or grouped on the basis of their exposure, then are followed to document occurrence of disease. Differences in disease rates between the exposed and unexposed groups lead investigators to conclude that exposure is associated with disease. In an observational case-control study, subjects are enrolled according to whether they have the disease or not, then are questioned or tested to determine their prior exposure. Differences in exposure prevalence between the case and control groups allow investigators to conclude that the exposure is associated with the disease. Cross-sectional studies measure exposure and disease status at the same time, and are better suited to descriptive epidemiology than causation.

Standard measures in epidemiological studies

Incidence

Incidence in epidemiology is a measure of the probability of occurrence of a given medical condition (disease) in a population within a specified period of time. It is sometimes expressed simply as the number of new cases during some time period, it is better expressed as a proportion or a rate.

Incidence rate (also known as cumulative incidence) is the number of new cases within a specified time period divided by the size of the population initially at risk. For example, if a population initially contains 1,000 non-diseased persons and 28 develop a condition over one year of observation, the incidence proportion is 28 cases per 1,000 persons per year, i.e. 2.8% in a year.

$$\text{Incidence rate} = \frac{\text{Number of new cases during a given time period}}{\text{Total number of people in the population}} \times 100 \%$$

Prevalence

It is the proportion of a particular population found to be affected by a disease or a risk factor. It is calculated with the number of people found to have the condition to the total number of people studied, and is usually expressed as a fraction, as a percentage, or as the number of cases per 1000 people. Point prevalence is the proportion of a population that has the condition at a specific point in time. Period prevalence is the proportion of a population that has the condition during a given period (e.g., 12 month prevalence), and

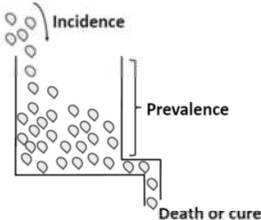
includes people who already have the condition at the start of the study period as well as those who acquire it during that period.

For example, consider a disease that takes a long time to cure and is widespread in 2002 but dissipated in 2003. This disease will have both high incidence and high prevalence in 2002, but in 2003 it will have a low incidence yet will continue to have a high prevalence (because it takes a long time to cure, so the fraction of individuals that are affected remains high). In contrast, a disease that has a short duration may have a low prevalence and a high incidence.

$$\text{Prevalence rate} = \frac{\text{Number of people with the disease (Existing+New)}}{\text{Total number of people in the population}} \times 100 \%$$

$$\text{Mortality rate} = \frac{\text{Number of deaths in a period}}{\text{Total number of people in the population}} \times 100 \%$$

Measures of disease frequency

<p>Measures of disease frequency</p> <ul style="list-style-type: none"> Measures of disease frequency in mathematical quantity <ul style="list-style-type: none"> Count Fraction <ul style="list-style-type: none"> Rate Ratio Proportion (percentage) Measures of disease frequency in epidemiology <ul style="list-style-type: none"> Prevalence Incidence 	<p>Counts</p> <ul style="list-style-type: none"> Simplest & most basic measure – absolute number of persons who have disease or characteristic of interest. Useful for health planners & administrators: for allocation of resources (e.g. quantity of ORS needed by diarrheal cases) Count of No. cases of a disease, is used for surveillance of infectious disease for early detection of outbreaks.
<p>Measurement fractions</p> <ul style="list-style-type: none"> Rate <ul style="list-style-type: none"> Measures the frequency of an event in a population. Time and multiplier Incidence Ratio <ul style="list-style-type: none"> A value obtained by dividing one number by another (either related or unrelated) Fraction that numerator is not a part of denominator Proportion <ul style="list-style-type: none"> Numerator and denominator have the same units (dimensionless). Prevalence 	<p>Prevalence VS. Incidence</p>  <ul style="list-style-type: none"> Prevalence can be viewed as describing a pool of disease in a population. Incidence describes the input flow of new cases into the pool. Fatality and recovery reflects the output flow from the pool.

Measurement of association

Expression	Question	Definition
Absolute risk	What is the incidence of disease in a group initially free of the condition?	$I = \frac{\# \text{ new case}}{\# \text{ People in group}}$
Attributable risk Risk difference	What is the incidence of disease attributable to exposure?	$AR = I_{E+} - I_{E-}$
Relative risk Risk ratio	How many times more likely are exposed persons to become disease, relative to nonexposed persons?	$RR = \frac{I_{E+}}{I_{E-}}$

		Outcome CA lung		
		Yes	No	
Risk	smoking	a	b	a+b
	non-smoking	c	d	c+d
Risks of CA in smoking (I_{E+})				= a/a+b
Risks of CA in non-smoking (I_{E-})				= c/c+d
Relative risk (risk ratio) (I_{E+}) / (I_{E-})				= $\frac{a/a+b}{c/c+d}$
Absolute risk reduction (ARR)				= (I_{E+}) - (I_{E-})
Number needed to treat (NNT)				= $\frac{1}{ARR}$

		Outcome CA lung		
		Yes	No	
Risk	smoking	a	b	
	non-smoking	c	d	
Odds of smoking in CA				= a/c
Odds of smoking in non-CA				= b/d
Odds Ratio				= $\frac{a/c}{b/d} = \frac{ad}{cb}$

Study designs in epidemiology studies. Intervention studies, Controlled clinical trials.
(Notes from Unit 2: pages 4 to 7 & notes from this unit: pages 1 to 3)

Errors in Epidemiological studies

- All epidemiological studies are mostly attempting to establish the presence or absence of a causal relationship, and the results are an estimate of the actual effect or degree of association.
- All studies are subject to error, which can obscure or minimize the truth- the size and nature of a causal relationship.
- Understanding common errors and the means to reduce them improves the precision of estimates.

There are two basic types of error in epidemiological studies:

Random error and **Systematic error**.

1. Random Error

- Random error or chance refers to the fluctuations around a true value.
- The effect of random error may result in either an underestimation or overestimation of the true value.
- It can produce type 1 or type 2 errors.
- **Type 1:** observing a difference when in truth there is none.
- **Type 2:** failing to observe a difference when there is one.

(Refer notes on Type I and Type II Errors from Unit 5)

- Random error occurs because of biologic variation, sampling error, and measurement error

Sources of Random Error

- i) Biologic Variation:** It refers to the fluctuation in biological processes in the same individual over time.
- ii) Sampling Error:** The part of the total estimation error caused by random influences on who or what is selected for the study.
- iii) Measurement Error:** The error resulting from random fluctuations in measurement.

2. Systematic Error

The systematic error refers to any difference between the true value and the actual value obtained in the study that is not the result of random error.

Systematic error, or bias, is more problematic, as it can significantly affect the validity of a study.

Sources of Systematic Error

- i) **Selection bias** can result when the selection of subjects into a study or their likelihood of being retained in the study leads to a result that is different from what you would have gotten if you had enrolled the entire target population.
- ii) **Information bias** results from systematic differences in the way data on exposure or outcome are obtained from the various study groups.
- iii) **Observation bias** may be a result of the investigator's prior knowledge of the hypothesis under investigation or knowledge of an individual's exposure or disease status.

Validity

Validity is used in epidemiology to assess the degree to which the information collected accurately answers the research question; i.e., the extent to which the results are accurate and the extent to which the conclusions derived can be generalized.

The term validity in epidemiological research is used in three different ways:

1. Internal Validity (Study Validity)
2. External Validity (Generalizability)
3. Measurement validity (Variable)

1. Internal Validity is the degree to which the observed findings lead to correct inferences about phenomena taking place in the study sample. A study is not valid if it cannot provide accurate information, or cannot enable inferences to be drawn from the population studied. For example, a study shows a higher risk of lung cancer among coffee drinkers. This increased risk attributed to coffee drinking is incorrect as coffee drinkers are more likely to smoke and therefore, show a higher risk of lung cancer.

2. External Validity is the degree to which the inferences drawn from a study can be generalized to a broader population beyond the study population. Example: A number of studies on white males in developed countries show that current smoking status increases the risk of fatal coronary heart disease. It remains to be judged whether these findings in white males can be generalized to other populations such as the males in Pakistan.

3. Measurement Validity is the degree to which a test actually measures what it is designed to measure. The process involves comparison with a technique known to be accurate (the gold standard). The validity of a measurement of body weight, for example, can be checked by calibrating the scale with standard weights. Similarly, a laboratory test must be appraised to establish validity, which is measured by the **sensitivity and specificity**

of the test. **Sensitivity** is the ability of the test to detect correctly those individuals who have a disease. While **specificity** is the ability to detect correctly those individuals who do not have the disease.

Reliability

Reliability refers to the degree to which a measurement procedure can be reproduced. Lack of reliability may arise from differences between observation or instruments of measurements or instability of the attribute being measured. A beam scale can measure body weight with great precision (that is with great reliability), on the other hand a questionnaire designed to measure quality of life is more likely to produce values that vary from one occasion to the next.

