BASIC BIOSTATISTICS FOR CLINICAL RESEARCHERS

Published by:
LUMHS Publishers

AUTHORS

PROF. DR. BINAFSHA MANZOOR SYED, PHD
PROF. DR. JAWAID NAEEM QURESHI, FRCS
PROF. IKRAM DIN UJJAN, M.PHIL, FCPS, PHD
DR. FAISAL HYDER SHAH, PHD
DR. FASIHA SHAH, PHD
DR. SIKANDER MUNIR MEMON, PHD

ISBN:978-627-7526-00-9

Basic Biostatics for clinical researchers

Authors

Prof. Binafsha Manzoor Syed, PhD

Prof. Jawaid Naeem Qureshi, FRCSIre, FRCSEd

Prof. Ikram Din Ujjan, M.Phil, FCPS, PhD,

Dr. Faisal Hyder Shah, PhD

Dr. Fasiha Faisal Shah, PhD

Dr. Sikandar Munir Memon, PhD

ISBN: 978-627-7526-00-9

Copyright:

LUMHS Publishers

Dedication

To

Ami (Kulsoom Manzoor Shah),

Abu (Syed Manzoor Ali Shah)

and

Chacha Bha (Syed Shabbir Ali Shah)

Preface

As medical students, we all are frightened by even the thought of mathematics; I chose medicine because I used to hate maths. As a postgraduate student, when I started my research work and found one portion of my research proposal with a heading of "statistical methods", that made me feel hypoglycemic. I did not know what it was. Then as a PhD student, I had no escape from statistics, so I finally decided to confront this fear. I took all available short books and started learning statistics. It did not take long, but statistics became one of my favourites in a few months. In 2012, I thought of writing a simple book to make biostatistics digestible for medical students and researchers. Then, all my co-authors helped in completing this book. I hope while reading this book; statistical methods will make complete sense.

All the best.

Prof. Dr. Binafsha Manzoor Syed, PhD
Director ORIC & MRC
Liaquat University of Medical & Health Sciences, Jamshoro, Sindh
Principal Research Officer, HRI (Ex-PHRC), Islamabad
Director, Asian Consortium on Arsenic research, China
Director, Global Health Consortium, China

Email: binafsha.syed@lumhs.edu.pk

Acknowledgment

Everyone who helped and inspired me

Table of content:

Chapter 1

Introduction to Biostatistics

1. Introduction to Biostatistics

Statistics is a branch of applied mathematics which deals with the study of a population of human beings living in a political union, as defined by Sir Ronald A. Fisher in his book *"statistical methods for research workers"*. However, Kuzma defined statistics as a body of techniques and procedures for the collection, organization, analysis, interpretation and presentation of the information which can be given numerically (quantitative). Though, this definition straight away excludes qualitative information. The statistics deal with the information, collectively called ***"data"***. Thus, it uses mathematical and statistical techniques to analyze and draw conclusions from data, making it a critical tool in numerous fields, including science, engineering, medicine, economics, and social sciences.

Statistics can describe a population or a sample, helps to make predictions, and test hypotheses about the relationships between different factors (i.e. variables).

Statistics is a critical tool for business, government, and organizations' decision-making. It provides a way to objectively evaluate and quantify information, assess risk, and make decisions based on data rather than intuition or guesswork. It is also used in research to test theories, analyze trends, and identify patterns and relationships in data.

Biostatistics is a branch of statistics which deals with the methods used in medical and health sciences.

Types of statistics

There are two main types of statistics: descriptive statistics and inferential statistics.

1. ***Descriptive statistics***: This type of statistics deals with a simple description of the data in a meaningful way without making future predictions; in other words, just a presentation of the observed facts. Descriptive statistics summarize large data from a sample into single-digit information, typically through measures like mean, median, mode, range, and standard deviation (SD). These statistics help us understand the shape and distribution of a dataset and can provide insights into parameters like central tendency, variability, and outliers. For example, presenting the average number of patients visiting outpatient department per month.
2. ***Inferential statistics***: This type of statistics involves using data to make inferences or predictions about a larger population. This is done through hypothesis testing, confidence intervals, and regression analysis. These statistics help us draw conclusions about how likely it is that specific findings are due to chance or whether they represent an actual effect or relationship—for example, getting the average number of patients visiting per month for the last six months and assuming the number of patients visiting the outpatient department

next month. It also deals with applying the information on the general population drawn from a sample.

Applications of biostatistics in Medical research

Medical science is a constantly evolving field. Every now and then, innovations are coming up for diagnostics, management and prevention of diseases, and some have revolutionised the healthcare system. Biostatistics is used to see the utilization of these advances and compare them with the available modalities for the same purpose. The application of the clinical trial data for approval of a new drug for its utilization in clinical practice is a typical example. Understanding the pattern of occurrence of a disease to prepare for its prevention when required is also the utilization of statistical tools. Using the data from case-control studies to analyse the risk factors for a particular disease is again a function of statistics. Clinicians need to decide to accept the change in the management pattern for any disease based on the data provided by pharmaceutical companies. This decision will be intelligent if the clinicians understand statistics to check if the data is appropriately analysed and presented. The Food and Drug Authority (FDA) 's approval needs quality data drawn from carefully designed studies with appropriate analysis and interpretation. To understand the logic for drug approval, clinicians should understand the basics of the statistics related to clinical research.

Data

"Data is a set of the relevant observations collected for each variable of interest from a selected sample/ population".

Types of data

There are at least two ways to categorize data: statistical and methodological. Statistically, there are two main types of data:

1. Qualitative data

This form of data is collected regarding the quality, such as the feelings, emotions, expectations etc. This kind of data can be summarized as qualitative data using particular methods or transformed into quantitative data for statistical analysis in numerical (numbers).

2. Quantitative data

This form of data is collected in numbers. The information gathered is then used for statistical analysis. Such numerical data include age in years, blood pressure, HBA1c measurements and many more.

While methodologically, the data can be divided as retrospective or prospective, Figure 1.1 (a and b) summarizes the methodological data types.

1. **Prospective data**: This data is collected after designing the study and fulfilling codal formalities such as funding and ethical approval. This data is appropriately planned and expensive in terms of time and money. This type of data has the beauty of following the definition of the variables as designed in the study with relatively less risk of missing variables.

2. **Retrospective data**: This form of data is collected for one study or for some other purpose or as an institutional database, which is later retrieved and utilised for another study. Since the data was collected for other reasons, there is a risk of missing variables. However, this data is less expensive in terms of time of money.