We can improve reliability by standardizing the measurement methods for example all study protocols should include operational definitions and precise instructions for recording measurements. Training and supervision of observers improves consistency of measurement techniques. Instruments can be designed to reduce variability. Variation in the way human observers make measurements can be eliminated with automatic mechanical devices and self-administered questionnaires.

Bias

Bias refers to systematic errors in any type of epidemiologic study that result in an incorrect estimate of the association between exposures and outcomes. Investigators can introduce bias into a study as a result of the procedures for identifying and enrolling subjects or from the procedures for collecting or analyzing information. Bias can also be introduced by errors in classification of outcomes or exposures. The two major types of bias are:

- a) Selection Bias
- b) Information Bias

(Refer notes on Systematic error in page 6 & 7)

Confounding

A variable that correlates both with the exposure variable and the outcome. Confounders may lead to establishment of associations between variables which do not exist. Mechanisms to prevent the effect of confounding variables include careful selection of subject populations and post- hoc multi-variable regression analysis to control for confounders.

For example, an association between low BMI and mortality may be confounded by smoking, which prevents weight gain, but also increases cancer risk. Confounders may increase or decrease the strength of an association between intervention and outcomes. Confounders can be controlled for in statistical analysis as long as they are recognized.

Short note on Biostatistics

Statistics is the numerically stated facts like the population of a country. It can also be referred to as the science dealing with data collection, tabulation and analysis and interpretation of data. In data analysis it can be classified into descriptive statistics and inferential statistics.

Descriptive Statistics – are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. It includes frequency counts, percentages, ranges (min-max), mean, median, mode, SD etc..

Inferential Statistics – are used to draw conclusions about a population by examining the sample. Accuracy of inference depends on representativeness of sample from population. Inferential statistics help researchers to test hypotheses and answer research questions, and derive meaning from the results. Researchers set the significance level for each statistical test they conduct and by using probability theory as a basis for their tests, researchers can assess how likely it is that the difference they find is real and not due to chance (*p-value*).

Biostatistics

Statistical methods applied in medicine, biology and public health are termed as **biostatistics**. **Biostatistics** is the term used when tools of statistics are applied to data that is derived from biological sciences such as medicine. It may be stated that the application of statistical methods to the solution of biological problems. Biostatistics is known by many names such as medical statistics, health statistics and vital statistics.

Medical statistics : Statistics related to clinical and laboratory parameters, their relationship, efficacy of drug, diagnostic analysis etc.

Health statistics : Statistics related to health of people in a community, epidemiology of diseases, association of occurrence of various diseases with socioeconomic and demographic variables, control and prevention of diseases etc.

Vital statistics : Statistics related to vital events in life such as of birth, death, marriages, morbidity etc. These terms are overlapping and not exclusive of each other.

Uses of Biostatistics

Statistical methods are widely used in almost all fields. Most of the basic as well as advanced statistical methods are applied in fields such as medicine, biology, public health etc.

Statistical methods are useful in planning and conducting meaningful and valid research studies on medical, health and biological problems in the population for the prevention of diseases, for finding effective appropriate treatment modalities etc. Statistical

methods needed in general are,

- > Collection of medical and health data scientifically
- > Summarizing the collected data to make it comprehensible
- > Generalizing the result from the sample to the entire population with scientific validity
- > Drawing conclusions from the summarized data and generalized results.

Example :

- > To determine the normal limits of various laboratory and clinical parameters such as BP, pulse rate, Cholesterol level, Blood sugar level etc.
- > To find difference between means and proportions of two different groups or places or periods.
- > To find correlation between variables such as cholesterol and BMI, exercise and obesity etc..
- > To find action of a drug or to compare between two drugs.
- > To find relative potency of a new drug with respect to a standard drug.
- > To find efficacy of a line of treatment or to compare between efficacies of two different line of treatments.
- > In community medicine and public health to compare the prevalence of deaths among vaccinated and unvaccinated in a community etc..

Descriptive statistics

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire population or a sample of a population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread). Measures of central tendency include the mean, median, and mode, while measures of variability include standard deviation, variance, minimum and maximum variables.

(Measures of central tendency & Measures of dispersion: Refer class notes)

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. In statistics, a central tendency is a central or typical value for a probability distribution. It may also be called a centre or location of the distribution. These values have the property that most of the observations in the data set accumulate around these values. The common measures of central tendency are the arithmetic mean, median and mode.

Arithmetic mean: It is the average of all the given observations

Arithmetic mean, $\bar{X} = \sum X_i / n$

Given scores: 6, 4, 5, 8, 2. Find the arithmetic mean.

$$\sum X_i = 6 + 4 + 5 + 8 + 2 = 25$$

$$\text{Arithmetic mean} = \sum X_i / n = 25/5 = 5$$

Median: It is the middlemost observation. ie, the observation in the $(n+1)/2^{\text{th}}$ position.

X_i : 6, 4, 5, 8, 2

Arrange: 2, 4, 5, 6, 8

Median = Middle observation = 5

Mode: It is the most frequent observation. ie, the observation which is repeated maximum number of times.

X_i : 6, 4, 5, 6, 2

Mode = Most repeated observation = 6

Measures of Dispersion / Measures of Variation

(also called variability, scatter, or spread)

It is the numeric value which gives the amount of variation in a distribution. That is, how the observations are spread among themselves. A measure of dispersion indicates the scattering of data. Measures of dispersion describe the spread of data around a central value. A high value indicates high variation and a low value indicates low variation. Zero indicates no variation, ie, all observations are same.

There are four measures of variation

(i) Range

(ii) Mean Deviation/Average Deviation

(iii) Standard Deviation (SD)

(iv) Quartile Deviation (QD)

(i) Range: It is the simplest method of measurement of dispersion and defines the difference between the largest and the smallest item in a given distribution.

Given the scores: 6, 4, 5, 8, 2. Find the range.

Smallest observation (minimum) = 2

Largest observation (maximum) = 8

$$\text{Range} = 8 - 2 = 6$$

(ii) Mean Deviation: It is the average of absolute values (positive values) of deviations from the arithmetic mean.

(Deviation: It is the difference of each observation from arithmetic mean).

For finding the mean deviation, first we have to find the mean \bar{x} .

Then find the deviation of each observation from the arithmetic mean. ie, difference of each observation from arithmetic mean.

Take the absolute values (positive) of the deviations $|x_i - \bar{x}|$.

Find the average of the deviations,

$$\text{Mean Deviation} = \frac{\sum |x_i - \bar{x}|}{n}$$

(iii) Standard Deviation: It is the square root of average of squares of deviations from the arithmetic mean.

(Deviation: It is the difference of each observation from arithmetic mean).

$$\text{Standard Deviation (SD)} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

For finding the standard deviation, first we have to find the mean \bar{x} .

Then find the deviation of each observation from the arithmetic mean. ie, difference of each observation from arithmetic mean.

Take the squares of the deviations $(x_i - \bar{x})^2$.

Find the average of the squares of deviations: $\frac{\sum (x_i - \bar{x})^2}{n}$.

Finding the square root of the above result will give the Standard Deviation.

Example: Given scores: 6, 4, 5, 8, 2. Find the Standard deviation.

$$\sum X_i = 6 + 4 + 5 + 8 + 2 = 25$$

$$\text{Arithmetic mean} = \frac{\sum X_i}{n} = \frac{25}{5} = 5$$

$$\frac{\sum (x_i - \bar{x})^2}{n} = \frac{1^2 + (-1)^2 + 0^2 + 3^2 + (-3)^2}{5} = 4$$

$$\begin{aligned} \text{Standard deviation} &= \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \\ &= \sqrt{4} = 2 \end{aligned}$$

(iv) Quartile Deviation: It is the deviation of the quartiles. Quartiles divide a distribution into four. There are three quartiles. The observation in $1/4^{\text{th}}$ position is the first quartile Q_1 , observation in $1/2^{\text{th}}$ position is the second quartile Q_2 and observation in $3/4^{\text{th}}$ position is the third quartile Q_3 .

$$\text{Quartile deviation} = \frac{(Q_3 - Q_1)}{2}$$

Inferential statistics

With inferential statistics you take that sample data from a small number of people and try to determine if the data can predict whether the drug will work for everyone (i.e. the population).

There are two main areas of inferential statistics:

Estimating parameters: This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).

Hypothesis tests: This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

(Write about research hypothesis, null hypothesis, Tests of significance – parametric & non-parametric)

Null Hypothesis:

A null hypothesis is a type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations. The null hypothesis attempts to show that no variation exists between variables or that a single variable is no different than its mean. It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis.

The null hypothesis, H_0 is the commonly accepted fact; it is the opposite of the alternate hypothesis. Researchers work to reject, nullify or disprove the null hypothesis. Researchers come up with an alternate hypothesis, one that they think explains a phenomenon, and then work to reject the null hypothesis.

Sampling Fundamentals:

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

Population:

A research population is generally a large collection of individuals or objects that is the main focus of interest of the researcher. It is for the benefit of the population that researches are done. Due to the large sizes of populations, researchers often cannot test every individual in the population because it is too expensive and time-consuming. This is the reason why researchers rely on samples.