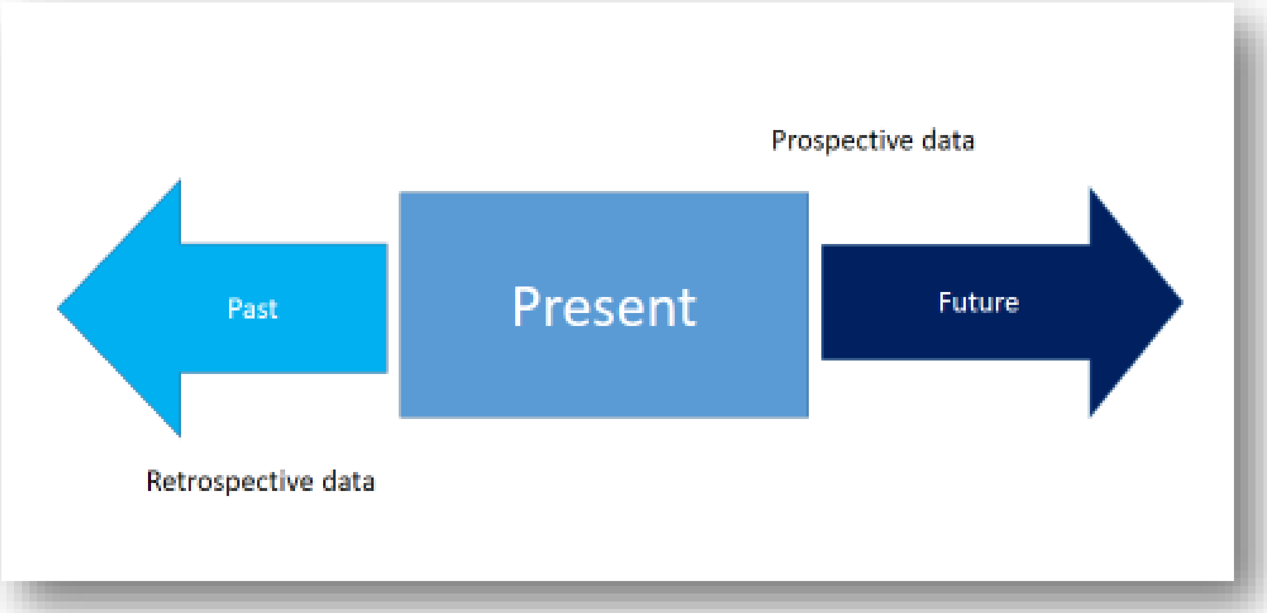
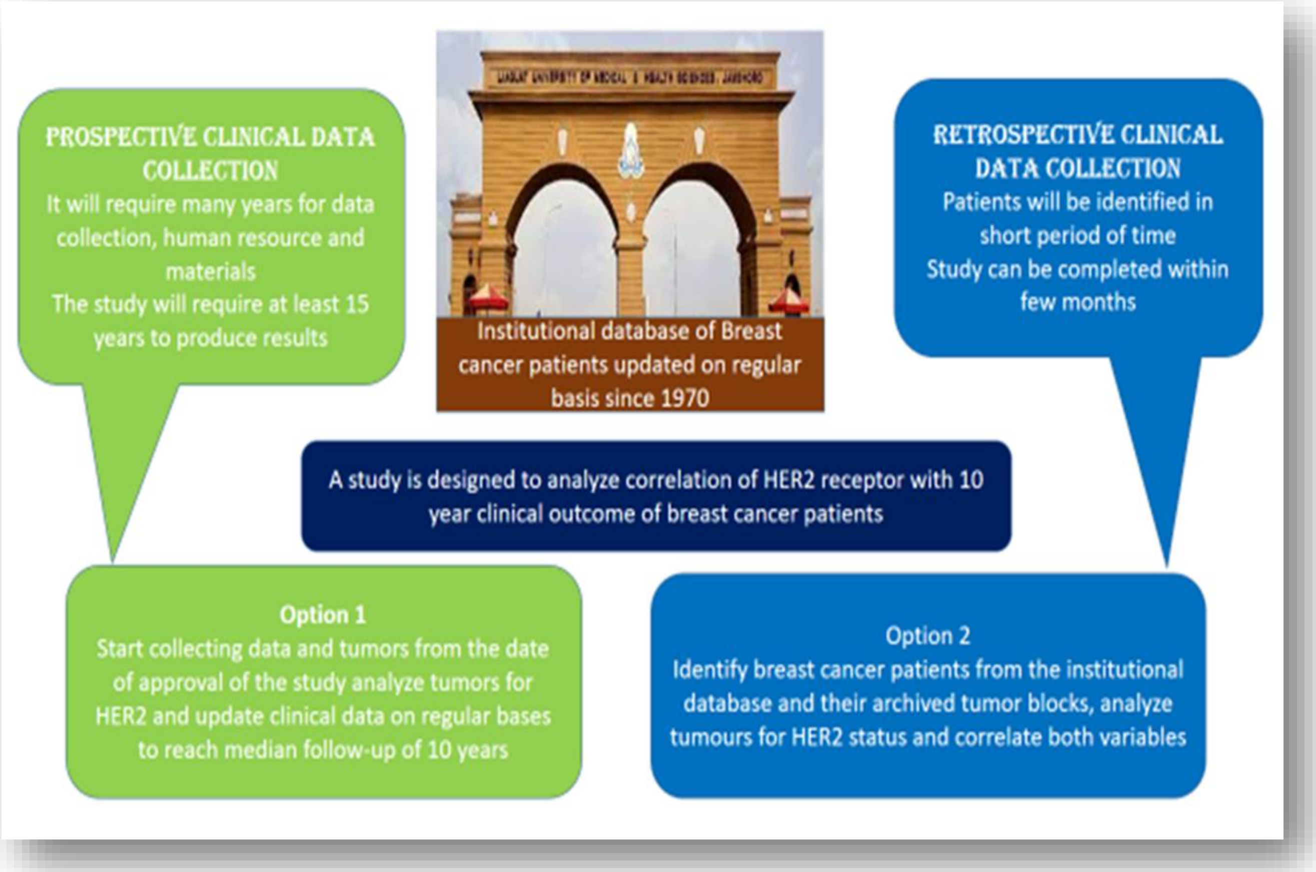


Figure 1.1.a. Methodological types of data



PROSPECTIVE CLINICAL DATA COLLECTION
It will require many years for data collection, human resource and materials
The study will require at least 15 years to produce results

Institutional database of Breast cancer patients updated on regular basis since 1970

RETROSPECTIVE CLINICAL DATA COLLECTION
Patients will be identified in short period of time
Study can be completed within few months

A study is designed to analyze correlation of HER2 receptor with 10 year clinical outcome of breast cancer patients

Option 1
Start collecting data and tumors from the date of approval of the study analyze tumors for HER2 and update clinical data on regular bases to reach median follow-up of 10 years

Option 2
Identify breast cancer patients from the institutional database and their archived tumor blocks, analyze tumours for HER2 status and correlate both variables

Figure 1.1.b. Example of the methodological types of data

Sources of data collection

1. <u>Surveys</u>

Surveys involve data collection without interfering with the natural progress of the disease process.

a. Direct interviews
b. Indirect investigation
c. Questionnaire-based surveys
d. Data collection from local sources
e. Web-based survey
f. Observational data from laboratories

2. <u>Experiments</u>

This mode of data collection involves the manipulation of the natural progression of the disease and then gathering information on the impact of the manipulation.

a. Laboratory experiments
b. Clinical trials

Variable

A single set of measurable information in a database which varies from one individual to the other is called a *variable*. Such as age and gender are typical examples of variables.