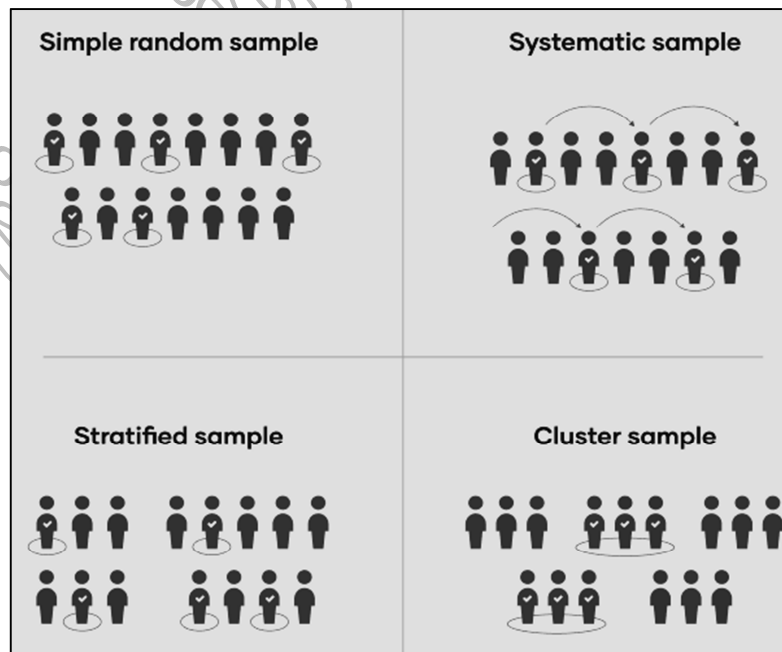
Sample:

A sample is simply a subset of the population. The sample must be representative of the population from which it was drawn and it must have good size for further statistical analysis. The main function of the sample is to allow the researchers to conduct the study to individuals from the population so that the results of their study can be used to derive conclusions that will apply to the entire population. The population “gives” the sample, and then it “takes” conclusions from the results obtained from the sample.

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group. Probability sampling means that every member of the population has a chance of being selected. It is mainly used in quantitative research. If you want to produce results that are representative of the whole population, probability sampling techniques are the most valid choice.

Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data. This type of sample is easier and cheaper to access, but it has a higher risk of sampling bias. That means the inferences you can make about the population are weaker than with probability samples, and your conclusions may be more limited. If you use a non-probability sample, you should still aim to make it as representative of the population as possible.

Probability sampling methods:



1. Simple random sampling

In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.

2. Systematic sampling

Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.

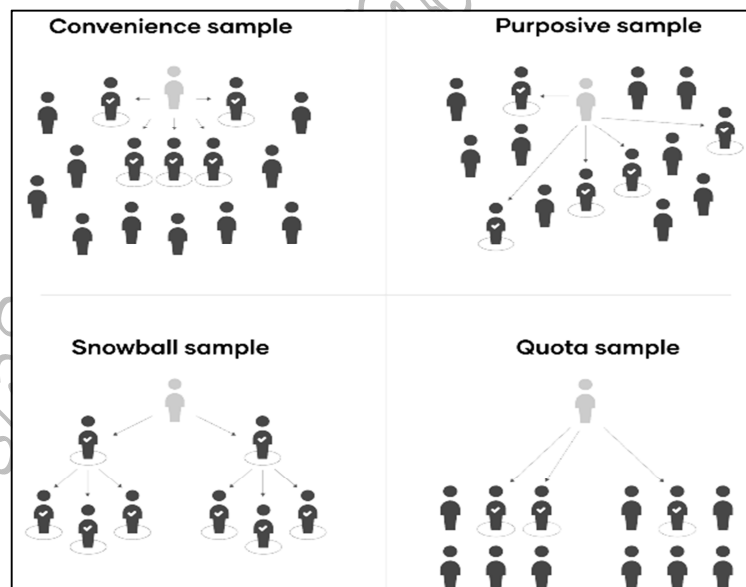
3. Stratified sampling

Stratified sampling involves dividing the population into subpopulations that may differ in important ways. It allows you draw more precise conclusions by ensuring that every subgroup is properly represented in the sample.

4. Cluster sampling

Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.

Non-Probability sampling methods:



1. Convenience sampling

A convenience sample simply includes the individuals who happen to be most accessible to the researcher. This is an easy and inexpensive way to gather initial data, but there is no way to tell if the sample is representative of the population, so it can't produce generalizable results.

2. Purposive sampling

This type of sampling, also known as judgment sampling, involves the researcher using their expertise to select a sample that is most useful to the purposes of the research.

3. Snowball sampling

If the population is hard to access, snowball sampling can be used to recruit participants via other participants. The number of people you have access to “snowballs” as you get in contact with more people.

4. Quota sampling

Quota sampling is defined as a non-probability sampling method in which researchers create a sample involving individuals that represent a population. Researchers choose these individuals according to specific traits or qualities. They decide and create quotas so that the market research samples can be useful in collecting data. These samples can be generalized to the entire population. The final subset will be decided only according to the interviewer’s or researcher’s knowledge of the population.

Factors involved in Sample size calculation:

- Type I error
- Type II error
- Effect size
- Standard deviation of population

(i) Type I error, α - error (Level of significance): cut-off level at which we say a p-value is significant. Probability of concluding that there is a statistically significant difference. Typically 5%.

(ii) Type II error (β - error) and Power ($1 - \beta$): Power is the ability of a statistical test to show if a significant difference truly exist symbolized as $1 - \beta$. In hypothesis testing, it is important to have a sizable sample to allow statistical tests to show significant differences where they exist. Typically 80%, 90%.

(iii) Effect Size: It is the difference the researcher expects to see. What has been seen previously from reviews. What is a clinically important difference?

Standard deviation of population: It is the standard deviation of the outcome variable, in most of the cases obtained from previous studies.

Estimating the sample size for a descriptive study based on a proportion

To calculate the sample size based on the sample required to estimate a proportion, the following formula is used:

$$n \geq \frac{(z)^2 pq}{N^2}$$

n is the required sample size, z is the normal distribution value corresponds to 95% limits (1.96) or 99% limits (2.58), p = proportion of population having that

characteristic, which can be known from previous studies or other sources, $q = 1 - p$ (or $100 - p$ if p and q are expressed in percentages), m is the allowable error.

Example: A study on anemic children in schools. Proportion of anemic children in a similar study is found to be 30%. Find the minimum sample size required at a confidence limit of 95% and accepting an error of 10% of the population.

$$n \geq \frac{(1.96)^2 30 \times 70}{10^2} = 80.67 \approx 81 \text{ or more samples}$$

Criteria for inclusion and exclusion

The investigator must specify inclusion and exclusion criteria for participation in a study. Inclusion criteria are characteristics that the prospective subjects must have if they are to be included in the study. Exclusion criteria are those characteristics that disqualify prospective subjects from inclusion in the study. Inclusion and exclusion criteria may include factors such as age, gender, race, ethnicity, type and stage of disease, the subject's previous treatment history, and the presence or absence (as in the case of the "healthy" or "control" subject) of other medical, psychosocial, or emotional conditions. Healthy, or control, subjects may be defined as those individuals who are free of certain specified attributes of non-health. Additional information on screening potential subjects for attributes of non-health is available in the Specific Guidance on Special Issues section of this module. Defining inclusion and exclusion criteria increases the likelihood of producing reliable and reproducible results, minimizes the likelihood of harm to the subjects, and guards against exploitation of vulnerable persons.

An example of inclusion criteria for a study of chemotherapy of breast cancer subjects might be postmenopausal women between the ages of 45 and 75 who have been diagnosed with Stage II breast cancer. An exclusion criterion for this study might be abnormal renal function tests, if the combination of study drugs includes one or more that is nephrotoxic. In this case it would be required to specify which tests of renal function are to be performed to evaluate renal function and the threshold values that would disqualify the prospective subject.

Furthermore, the investigator must be prepared to provide a rationale in case one or more of the inclusion or exclusion criteria is questioned. The investigator should review the inclusion and exclusion criteria and decide if any group(s) is inappropriately excluded. If the justification for the exclusion of this group is not reasonable with regard to the risks, benefits, and purpose of the research, then this group should be included. In the breast cancer study example discussed above, there would be no justification for the exclusion of minority women.

Dropouts

Dropout in the medical research refers to a state in which observation is suspended or lost because a study participant cannot or does not attend the scheduled visits required by the research plan. When data are collected repeatedly over a period of time, a proportion of the data will be lost if participants drop out of the study. Dropping out is more common in subjects receiving interventions with potentially negative effects and might lead to incorrect estimation of the true effects of an intervention. Loss of study data due to dropouts potentially introduces a risk of bias, so reducing the dropout rate is essential for a successful clinical trial.

Dropout in randomised controlled trials is common and threatens the validity of results, as completers may differ from people who drop out. Differing dropout rates between treatment arms is sometimes called differential dropout or attrition. Although differential dropout can bias results, it does not always do so.

vipinsjcp@gmail.com

Hypothesis

A hypothesis is an educated guess that's formed at the beginning of a research experiment and can be used in almost every field. A hypothesis is a potential explanation for something that happens or that you observe and think to be true. It can also be used to determine the relationship between two or more variables that you think might be related to each other. A **scientific hypothesis** must be about something that can be proved or disproved through experimentation or observation and as such require extensive research as well as controlling dependent and independent variables.

A **research hypothesis** states your predictions about what your research will find. It is a tentative answer to your research question that has not yet been tested. For some research projects, you might have to write several hypotheses that address different aspects of your research question. A **research hypothesis** is a statement of expectation or prediction that will be tested by research. Before formulating your research hypothesis, read about the topic of interest to you.

A Hypothesis is a tentative statement about the relationship between two or more variables. A hypothesis is a specific, testable prediction about what you expect to happen in your study.

To be complete, the hypothesis must include three components –

- The variables;
- The population
- The relationship between the variables.

Null Hypothesis:

A null hypothesis is a type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations. The null hypothesis attempts to show that no variation exists between variables or that a single variable is no different than its mean. It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis.

The null hypothesis, H_0 is the commonly accepted fact; it is the opposite of the alternate hypothesis. Researchers work to reject, nullify or disprove the null hypothesis. Researchers come up with an **alternate hypothesis**, one that they think explains a phenomenon, and then work to reject the null hypothesis.

Characteristics of a good hypothesis

A good Hypothesis must possess the following characteristics –

1. It is never formulated in the form of a question.
2. It should be empirically testable, whether it is right or wrong.

3. It should be specific and precise.
4. It should specify variables between which the relationship is to be established.
5. It should describe one issue only. A hypothesis can be formed either in descriptive or relational form.
6. It should not conflict with any law of nature which is known to be true. It guarantees that available tools and techniques will be effectively used for the purpose of verification.
7. It should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned.
8. It must explain the facts that gave rise to the need for explanation.
9. It should be amenable to testing within a reasonable time.
10. It should not be contradictory.

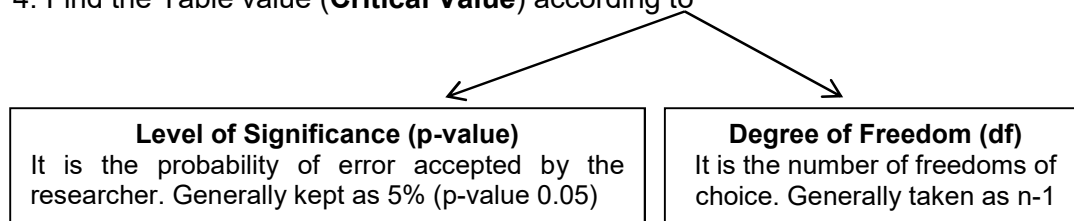
Testing of Hypothesis:

Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.

- Hypothesis testing is used to assess the plausibility (reasonability) of a hypothesis by using sample data.
- The test provides evidence concerning the plausibility of the hypothesis, given the data.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

Procedure/Steps of Hypothesis Testing:

1. Set the **Null Hypothesis H_0**
2. Choose appropriate Statistical Test according to Study Design.
3. Carry out the statistical test with the formula and find the calculated value (**Test Statistic**) - The value depends on data collected from samples
4. Find the Table value (**Critical Value**) according to



5. Compare the calculated value with the table value:

If calculated value (test statistic) > table value (critical value):

Reject **Null Hypothesis H_0**

If calculated value (test statistic) < table value (critical value):

Accept **Null Hypothesis H_0**

Hypothesis Testing / Tests of Significance

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favour or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance.

Every test of significance begins with a **Null Hypothesis H_0** . H_0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the **null hypothesis** might be that the new drug is not better than the current drug. We would write H_0 : there is no significant difference between the two drugs.

The **Research Hypothesis/alternative hypothesis, H_a** , is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write **H_a** : there is significant difference between the two drugs. The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write H_a : the new drug is better than the current drug.

The final conclusion once the test has been carried out is always given in terms of the **null hypothesis**. We either "**reject H_0 in favour of H_a** " or "**do not reject H_0** ".

Level of Significance (p-value):

The null hypothesis is rejected if the p-value is less than a predetermined level, α . α is called the significance level, and is the probability of rejecting the null hypothesis given that it is true (a type I error). It is usually set at or below 5%. Statistical significance plays an important role in statistical hypothesis testing. It is used to determine whether the null hypothesis should be rejected or retained. The null hypothesis is the default assumption that nothing happened or changed. For the null hypothesis to be rejected, an observed result has to be statistically significant, i.e. the observed **p-value** is less than the pre-specified significance level.

To determine whether a result is statistically significant, a researcher calculates a p-value, which is the probability of observing an effect of the same magnitude or more extreme given that the null hypothesis is true. The null hypothesis is rejected if the p-value is less than a predetermined level, α . α is called the significance level, and is the probability of rejecting the null hypothesis given that it is true (a type I error). It is usually set at or below 5%.

Tests of Significance

Parametric tests	Non-parametric tests
1. Tests involving comparison of parameters like mean, SD etc.	1. No comparison of parameters like mean, SD etc.
2. Conditions of normality and homogeneity are to be satisfied.	2. No conditions of normality and homogeneity are to be satisfied.
3. Relies on statistical distributions	3. The procedures are distribution-free
4. Quantitative approach	4. Qualitative approach
5. Examples: t-test, z-test, ANOVA	5. Examples: Mann Whitney U-test, Wilcoxon test

Parametric tests

Parametric tests are those that make assumptions about the parameters of the population distribution from which the sample is drawn.

Student's t-test

It is a parametric test used to analyse the significant difference between two means. The null hypothesis states that there is no significant difference between the means. We have two types of t-test:

1. Paired t-test
2. Unpaired/Independent t-test

1. Paired t-test:

It is used to analyse the significant difference between two paired observations (pre-test and post-test or the observations on the same set of samples taken at two different time periods). In a paired sample t-test, each subject or entity is measured twice, resulting in pairs of observations.

Procedure:

1. Set the Null Hypothesis H_0 (There is no significant difference between pre-test and post-test means)
2. Find the difference between each paired observation taken as d_i .

Pre-test	Post-test	Pre-post difference(d_i)	d_i^2
X_{11}	X_{12}	d_1	d_1^2
X_{21}	X_{22}	d_2	d_2^2
....
....

- Find the mean of differences \bar{d} .
- Find the square of differences d_i^2
- Find the standard deviation of differences, $S.D. = \sqrt{\frac{\sum d_i^2 - n(\bar{d})^2}{n-1}}$
- Find the **Standard Error (S.E.)** = SD/\sqrt{n}
- Find the calculated test statistic, $t = \bar{d} / SE$
- Find the table value/critical value according to level of significance 5% (0.05) and degree of freedom ($n - 1$).

Conclusion

Compare the calculated value with the table value:

If calculated value (test statistic) > table value (critical value):

Reject Null Hypothesis H_0

If calculated value (test statistic) < table value (critical value):

Accept Null Hypothesis H_0

Example:

Given the following BP before and after administering an anti-hypertensive. Test whether the drug is effective in reducing BP.

Pre-test	Post-test	Pre-post difference(d_i)	d_i^2
136	130	6	36
140	136	4	16
130	125	5	25
128	121	7	49
132	129	3	9
		$\bar{d} = 25/5 = 5$	$\sum d_i^2 = 135$

Null Hypothesis H_0 : The drug is not effective in reducing BP.

Mean difference, $\bar{d} = 5$

$$\begin{aligned}\text{Standard deviation of differences, S.D.} &= \sqrt{\frac{\sum d_i^2 - n(\bar{d})^2}{n-1}} \\ &= \sqrt{\frac{135 - 5 \times 25}{4}} \\ &= 1.58\end{aligned}$$

$$\text{Standard Error (S.E.)} = SD/\sqrt{n} = 1.58/\sqrt{5} = \mathbf{0.707}$$

$$\text{Test statistic, } t = \bar{d} / SE = 5/0.707 = \mathbf{7.071}$$

Find the table value/critical value according to level of significance 5% (0.05) and degree of freedom ($5 - 1 = 4$). Critical value = **2.78**

Conclusion:

The test statistic (7.071) is greater than the critical value (2.78) at 5% level of significance. Hence the null hypothesis is rejected. So the drug is effective in reducing BP.

2. Unpaired/independent t-test:

It is used to analyse the significant difference between two different groups (control-experimental or comparing effect of two types of drugs). There will be two set of observations (observations of samples on two levels of the independent variable).