Types of variables

Mathematically variables can be discrete or continuous (Figure 1.2). The discrete are the whole numbers, such as the number of households in a town, while continuous is a scale having a number at a fixed interval; it simply explained that discrete are those where decimal does not make any sense, such as the number of children so someone can have 3 or 4 children but not 3.4 children similarly continuous are those where decimals make sense such as haemoglobin level 9.2.

1. *Continuous variables*:

These variables are measured on a continuous scale. Such variables include age (in units of days, weeks, months and years).
The continuous variables are further defined as

i. **Interval scale**: This continuous variable type has a constant size/distance but not a zero point or point of origin. Temperature is a typical example of an interval scale.

ii. **Ratio scale**: This type of continuous variable with a constant distance with a zero point and a point of origin. Ratio scales are primarily used in laboratory measurements.

2. *Categorical variables*

This is the variable type where the observations are divided into groups/categories, such as nationality, race, social class etc.

i. **Binary / Dichotomous variable**:
Where there are only two types of observations are called binary or dichotomous variables. Such variables include gender (i.e. male and female) and vital status (i.e. dead and alive). Figure 1.3 presents examples of binary variables.

ii. **Nominal variable**:
Where the observations are divided into categories, there is no difference in the status of the categories. Such variables include nationality, race and students studying in class iii in different schools etc (Figure 1.4).

iii. **Ordinal variable**:
These are rank-order variables where one category is superior to the other. Such variables include social classes, level of education, school grade etc (Figure 1.5).