Procedure:

1. Set the Null Hypothesis H_0 (There is no significant difference between the groups)
2. Find the mean in two groups \bar{x}_1 and \bar{x}_2 .
3. Find the deviations and the square of deviations in both groups as given below:

Group A(x_1)	Group B(x_2)	$(x_{1i} - \bar{x}_1)$	$(x_{1i} - \bar{x}_1)^2$	$(x_{2i} - \bar{x}_2)$	$(x_{2i} - \bar{x}_2)^2$
x_{11}	x_{21}
x_{12}	x_{22}
....
....

$$4. \text{ Find the pooled standard deviation, S.D.} = \sqrt{\frac{\sum (x_{1i} - \bar{x}_1)^2 + \sum (x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

$$5. \text{ Find the Standard Error (S.E.)} = SD / \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

6. Find the calculated test statistic, $t = |\bar{x}_1 - \bar{x}_2| / SE$

7. Find the table value/critical value according to level of significance 5% (0.05) and degree of freedom ($n_1 + n_2 - 2$).

Conclusion

Compare the calculated value with the table value:

If calculated value (test statistic) > table value (critical value):

Reject Null Hypothesis H_0

If calculated value (test statistic) < table value (critical value):

Accept Null Hypothesis H_0

Example:

Given the following are the weight gain in children on usual diet and children on usual diet + vitamins. Test whether there is any significant differences in the weight gain among children in two groups.

Usual diet (x_1)	Usual diet + Vitamins (x_2)	$(x_{1i} - \bar{x}_1)$	$(x_{1i} - \bar{x}_1)^2$	$(x_{2i} - \bar{x}_2)$	$(x_{2i} - \bar{x}_2)^2$
1	4	-1	1	-1	1
2	6	0	0	1	1
1	3	-1	1	-2	4
4	7	2	4	2	4
2	5	0	0	0	0

Null Hypothesis H_0 : There is no significant difference in the weight gain among children in two groups.

Mean in two groups $\bar{x}_1 = 10/5 = 2$ and $\bar{x}_2 = 25/5 = 5$

$$\begin{aligned} \text{Pooled standard deviation, S.D.} &= \sqrt{\frac{\sum (x_{1i} - \bar{x}_1)^2 + \sum (x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}} \\ &= \sqrt{\frac{6 + 10}{5 + 5 - 2}} = \sqrt{2} = 1.41 \end{aligned}$$

$$\text{Standard Error (S.E.)} = SD / \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 1.41 / \sqrt{2.5} = 1.41 / 1.58 = 0.89$$

$$\text{Test statistic, } t = |\bar{x}_1 - \bar{x}_2| / SE = |2 - 5| / 0.89 = 3 / 0.89 = \underline{\underline{3.36}}$$

Find the table value/critical value according to level of significance 5% (0.05) and degree of freedom $n_1 + n_2 - 2 = 5 + 5 - 2 = 8$. $df = 8$.

Critical value = **2.31**

Conclusion:

The test statistic (3.36) is greater than the critical value (2.31) at 5% level of significance. Hence the null hypothesis is rejected. So there is a significant difference in the weight gain among children in two groups. Children supplemented with vitamins along with usual diet had a significant high weight gain.

ANOVA (ANalysis Of VAriance)

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures such as the "variation" among and between groups used to analyze the differences among more than two group means in a sample. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.

One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test. One-way has one independent variable (with 2 levels) and two-way has two independent variables (can have multiple levels). For example, a one-way Analysis of Variance could have one independent variable (brand of cereal) and a two-way Analysis of Variance has two independent variables (brand of cereal, calories).

ANOVA – key features

- It is a parametric test.
- It can be used in comparing means of more than two groups of samples.
- The null hypothesis states that there are no differences among the population means.
- It is assumed that samples are drawn from population following Normal Distribution
- The variances of different samples are relatively equal
- The variance between the samples is calculated.
- The variance within samples is calculated.
- In ANOVA we use the F-distribution.

- The test statistic F is the ratio of variance between samples to the variance within samples. $F = \frac{\text{Variance between samples (groups)}}{\text{Variance within samples (groups)}}$

Procedure:

1. Set the Null Hypothesis H_0 (There is no significant difference between the groups)
2. Find **T** = Sum of all items in various samples.
3. Find the Correction Factor (**C.F.**) = T^2/n , n is the total number of samples.
4. Find Total Sum of Squares (**TSS**) = $\sum X_i^2 - CF$
5. Sum of Squares between Samples (**SSS**) = $\sum \left[\frac{(\sum X_i)^2}{N} \right] - CF$,
N is the number of samples within the corresponding group.
6. Sum of Squares within Samples = **TSS – SSS**
7. There are two degrees of freedom, **df₁** = no. of groups – 1 and **df₂** = n – no. of groups
8. Prepare the table :

Variance	SS	d.f.	SS/d.f.
b/w groups	SSS	df ₁ = no. of groups – 1	SSS/ df ₁
within groups	TSS – SSS	df ₂ = n – no. of groups	(TSS – SSS)/ df ₂

Variance between groups = SSS/ df₁

Variance within groups = (TSS – SSS)/ df₂

9. The test statistic is given by $F = \frac{\text{Variance between samples (groups)}}{\text{Variance within samples (groups)}}$
- $$= \frac{\text{SSS/ df}_1}{(\text{TSS} - \text{SSS}) / \text{df}_2}$$

10. Find the table value/critical value according to level of significance 5% (0.05) and degree of freedoms df₁ and df₂.

Conclusion

Compare the calculated value with the table value:

If calculated value (test statistic) > table value (critical value):

Reject Null Hypothesis H_0

If calculated value (test statistic) < table value (critical value):

Accept Null Hypothesis H_0

Example:

Test whether there is a significant difference between the three drugs.

Drug A (x_1)	Drug B (x_2)	Drug C (x_3)	x_1^2	x_2^2	x_3^2
2	8	5	4	64	25
1	12	6	1	144	36
2	10	4	4	100	16
2	11	5	4	121	25
3	9	5	9	81	25
$\Sigma x_1 = 10$	$\Sigma x_2 = 50$	$\Sigma x_3 = 25$	$\Sigma x_1^2 = 22$	$\Sigma x_2^2 = 510$	$\Sigma x_3^2 = 127$

Null Hypothesis H_0 : There is no significant difference between the groups.

T = Sum of all items in various samples = $\Sigma x_1 + \Sigma x_2 + \Sigma x_3 = 10 + 50 + 25 = 85$

Correction Factor (**C.F.**) = $T^2/n = 85^2/15 = 481.67$

Total Sum of Squares (**TSS**) = $\Sigma X_i^2 - CF = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 - CF = 22 + 510 + 127 - 481.67$
 $= 659 - 481.67 = 177.33$

Sum of Squares between Samples (**SSS**) = $\Sigma \left[\frac{(\Sigma X_i)^2}{N} \right] - CF$, (N = 5 in each group)
 $= \left[\frac{10^2}{5} + \frac{50^2}{5} + \frac{25^2}{5} \right] - 481.67$
 $= 645 - 481.67$
 $= 163.33$

Sum of Squares within Samples = **TSS – SSS** = $177.33 - 163.33 = 14$

Degrees of freedom,

df₁ = no. of groups – 1 = $3 - 1 = 2$

df₂ = n – no. of groups = $15 - 3 = 12$

Variance	SS	d.f.	SS/d.f.
b/w groups	SSS = 163.33	df ₁ = 2	SSS/ df ₁ = $163.33/2 = 81.67$
within groups	TSS – SSS = 14	df ₂ = 12	(TSS – SSS)/ df ₂ = $14/2 = 1.17$

Variance between groups = $SSS / df_1 = 81.67$

Variance within groups = $(TSS - SSS) / df_2 = 1.17$

The test statistic is given by $F = \frac{SSS / df_1}{(TSS - SSS) / df_2} = \frac{81.67}{1.17} = \underline{\underline{70}}$

Table value/critical value according to level of significance 5% (0.05) and degree of freedoms $df_1 = 2$ and $df_2 = 12$,

Critical value = **3.88**

Conclusion:

The test statistic (70) is greater than the critical value (3.88) at 5% level of significance. Hence the null hypothesis is rejected. So there is a significant difference between the three drugs.

Non-Parametric tests

Non-Parametric tests are those do not make assumptions about the parameters of the population distribution from which the sample is drawn. They are distribution-free.

Chi-square test (χ^2)

It is a non-parametric test used to analyse the significance of association between two discrete (categorical) random variables such as smoking and lung cancer, alcohol consumption and liver cirrhosis, vaccination and disease, stress and BP etc..

The test gives a numeric value of test statistic, which gives the probability of association between the variables.

Key points

- Non-parametric test
- Association between two discrete (categorical) random variables
- Variables must have two or more categories
- The test finds a test statistic which gives the probability of association between the variables.

Procedure:

1. Prepare the contingency table (cross table consisting of frequencies in categories)
2. Set the **Null Hypothesis H_0** : There is no significant association between the variables.
3. Observed values are taken as (**O_i**) – the observed frequencies in the contingency table.
4. Find the expected values (**E_i**) corresponding to each **O_i**

$$\text{Expected value } (E_i) = \frac{\text{Corresponding row total} \times \text{column total}}{\text{Grand total}}$$

5. Find the difference between each observed (O_i) and expected (E_i) ie, ($O_i - E_i$)

Also find the square of differences ($O_i - E_i$)²

6. Divide it with corresponding expected (E_i) ie, ($O_i - E_i$)² / (E_i)

7. The test statistic is given by

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

8. Find the table value/critical value according to level of significance 5% (0.05) and degree of freedom, $df = (c - 1) \times (r - 1)$, where c is the number of columns and r number of rows.