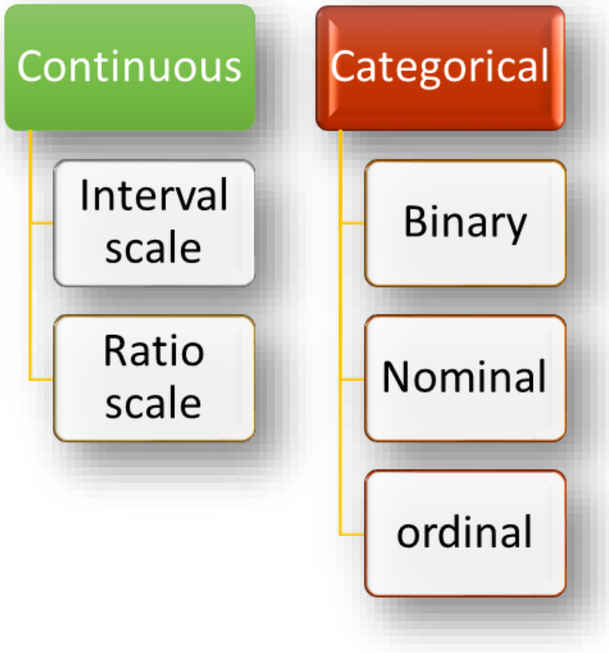


Figure 1.2. Summary of the types of variables

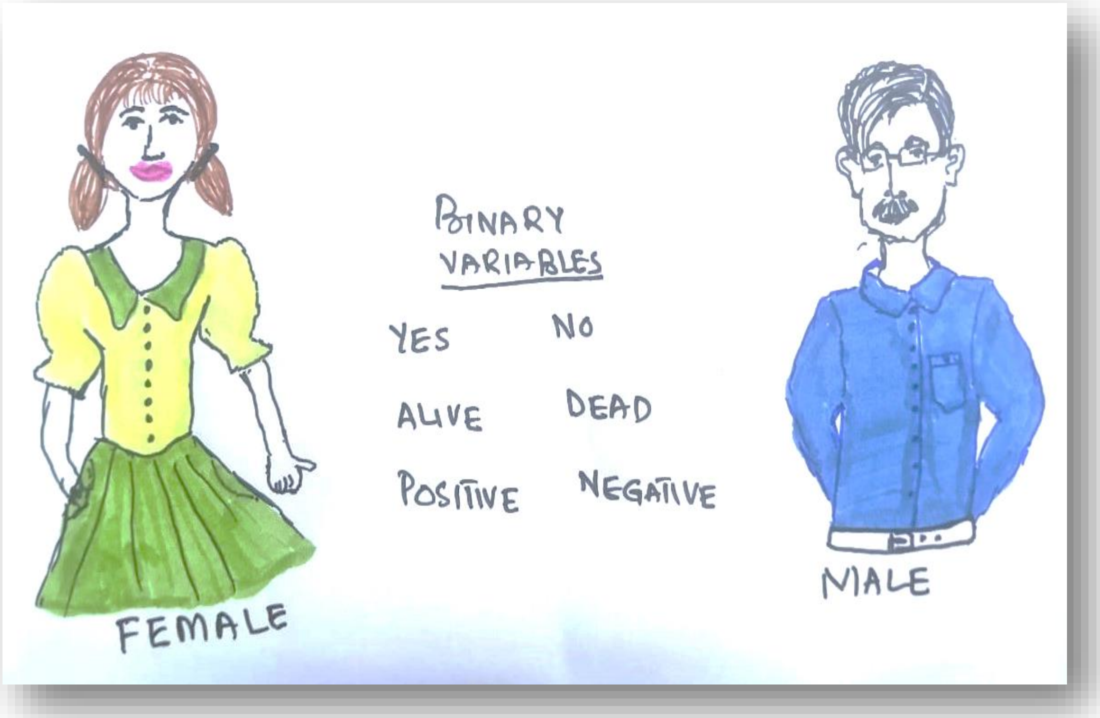


Figure 1.3. Binary – Dichotomous variables

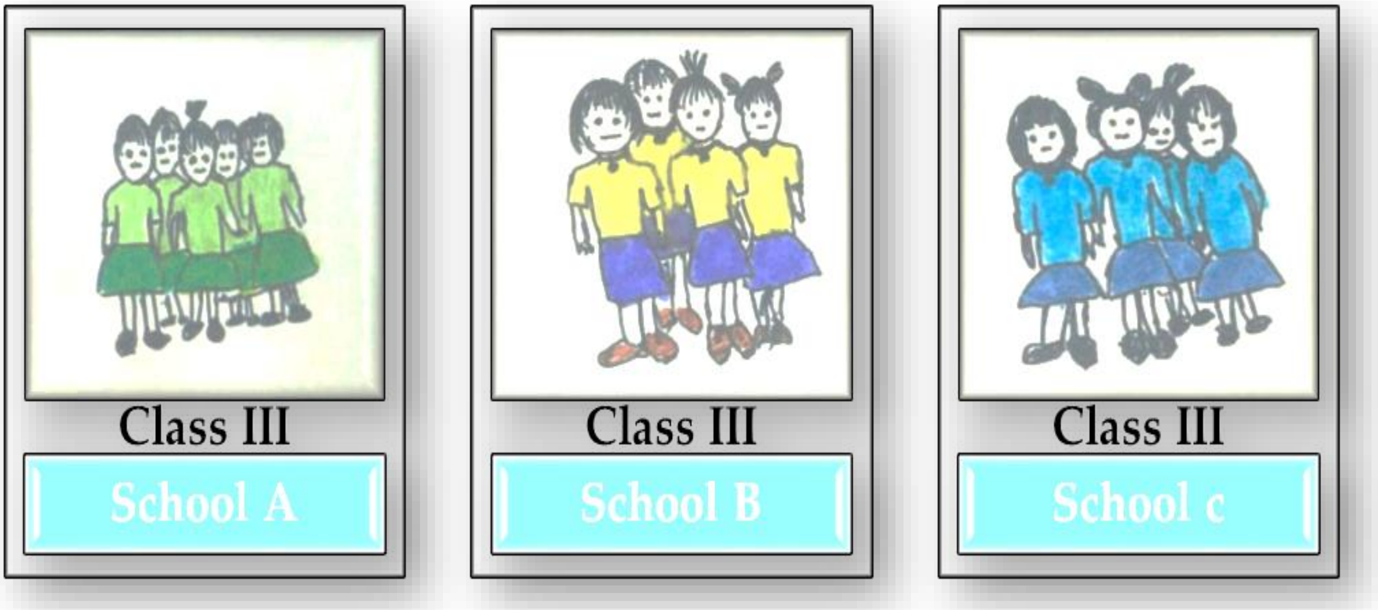


Figure 1.4. Example of Nominal variable- a variety of categorical variable

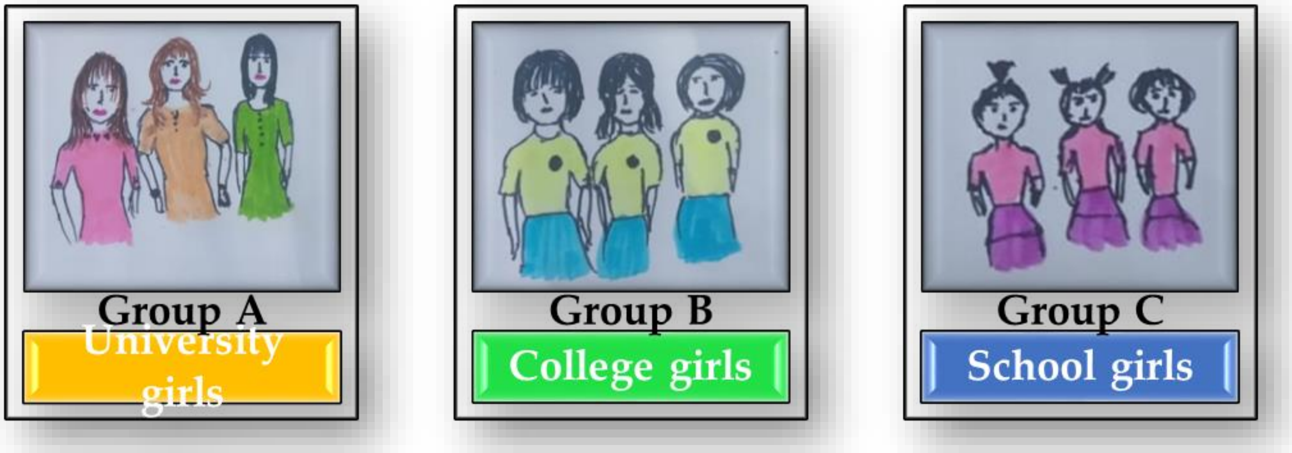


Figure 1. 5. Example of ordinal variable- a variety of categorical variable

Methodological classification of the variables:

1. ***Independent variable***: These are the variables that remain stable; if there is any change in these variables, they will change the other associated variables (i.e., dependant variables). These are also called predictor variables. For example, Amlodipine was assessed for its efficacy in controlling hypertension in a study. The change in the dose from 5 to 10 mg will change the blood pressure measurement. Thus the dose of amlodipine is an independent variable.
2. ***Dependant variable***: This type of variable remains unstable, and any change in the independent variable will change this variable. In the above example, blood pressure measurement is a dependent or outcome variable.

In further simplified form, an independent variable in a statistical model is presumed to cause or influence another variable. It is a variable that the researcher can manipulate or control to observe the effect on the dependent variable. On the other hand, a dependent variable is a variable in a statistical model that is presumed to be influenced by another variable. It is the variable that researchers measure to observe the independent variable's effect.

For example, in a study investigating the effect of exercise on weight loss, exercise is the independent variable because it is being manipulated, and weight loss is the dependent variable because it is being measured to observe the effect of exercise. In daily life examples, children are called dependent on their parents; thus, parents' financial and social status directly affect their children.

It is important to remember that dependent and independent variables are not constant in each project. If an independent variable influences another variable in one study, this might get influenced by another variable in the other study. For example, the number of cigarettes smoked per day during pregnancy influences the baby's birth weight; thus, the number of cigarettes is an independent variable, and birth weight is a dependent variable. Yet another study was conducted to correlate stress level and the number of cigarettes smoked per day; thus, the number of cigarettes becomes dependent on the stress level in this study.

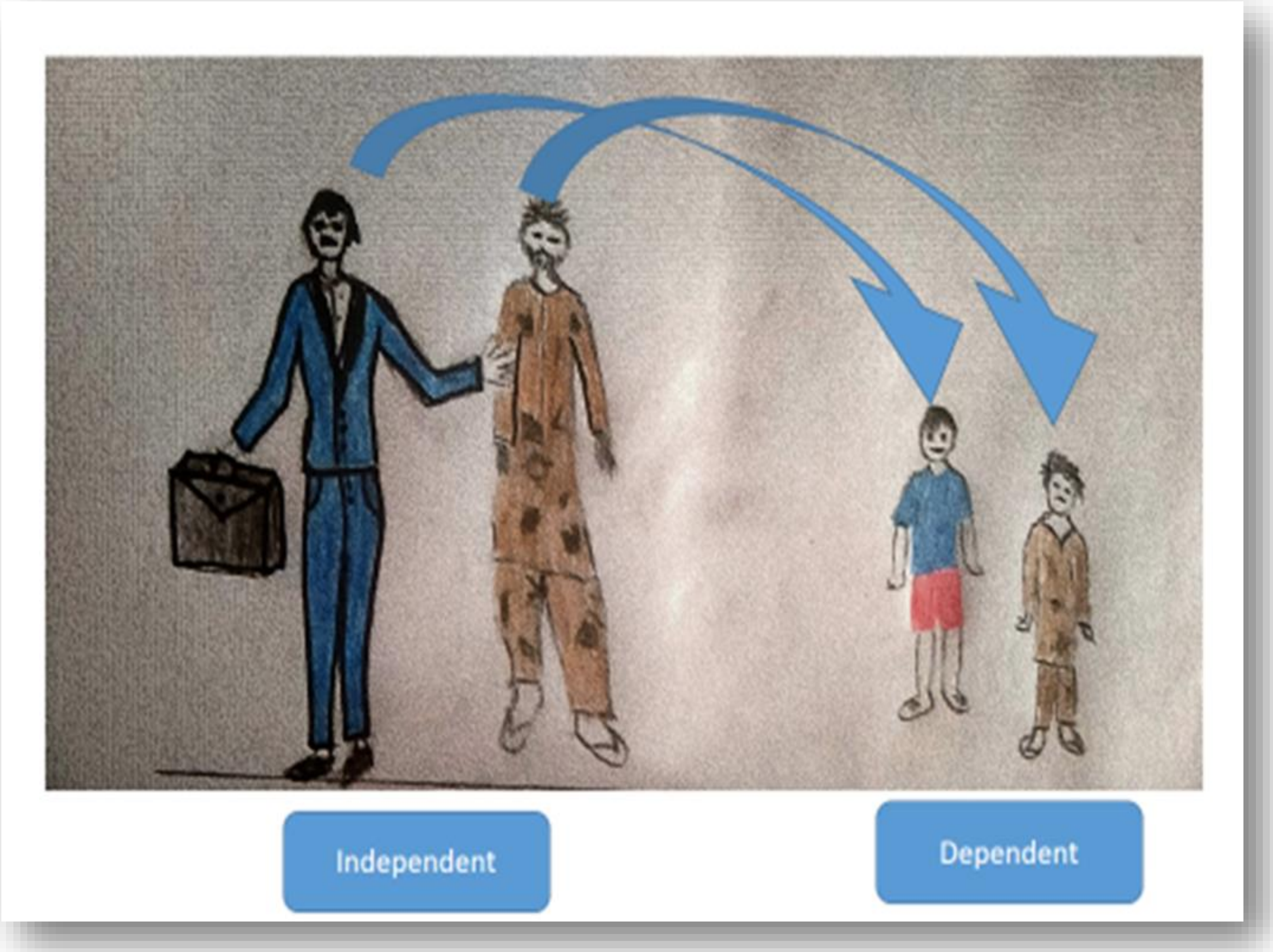


Figure 1.6. Example of an independent and dependent variable

Research studies in clinical sciences

There are several reasons for research in clinical sciences. Therefore, research questions also vary; thus, depending on the research question, many study/ research options are available. A summary of the typical studies done in clinical sciences is given in Figure 1. 7.

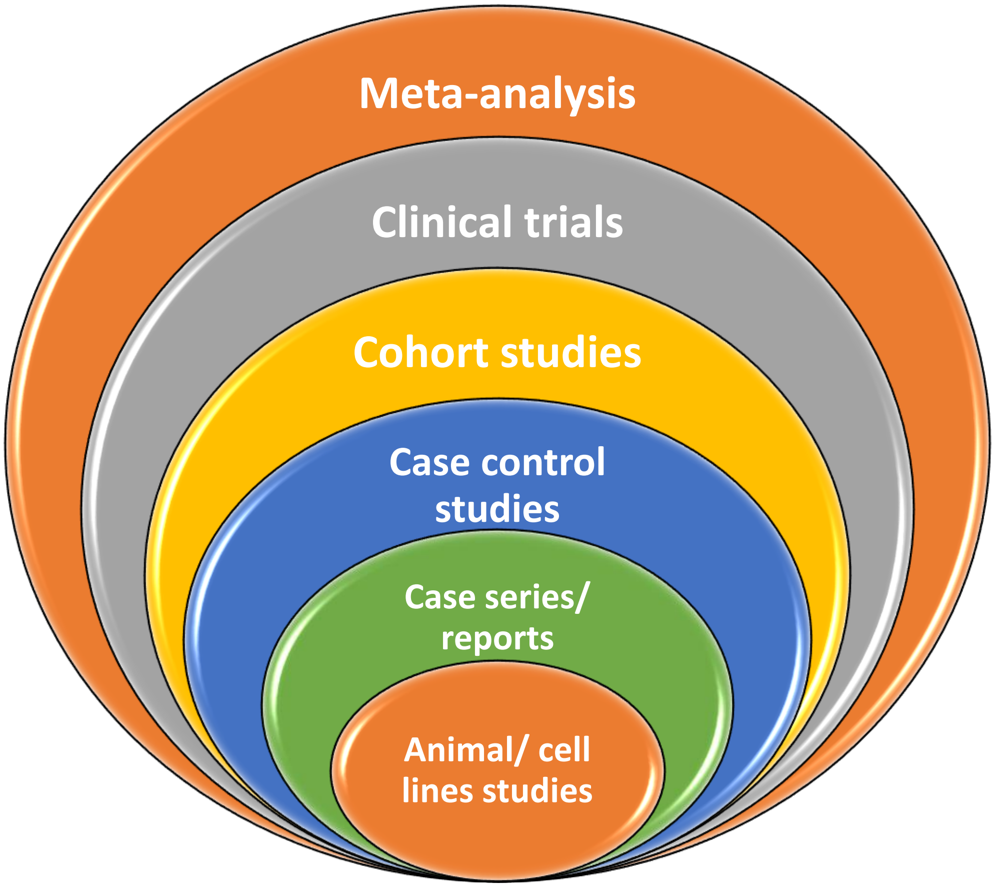


Figure 1.7. Summary of the common type of studies done in clinical sciences

Meta-analysis

This is the re-analysis of the data driven from more than one study following a comprehensive systematic literature review process. There are situations where more than one study is conducted in a controlled environment but drew different conclusions, such as many randomized controlled trials were conducted to test the efficacy of a new drug in different continents and draw different results, thus all studies will be analyzed together (Figures 1.8, 1.9 and 1.10). In these situations, raw data from all the relevant studies are collected from the investigators and re-analyzed to make a consensus on the results. In a meta-analysis, there is a considerable increase in the sample size, and there is also a careful check of the conduct of the study. Thus, any bias can be identified. Nonetheless, reliable results are obtained, providing the highest level of evidence. The meta-analysis then paves the way for robust clinical guidelines.