Conclusion

Compare the calculated value with the table value:

If calculated value (test statistic) > table value (critical value):

Reject Null Hypothesis H_0

If calculated value (test statistic) < table value (critical value):

Accept Null Hypothesis H_0

Example:

Test whether there is a significant association between gender and hypertension.

Gender	Hypertension		Total
	Present	Absent	
Male	20	30	50
Female	35	15	50
Total	55	45	100

Null Hypothesis H_0 : There is no significant association between gender and hypertension.

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
20	$(50 \times 55) / 100 = 27.5$	$20 - 27.5 = -7.5$	$(-7.5)^2 = 56.25$	$56.25 / 27.5 = 2.05$
30	$(50 \times 45) / 100 = 22.5$	$30 - 22.5 = 7.5$	$(7.5)^2 = 56.25$	$56.25 / 22.5 = 2.50$
35	$(50 \times 55) / 100 = 27.5$	$35 - 27.5 = 7.5$	$(7.5)^2 = 56.25$	$56.25 / 27.5 = 2.05$
15	$(50 \times 45) / 100 = 22.5$	$15 - 22.5 = -7.5$	$(-7.5)^2 = 56.25$	$56.25 / 22.5 = 2.50$
				$\Sigma(O_i - E_i)^2 / E_i = 9.1$

Test statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \underline{\underline{9.1}}$$

Table value/critical value according to level of significance 5% (0.05) and degree of freedom $df = (2 - 1) \times (2 - 1) = 1$,

Critical value = **3.84**

Conclusion:

The test statistic (9.1) is greater than the critical value (3.84) at 5% level of significance. Hence the null hypothesis is rejected. So there is a significant association between gender and hypertension.

Mann Whitney U test

It is a non-parametric test used to analyse the significant difference between two groups similar to that of independent t-test. There is an independent variable having two discrete (categories) levels and there is a continuous dependent variable.

Procedure:

1. The null hypothesis states that there is no difference between two groups.
2. Data are ranked in ascending order considering the data in both the groups.
3. The sums of the ranks of the dependent variables is calculated for one level of the independent variable (sum of ranks in only one group is considered)

Group A(x_1)	Rank	Group B(x_2)	Rank
x_{11}	R_{11}	x_{21}	R_{21}
x_{12}	R_{12}	x_{22}	R_{22}
x_{13}	R_{13}	x_{23}	R_{23}
.....
.....
	ΣR_{1i}		ΣR_{2i}

4. Either the first or second ranking could be used for statistical sum of the ranks (ΣR_{1i} or ΣR_{2i}). The statistical values are calculated using the following formula:

$$5. \quad U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \Sigma R_{1i} \quad , \text{ here } \Sigma R_{1i} \text{ and } n_1 \text{ are used for calculation.}$$

n_2 is the number of observations in second group.

(ΣR_{2i} and n_2 can also be used for calculation)

6. The U value is converted to a z-value by applying a second formula:

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

7. The calculated z-value is then compared to values of the normalized z-values 1.96 (5% level of significance) or 2.58 (1% level of significance).

Conclusion

Compare the calculated value with the critical values at 5% level of significance:

If calculated z value (test statistic) > 1.96 (critical value):

Reject Null Hypothesis H_0

If calculated z value (test statistic) < 1.96 (critical value):

Accept Null Hypothesis H_0

Example:

The following data gives the weight gain in children on usual diet (Group A) and usual diet + vitamins (Group B). Test whether the children on vitamins have a significant weight gain.

Group A	Rank R_1	Group B	Rank R_2
6	5	12	10
3	2	9	7
2	1	8	6
4	3	10	8
5	4	11	9
	$\Sigma R_{1i} = 15$		$\Sigma R_{2i} = 40$

Null Hypothesis H_0 : There is no significant difference between groups

We can consider $\Sigma R_{1i} = 15$, for the sum of ranks.

$$\begin{aligned}
 U &= n_1 n_2 + \frac{n_1(n_1+1)}{2} - \Sigma R_{1i} \\
 &= 5 \times 5 + \frac{5(5+1)}{2} - 15 \\
 &= \underline{\underline{25}}
 \end{aligned}$$

Finding the z value from U,

$$z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

$$= \frac{25 - 5 \times 5 / 2}{\sqrt{5 \times 5 (5 + 5 + 1) / 12}} = 12.5 / \sqrt{275 / 12}$$

$$= 12.5 / 4.79$$

$$= \underline{\underline{2.61}}$$

Conclusion:

The test statistic (2.61) is greater than the critical value (1.96) at 5% level of significance. Hence the null hypothesis is rejected. So there is a significant difference between the groups.

Wilcoxon rank sum test (Matched pairs)

It is a non-parametric test used to analyse the significant difference between two paired observations similar to that of paired t-test.

Procedure:

1. The null hypothesis states that there is no difference between pre-test and post-test observations.
2. The pre-test and post-test differences of each sample is taken as d_1, d_2, \dots (put +ve sign for positive changes and -ve sign for negative changes)

Pre-test	Post-test	Pre-post difference(d_i)	Rank
X_{11}	X_{12}	d_1	R_1
X_{21}	X_{22}	d_2	R_2
....
....

3. The absolute differences (regardless of sign, positive or negative) are then ranked from smallest to largest R_1, R_2, \dots . If a difference is zero, it is not considered. The data associated with no differences are eliminated and number of samples (n) reduces appropriately.
4. A Wilcoxon T-statistic, **T-value** is calculated for the sum of the ranks associated with

the *least frequent* sign +/ – (if all the patients have +ve / –ve changes, T will be 0).

T+ = Sum of ranks of positive changes

T– = Sum of ranks of negative changes

T - value = Minimum (T+, T–)

5. Find the normal approximation z-value using T-value,

$$z = \frac{|T - \frac{n(n+1)}{4}|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

6. The calculated z-value is then compared to values of the normalized z-values 1.96 (5% level of significance) or 2.58 (1% level of significance).

Conclusion

Compare the calculated value with the critical values at 5% level of significance:

If calculated z value (test statistic) > 1.96 (critical value):

Reject Null Hypothesis H_0

If calculated z value (test statistic) < 1.96 (critical value):

Accept Null Hypothesis H_0

Example:

The following data gives the BP before and after a drug is administered. Test whether the drug is effective in reducing BP.

Pre-test	Post-test	Pre-post difference(di)	Rank
136	130	+6	4
140	132	+8	5
128	130	–2	1
135	130	+5	3
140	136	+4	2

Null Hypothesis: There is no significant effect of the drug.

T+ = Sum of ranks of positive changes = 4 + 5 + 3 + 2 = 14

T– = Sum of ranks of negative changes = 1

T - value = Minimum (T+, T–) = 1

Normal approximation for the Wilcoxon T-statistic,

$$z = \frac{|T - \frac{n(n+1)}{4}|}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{|1 - \frac{5(5+1)}{4}|}{\sqrt{\frac{5(5+1)(10+1)}{24}}} = \frac{|-6.5|}{\sqrt{\frac{330}{24}}}$$

$$= 6.5/3.71 = \underline{\underline{1.75}}$$

Conclusion:

The test statistic (1.75) is less than the critical value (1.96) at 5% level of significance. Hence the null hypothesis is accepted. So there is no significant effect of the drug on hypertension.

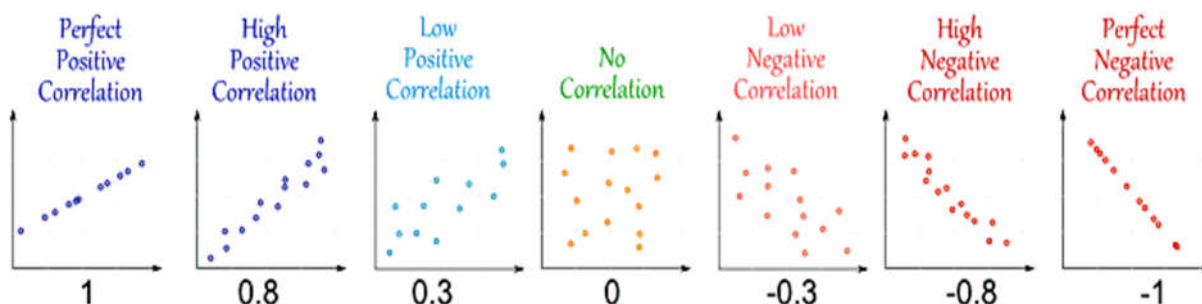
Correlation Analysis

It is a statistical measure which shows the relationship between two or more variables moving in the same direction or in opposite direction. With correlation, two or more variables may be compared to determine if there is a relationship and to measure the strength of that relationship. The correlation coefficient gives the strength of relationship between the variables.

- Correlation gives degree and direction of relationship
- Correlation does not require an independent (predictor) variable
- Correlation results do not explain why the relation occurs

The correlation may be either positive, negative or zero. The first role of correlation is to determine the strength of relationship between the two variables represented on the x-axis and y-axis. The measure of this magnitude is called the correlation co-efficient. The data required to compute this coefficient are two continuous measurements (x, y) obtained on the same entity.

If there is a perfect relationship, a straight line can be drawn through all the data points. The greater the change in y for a constant change in x, the steeper the slope of the line. In a less than perfect relationship between two variables, the closer the data points are located on a straight line, the stronger the relationship and greater the correlation coefficient. In contrast, a zero correlation would indicate absolutely no linear relationship between the two variables.



Positive Correlation

One variable increases with increase of the other or decreases with decrease of the other. Eg: temperature and filter rate, water intake and urine output, exercise and heart rate.

Negative Correlation

One variable increases with decrease of the other or decreases with increase of the other. Eg: Pressure and volume, Particle size and filter rate, insulin and blood sugar.

Zero Correlation

There is no relation between the variables.

The Coefficient of Correlation

A measure of the strength of linear relationship between two variables that is defined in terms of the covariance of the variables divided by their standard deviations.

$$\text{Correlation coefficient, } r = \frac{\text{Covariance } (x, y)}{(\text{S.D. of } x) (\text{S.D. of } y)}$$

The following formula gives the result of correlation coefficient.

$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

Regression Analysis

In regression analysis, researchers control the values of at least one of the variables and assign objects at random to different levels of these variables. Where correlation simply described the strength and direction of the relationship, regression analysis provides a method for describing the nature of the relationship between two or more continuous variables. Correlation coefficient can support the interpretation associated with regression. If a linear relationship is established, the magnitude of the effect of the independent variable can be used to predict the corresponding magnitude of the effect on the dependent variable.

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (response) and independent variable(s) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables.

Regression analysis is a statistical method to estimate or predict the values of one variable (dependent variable) for the given values of independent variable.

- > Dependent variable is to be estimated or predicted (response)
- > Independent variable is the given variable (predictor)

Example: weight of a baby depends on age.

So age is the independent variable whereas weight is dependent variable.

Uses of Regression Analysis

- Describe one variable with level of other
- Understanding association eg: birth wt. & gestation
- Identify the variable which influence a particular one
- Prediction of dependent variable for given values of independent variable
- To identify the abnormal values or outliers

Types

- Simple Linear Regression (1 response – 1 predictor)
- Multiple Regression (1 response – Many predictors)
- Logistic Regression (Any response or predictors – Nominal / Ordinal)

1. Simple Linear Regression (1 response – 1 predictor)

The dependent variable is continuous, independent variable can be continuous or discrete, and nature of regression line is linear. Linear is used to denote that the relationship between two variables can be described by a straight line. With linear regression, a relationship is established between the two variables and a response for the dependent variable can be made based on a given value for the independent variable. For example Injury Severity Score can be used to predict length of hospital stay.

2. Multiple Regression (1 response – Many predictors)

The dependent variable (response) is predicted by using several independent variables (predictors) You could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety and lecture attendance.

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.

3. Logistic Regression (Any response or predictors – Nominal / Ordinal)

This is the regression model in which the dependent variable is not continuous, ie, it is categorical. Independent variables can be continuous or discrete, and nature of regression line is linear. For example Smoking habit (Yes/No) can be used to predict COPD (Yes/No).

Applications of parametric and non-parametric tests:

(Refer Pages 1 & 2: Uses of Biostatistics, Example in Unit 4)

Interpretation of results:

(Refer Pages 2 & 3: Testing of Hypothesis, Steps of Hypothesis testing in this unit)

Type I and Type II Errors:

There are two possible errors associated with hypothesis testing. When considering a hypothesis testing, we either reject the null hypothesis or accept the null hypothesis. When considering the reality into the scenario, we have four potential outcomes:

$1 - \alpha$: Accept a null hypothesis H_0 when in fact it is to be accepted

α : Reject a null hypothesis H_0 when in fact it is to be accepted

$1 - \beta$: Reject a null hypothesis H_0 when in fact it is to be rejected

β : Accept a null hypothesis H_0 when in fact it is to be rejected

Type I error is the probability of rejecting a true null hypothesis (H_0). For example the null hypothesis is “The drug is not effective” which is true – in reality the drug is not effective, but the hypothesis testing resulted in rejecting the null hypothesis.

Type II error is the probability of accepting a false null hypothesis (H_0). For example the null hypothesis is “The drug is not effective” which is false – in reality the drug is effective, but the hypothesis testing resulted in accepting the null hypothesis.

		The Real World (All the facts known)	
		H_0 is true Person Innocent	H_0 is false Person Guilty
Statistical Test Result (Judge's Decision)	Fail to Reject H_0 Not Guilty	$1 - \alpha$	Type II Error β
	Reject H_0 Guilty	Type I Error α, ρ	$1 - \beta$

In Hypothesis testing we always want to minimize the α and maximize $1 - \beta$.

Computer software's in Bio statistical Analysis

A statistical package is a suite of computer programs that are specialised for statistical analysis. It enables people to obtain the results of standard statistical procedures and statistical significance tests, without requiring low-level numerical programming. Most statistical packages also provide facilities for data management.

(1) SPSS (Statistical Package for Social Sciences):

SPSS is a computer program used for statistical analysis. **SPSS (Statistical Package for the Social Sciences)** was released in its first version in 1968 after being

developed by Norman H. Nie and C. Hadlai Hull. It is used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others.

The many features of SPSS are accessible via pull-down menus or can be programmed with command syntax language. Additionally, some complex applications can only be programmed in syntax and are not accessible through the menu structure.

SPSS has got two views: **Data View and Variable View**. In the variable view, the variables such as age, sex, stress score, knowledge score etc... are declared. In the Data view, we enter the data collected from the samples. SPSS datasets have a 2-dimensional table structure where the rows typically represent **cases (individuals)** and the columns represent **measurements (such as age, sex or household income)**. SPSS Statistics data files are organized by **cases (rows) and variables (columns)**.

Rather than typing all of your data directly into the Data Editor, you can read data from applications such as Microsoft Excel. The Opening Excel Data Source dialog box is displayed, allowing you to specify whether variable names are to be included in the spread sheet, as well as the cells that you want to import. If you want to import only a portion of the spread sheet, specify the range of cells to be imported in the Range text box.

Statistical Analysis such as **Paired t test, Unpaired t-test, Chi-Square test, Test of normality, ANOVA** etc... can be carried out very easily in the Software. For the two group studies, we can do the separation into groups and do the analysis separately in groups.

Applications:

1. Calculation is very fast and no need of paper work in computing is required.
2. Different significance tests can be performed with one time entry of data.
3. Graphical presentation of data can also be done with the data.
4. Different comparisons and associations can be made at a time.
5. Documentation of the whole data analysis is easy.

(2) MINITAB:

Minitab is statistical analysis software. It can be used for learning about statistics as well as statistical research. Statistical analysis computer applications have the advantage of being accurate, reliable, and generally faster than computing statistics and drawing graphs by hand. Minitab contains various visual tools, including histograms, boxplots and scatterplots, that help professionals perform statistical analysis more efficiently and visualize what the data is telling them. It also enables users to calculate descriptive statistics for their data.

It offers user convenient ways to input statistical data, manipulate that data, identify patterns and trends and then analyze the data to solve real-world problems. It provides

simplified data analysis ideal for use in statistical interpretation at the business level. When Minitab displays charts, patterns or trends, it also provides an accompanying analysis and interpretation to help users draw helpful, practical conclusions.

Numerical data is the only type Minitab will use for statistical calculations. Text cannot be used for computations. Minitab will conduct a variety of statistical calculations. These are found under the main menu option of **STAT**.

(3) Excel:

Microsoft Excel is a commercial spreadsheet application written and distributed by Microsoft for Microsoft Windows and Mac OS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. It has a better supply of functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display.

Excel 2007 onwards uses the new Ribbon menu system. This is different from what users are used to, but the number of mouse-clicks needed to reach a given functionality is generally less.

Tracking and analyzing results and patient data can consume much of the workday of medical professionals. We can simplify tracking and analysis duties so that more time is focused on patient care. By using Excel, we can create reports to analyze data in many meaningful ways — for example, you can analyze each unit by month, compare completion of tasks by medical representatives, and audit charts by month or unit. We can customize the electronic medical record (EMR) system to track and analyze medical data in the ways that we need.

Most of Excel's statistical procedures are part of the **data analysis tool pack**, which is in the tools menu. It includes t-tests, correlations, descriptive statistics, one or two-way analysis of variance, regression, etc. Microsoft Excel 2000 (version 9) provides a set of data analysis tools (Analysis ToolPak), which can be used for the development of complex statistical analysis. The Data Analysis ToolPak has a descriptive statistics tool that provides summary statistics for a set of sample data. Summary statistics include mean, mode, median, minimum, maximum, standard error, standard deviation, variance, skewness, range, etc. **Descriptive analysis** is observed by going to Excel Data → Data Analysis → Descriptive statistics.