Cochrane library (https://www.cochranelibrary.com/) meta-analyses all the clinical trials conducted to find the best possible management of different diseases and revise these every five years. Thus it also guides for further research on the subject.

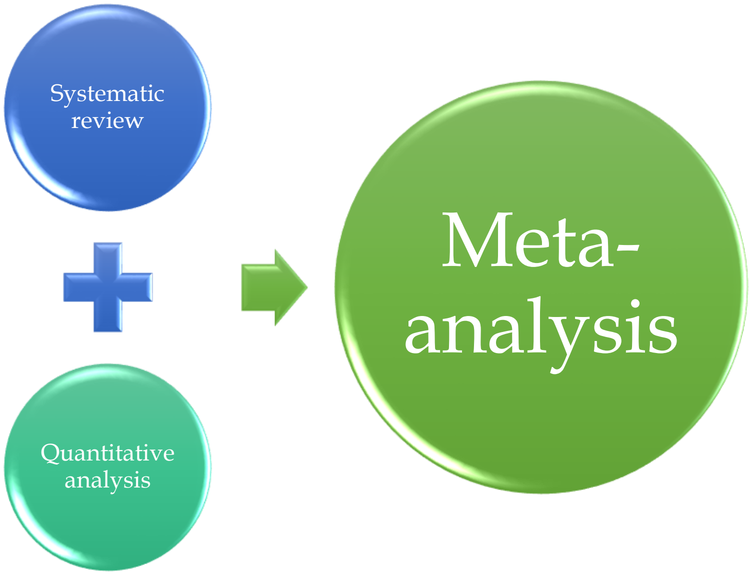
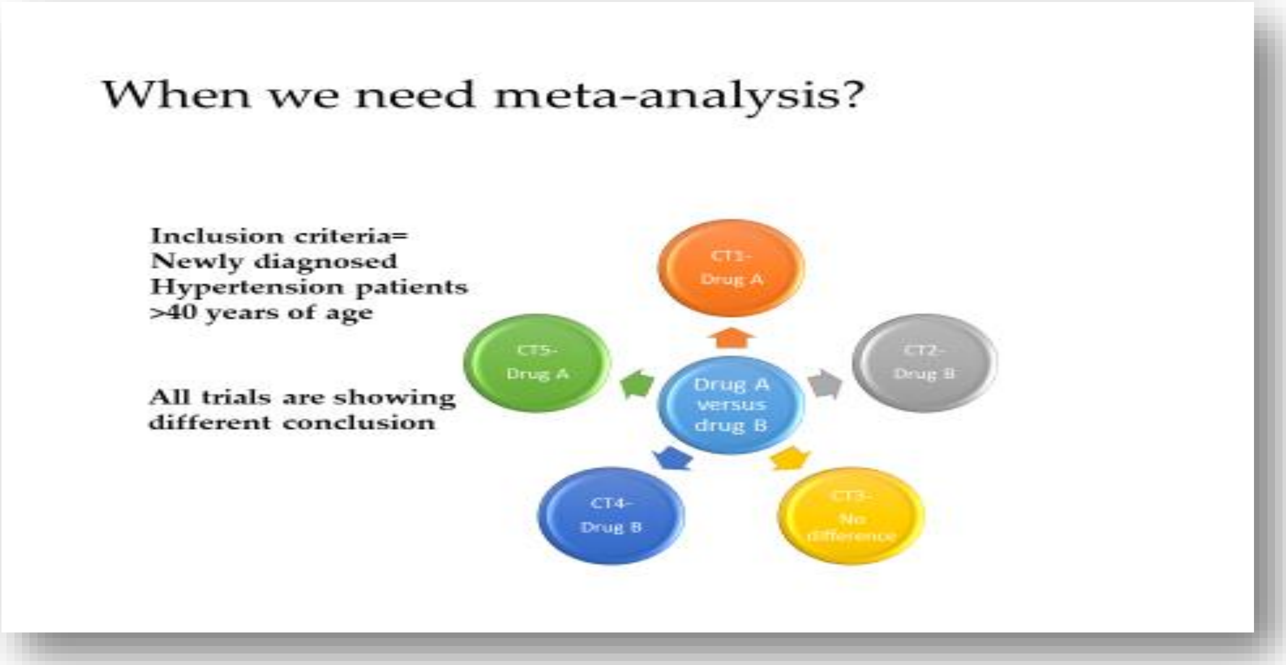