Thesis Writing

Components of Thesis

The basic elements of a thesis are: Abstract, Introduction, Literature Review, Methods, Results, Discussion, and Conclusion.

1. **Abstract:** The abstract is the overview of your thesis and generally very short. It is recommended to write it last, when everything else is done.
2. **Introduction:** The introduction chapter is there to give an overview of your thesis' basics or main points. It should answer the following questions:
 - a. Why is the topic being studied?
 - b. How is the topic being studied?
 - c. What is being studied?
3. **Literature review:** Literature review is often part of the introduction, but it can be a separate section. It is an evaluation of previous research on the topic showing that there are gaps that your research will attempt to fill. A few tips for your literature review:
 - a. Use a wide array of sources
 - b. Show both sides of the coin
 - c. Make sure to cover the classics in your field
 - d. Present everything in a clear and structured manner
4. **Methods:** The methodology chapter outlines which methods you choose to gather data, how the data is analyzed and justifies why you chose that methodology. It shows how your choice of design and research methods is suited to answering your research question. Make sure to also explain what the pitfalls of your approach are and how you have tried to mitigate them. Discussing yourself where your study might come short can give you more credibility as it shows the reader that you are aware of the limitations of your study.
5. **Results:** The results chapter outlines what you found out in relation to your research questions or hypotheses. It generally contains the facts of your research and does not include a lot of analysis, because that happens mostly in the discussion chapter. What helps making your results chapter better is to clearly visualize your results, using tables and graphs, especially when summarizing, and to be consistent in your way of reporting. This means sticking to one format to help the reader evaluate and compare the data.
6. **Discussion:** The discussion chapter includes your own analysis and interpretation of the data you gathered, comments on your results and explains what they mean. This

is your opportunity to show that you have understood your findings and their significance. Point out the limitations of your study, provide explanations for unexpected results, and note any questions that remain unanswered.

7. **Conclusion:** This is probably your most important chapter. This is where you highlight that your research objectives have been achieved, and how you have contributed to all parties involved with your research. In this chapter you should also point out the limitations of your study, because showing awareness of your limitation gives a better grounding on your thesis. You can talk about your personal learnings here and also make suggestions for future research.

Paraphrasing

Paraphrasing is putting someone else's ideas into your own words. Paraphrasing a source involves changing the wording while preserving the original meaning. Paraphrasing is an alternative to quoting (copying someone's exact words and putting them in quotation marks). The following steps can be used for paraphrasing:

- ➔ Read the original text until you grasp its meaning; then set it aside.
- ➔ Using your memory, write down the main points or concepts. Do not copy the text verbatim.
- ➔ Change the structure of the text by varying the opening, changing the order of sentences, lengthening or shortening sentences, etc.
- ➔ Replace keywords within the sentences with synonyms or phrases with similar meanings.
- ➔ Check your notes against the original to ensure you have not accidentally plagiarized.

Plagiarism

Plagiarism is presenting someone else's work or ideas as your own, with or without their consent, by incorporating it into your work without full acknowledgement. All published and unpublished material, whether in manuscript, printed or electronic form, is covered under this definition.

Although plagiarism in some contexts is considered theft or stealing, the concept does not exist in a legal sense, although the use of someone else's work in order to gain academic credit may meet some legal definitions of fraud. Some cases may be treated as unfair competition or a violation of the doctrine of moral rights. In short, people are asked to use the guideline, "if you did not write it yourself, you must give credit".

Plagiarism is concerned with the unearned increment to the plagiarizing author's reputation, or the obtaining of academic credit, that is achieved through false claims of authorship. Thus, plagiarism is considered a moral offense against the plagiarist's audience

(for example, a reader, listener, or teacher).

Types:

Direct plagiarism is the word-for-word transcription of a section of someone else's work, without attribution and without quotation marks.

Self-plagiarism occurs when a student submits his or her own previous work, or mixes parts of previous works, without permission from all professors involved.

Mosaic Plagiarism occurs when a student borrows phrases from a source without using quotation marks, or finds synonyms for the author's language while keeping to the same general structure and meaning of the original.

Accidental plagiarism occurs when a person neglects to cite their sources, or misquotes their sources, or unintentionally paraphrases a source by using similar words.

Plagiarism checker Software:

Software system that takes a submitted text as input, and compares the text against a set of publicly available and privately held documents, resulting in a similarity report. The similarity report includes marking of similar or identical text, hyperlinks or other references to sources that match it, and an overall similarity report. In some cases, the similarity report can be customized to include or not quoted text, and to include or not the bibliography (where the use of a style will ensure that the text of any citation matches the same text in any other document citing the same source).

Online computer programs or web interfaces of such programs developed specifically to identify cases where someone's work is presented totally or partially without giving credit to its owner and without applying proper citation practices.

References

A references page is the last page of a research paper which lists all the sources you've used in your project, so readers can easily find what you've cited. Referencing allows you to acknowledge the contribution of other writers and researchers in your work. Any university assignments that draw on the ideas, words or research of other writers must contain citations. Referencing is also a way to give credit to the writers from whom you have borrowed words and ideas.

Referencing is a way to provide evidence to support the assertions and claims in your own assignments. By citing experts in your field, you are showing your marker that you are aware of the field in which you are operating. Your citations map the space of your discipline and allow you to navigate your way through your chosen field of study, in the same way that sailors steer by the stars.

References should always be accurate, allowing your readers to trace the sources of information you have used. The best way to make sure you reference accurately is to keep a

record of all the sources you used when reading and researching for an assignment.

Bibliography

Reference implies the list of sources that has been referred in the research work. Bibliography is about listing out all the materials which has been consulted during the research work. A bibliography is a list of works on a subject or by an author that were used or consulted to write a research paper, book or article. It can also be referred to as a list of works cited. It is usually found at the end of a book, article or research paper.

The general practice of entering sources may be mentioned as follows:

- (i) The author name comes first in the bibliographical arrangement.
- (ii) The title of the book which comes next to the author's name is underlined or italicized. In case of article it is put within inverted comma.
- (iii) Then details of publication are mentioned.
- (iv) Each entry begins flush with the left margin. The subsequent lines are indented in case the entry exceeds the line.
- (v) The entries are separated by a double space.
- (vi) In case two or more works of the same author consecutively appear the author's name is represented by a short solid line which may be termed as continuous line.

Research publication

Publications make scientific information publically available, and allow the rest of the academic audience to evaluate the quality of the research. There are various types of publications like Scholarly Journals, Professional or Trade Publications, Popular and General Interest Magazines. Scientific publications have their own identity, place and necessity. The Academic publications have a peer review system which maintains novelty, applicability and advancement in a given field of knowledge. There is fast online publication process to enhance the publication frequency and reduce the in process time expenditure.

Publishing your research is an important step in your academic career. **Choosing which journal** to publish your research paper in is one of the most significant decisions you have to make as a researcher. Where you decide to submit your work can make a big difference to the reach and impact your research has. **Writing** an effective, compelling **research paper** is vital to getting your research published. Everything from the style and structure you choose to the audience you should have in mind while writing will differ, so it's important to think about these things before you get stuck in. Once you've chosen the right journal and written your manuscript, the next step in publishing your research paper is to make your **submission**. Each journal will have specific submission requirements. To submit your manuscript you'll need to ensure that you've gone through all the steps in our making your submission guide. This includes thoroughly understanding your chosen journal's

instructions for authors, writing an effective cover letter, navigating the journal's submission system, and ensuring your research data is prepared as required. **Peer review** is the independent assessment of your research article by independent experts in your field. Reviewers, also sometimes called 'referees', are asked to judge the validity, significance, and originality of your work. If your paper is accepted for publication, it will then head into **production**. At this stage of the process, the paper will be prepared for publishing in your chosen journal.

Impact factor

The impact factor (IF) is a measure of the frequency with which the average article in a journal has been cited in a particular year. It is used to measure the importance or rank of a journal by calculating the times its articles are cited.

Impact factor can be calculated after completing the minimum of 3 years of publication; for that reason journal IF cannot be calculated for new journals. The journal with the highest IF is the one that published the most commonly cited articles over a 2-year period.

It is used for the relative importance of a journal within its field, with journals with higher impact factors deemed to be more important than those with lower ones. Other related Indices are h-Index, I10-index etc.

Publication ethics

Ethical standards for publication exist to ensure high-quality scientific publications, public trust in scientific findings, and that people receive credit for their ideas. It is important to avoid:

- (i) **Data fabrication and falsification:** Data fabrication means the researcher did not actually do the study, but made up data. Data falsification means the researcher did the experiment, but then changed some of the data.
- (ii) **Plagiarism:** Taking the ideas and work of others without giving them credit is unfair and dishonest.
- (iii) **Multiple submissions:** It is unethical to submit the same manuscript to more than one journal at the same time.
- (iv) **Redundant publications:** This means publishing many very similar manuscripts based on the same experiment. It can make readers less likely to pay attention to your manuscripts.