Figure 1.8. Simple definition of meta-analysis



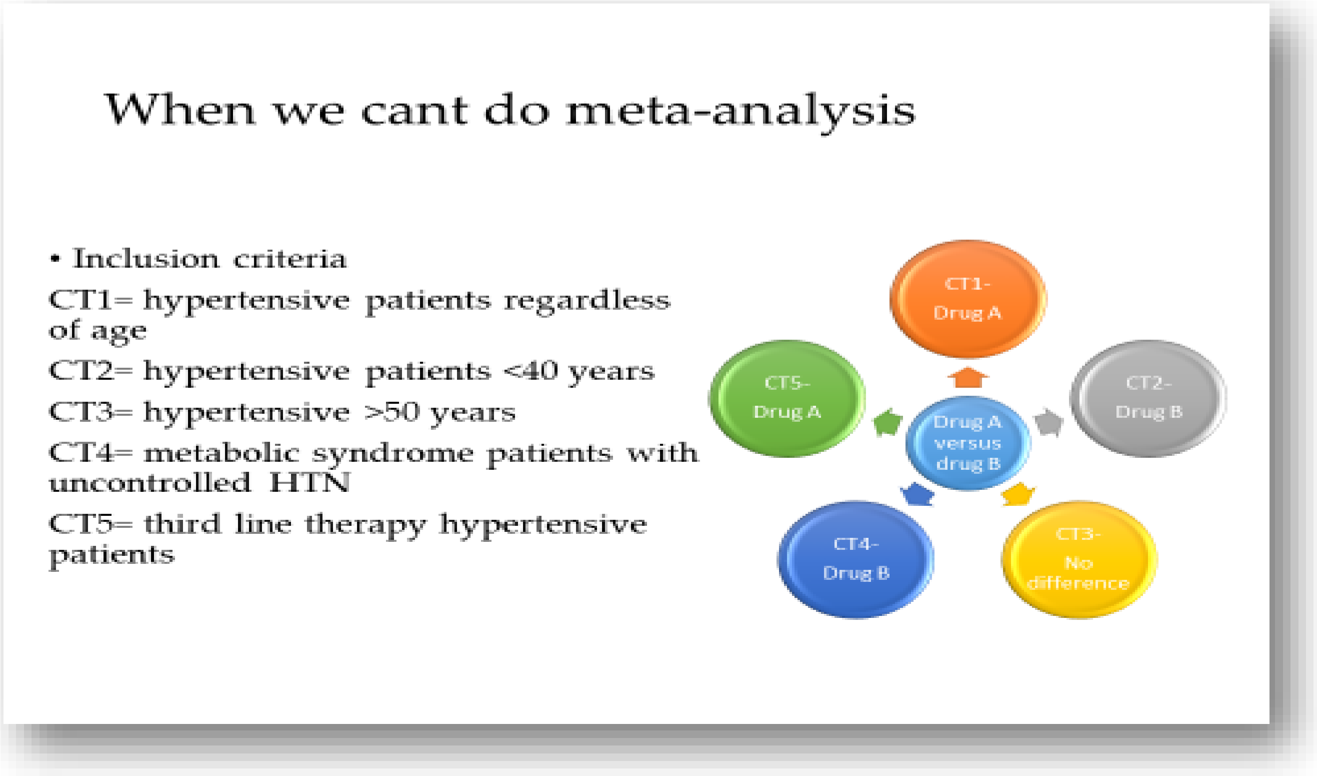
When we need meta-analysis?

Inclusion criteria=
Newly diagnosed
Hypertension patients
>40 years of age

All trials are showing
different conclusion

CT1-
Drug A
CT5-
Drug A
Drug A
versus
drug B
CT2-
Drug B
CT4-
Drug B
CT3-
No
difference

Figure 1.9. Example of the situation when we need meta-analysis



When we cant do meta-analysis

- Inclusion criteria

CT1= hypertensive patients regardless of age

CT2= hypertensive patients <40 years

CT3= hypertensive >50 years

CT4= metabolic syndrome patients with uncontrolled HTN

CT5= third line therapy hypertensive patients

CT1-
Drug A
CT2-
Drug B
CT3-
No difference
CT4-
Drug B
CT5-
Drug A
Drug A versus drug B

Figure 1.10. An example of when we cannot do a meta-analysis

Methodology of Meta-analysis

The first step in conducting a meta-analysis is to set a research question or the research area where it is needed to be done. The next step is the systematic review of literature based on the inclusion and exclusion criteria of the studies and finally compiling the results by re-analysing the raw data of available studies.

Here is the step-by-step method of conducting a meta-analysis:

1. **Define the research question:**

Before conducting a meta-analysis, it is essential to define the research question. Defining the research question appropriately will help choose keywords and help finalise the studies' selection criteria to be included in the meta-analysis.

2. **Conduct a comprehensive literature search:**

Before jumping to the literature review, selecting appropriate keywords, deciding the inclusion and exclusion criteria of the studies, and selecting search engines is essential. Then conduct a thorough search of relevant databases (such as PubMed, Cochrane

Library, or PsycINFO) for studies related to the research question. Use specific keywords and search terms to ensure a comprehensive literature search.

3. **Screen the studies:**

The first step starts with screening through titles, followed by abstracts of all the selected studies, and finally, reviewing full articles of the shortlisted studies which fulfilled inclusion criteria.

4. Evaluate the quality of the studies:

It is essential to evaluate the quality of each study included in the meta-analysis. This can be done by assessing the study design, sample size, and methodology. Studies that are of low quality may be excluded from the meta-analysis.

5. Extract the data:

Extract relevant data from each study, including study design, sample size, statistical data and other relevant information. This data will be used for the meta-analysis. (Here, if it is planned to be a systematic review, then the studies' summary can be presented without re-analysis of the data driven from the selected studies).

6. Analyse the data:

Conduct a statistical analysis of the data using appropriate techniques. This may include calculating effect size, conducting subgroup analyses, and evaluating potential biases.

7. Interpret the results:

Interpreting the results of the meta-analysis includes presenting any significant findings, limitations, and potential implications. This can be done by creating a forest plot or other visual data representation.

8. Conclude and make recommendations:

Based on the meta-analysis results, conclude and make recommendations for future research or clinical practice. This also includes highlighting research gaps and directing future research in the areas where it is needed or making a robust conclusion to guide or modify clinical guidelines. The conclusion from the meta-analysis provides the highest evidence for clinical guidelines.

In general, conducting a meta-analysis is a comprehensive and complex process that requires careful attention and a thorough understanding of statistical methods. A detailed guide for conducting a meta-analysis is available from PRISMA (http://prisma-statement.org/).

Clinical trials

The clinical trial is a scientific method to evaluate the effectiveness and safety of new medical treatments, drugs, and devices. It involves collecting data from human subjects to determine the safety, efficacy, and appropriate dosage of a new treatment or drug. Before clinical trials, a new molecule must have passed efficacy and safety trials in animal models. Throughout the clinical trial process, participants are closely monitored for any changes in their condition or health, and their experiences are documented and analyzed by researchers. The goal is to determine whether the new treatment or drug is safe and effective and offers advantages over existing treatments.

The clinical trial research method typically involves four phases:

Phase 1 trials are the initial testing of a new drug or treatment in a small group of healthy volunteers. Researchers evaluate the drug's safety, along with any possible side effects, and determine its appropriate dosage. The dose decision involves testing the drug at different levels of safety in a controlled environment.

Phase 2 trials this phase involves testing the drug or treatment on a larger group of people. In most situations, this phase recruits patients who are treatment exhausted and suffering from the disease of interest. In Phase II the efficacy of the new drug is evaluated using the effective dose driven from the results of the Phase I trial. The safety of the patients remains the second primary objective.

Phase 3 trials this phase is quite popular in clinical research as it compares new effective drugs as proved in Phase II will be tested against standard available therapy. When the standard drug is unavailable or developed for the first time for a newly discovered disease, the new drug will be compared with a placebo. These are generally called randomized controlled trials or in situations where a comparison is made with a placebo; these are called placebo-controlled trials. These studies can involve thousands of participants and take several years to complete. Once the drug is shown to be effective in phase 3 trials, approval from regulatory authorities is processed for marketing of the drug. Figure 1.11 summarizes the method of a clinical trial.

Phase 4 trials are postmarketing surveillance studies involving large-scale data collection, particularly looking at the safety of all patients taking the drug. These patients in post-marketing surveillance take drugs without any controlled environment. This phase is conducted after the drug has been approved by regulatory agencies (i.e. FDA). These studies continue to monitor the safety and effectiveness of the treatment in a larger population over a more extended period till the drug is available in the market.

Following is the step-by-step process of the clinical trial:

1. **Development of the Drug:** The first and crucial step is developing the new drug molecule; that molecule can be synthetic or extracted from a natural compound. Then studying its biochemical structure, followed by cell line and animal model studies. Mouse models are frequently used. Preclinical animal testing is essentially done to ensure safety and efficacy before human trials begin.

2. **Design of the Clinical Trial:** Once the new molecule is found safe and effective in animal models, the next step is to design clinical trials on humans. The trial should be designed to answer specific questions about the drug, including safe and effective doses. This involves identifying the population, such as age, sex, or pre-existing medical conditions, and selecting the endpoints with specific outcomes the trial is designed to measure.

3. **Ethics Approval:** Before the clinical trial can begin, a research ethics committee must approve the study protocol to ensure that the study is ethical and the rights of all the participants (volunteers and patients both) are protected. The ethical committee will also ensure that the patients are not deprived of standard treatment.

4. **Recruitment of Participants:** Potential participants are identified and recruited. They must meet inclusion criteria to qualify for the trial. Informed consent has to be taken from all participants. They must be informed about all potential risks associated with using the new drug, and the benefits are also explained before deciding to participate in the trial.

5. **Randomization:** Participants are randomly assigned to the treatment groups (taking experimental drugs) or the control group (taking standard therapy or the placebo).

6. **Data Collection:** Careful monitoring must be done for patients' compliance in both groups, and data is carefully collected throughout the trial duration. Recording adverse events and any change in the health status are essential.

7. **Analysis of Results:** Once the trial is complete, the data is analysed to determine the efficacy and safety of the drug. The results are compared between the treatment group and the control group.

8. **Publication of Results:** The results are published in a scientific journal, and the dru g is then submitted for approval from regulatory authorities such as the FDA. Once FDA approves the drug, only then can it be marketed.

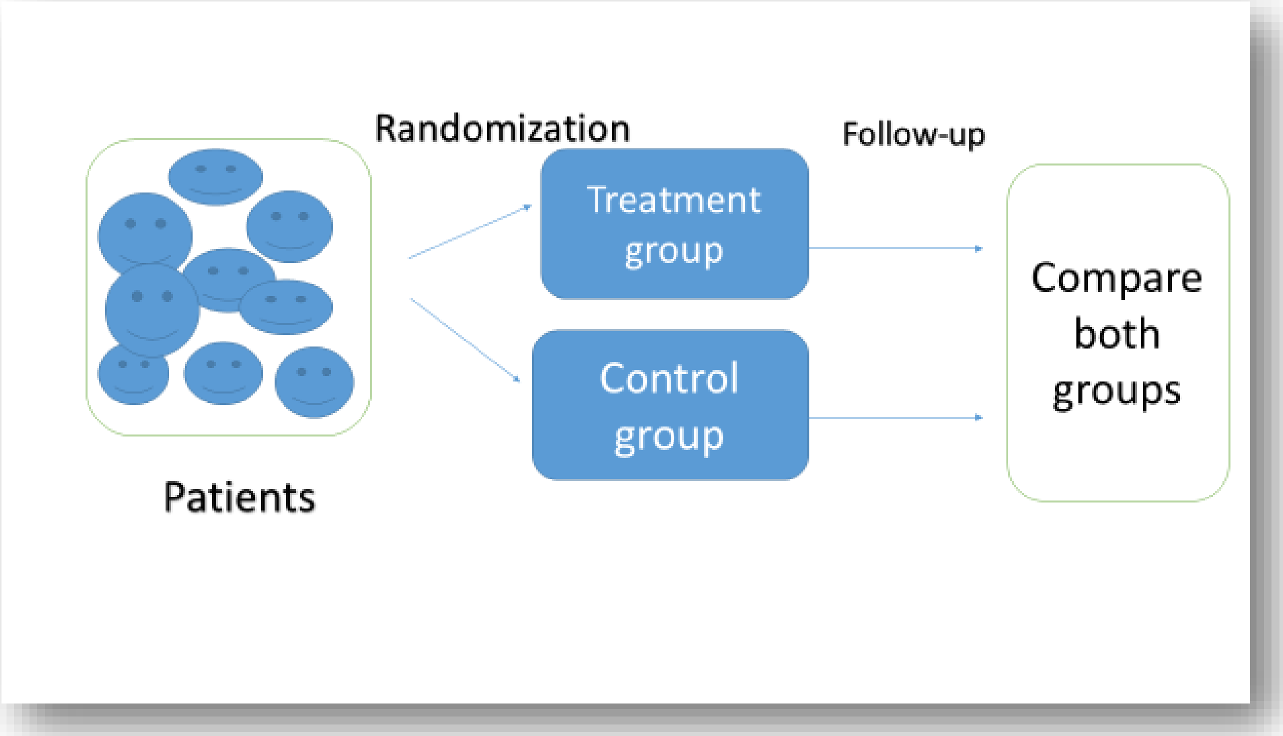


Figure 1.11. A summary of the method of a clinical trial

Cohort study

A cohort study is a type of research design that involves a specific group of patients (called a cohort) showing similar characteristics, such as suffering from a particular disease, then divided into two subgroups based on the difference in the status of a factor of interest. These patients will be followed up for a certain period, and the groups will be compared for the outcome variables. This type of study is often used to investigate the causes of a particular disease or condition. In clinical research, the factors for comparison are studied to evaluate their potential to be a prognostic or predictive factor. These factors may include lifestyle, medical history, laboratory results, or environmental exposures (Figure 1.12).

There are two main types of cohort studies: prospective and retrospective. In a prospective cohort study, participants are identified and followed up forward in time. In a retrospective cohort study, participants are identified, and their medical records are reviewed to collect data on various factors. In chronic illnesses such as cancers, a mixed method is also adopted, such that clinical data is retrieved from existing records, tumour blocks retrieved from archives (retrospective part), and tumour analysis and long-term follow-up are carried out prospectively. Retrospective cohort studies are much less expensive than prospective cohort studies in terms of time and money.

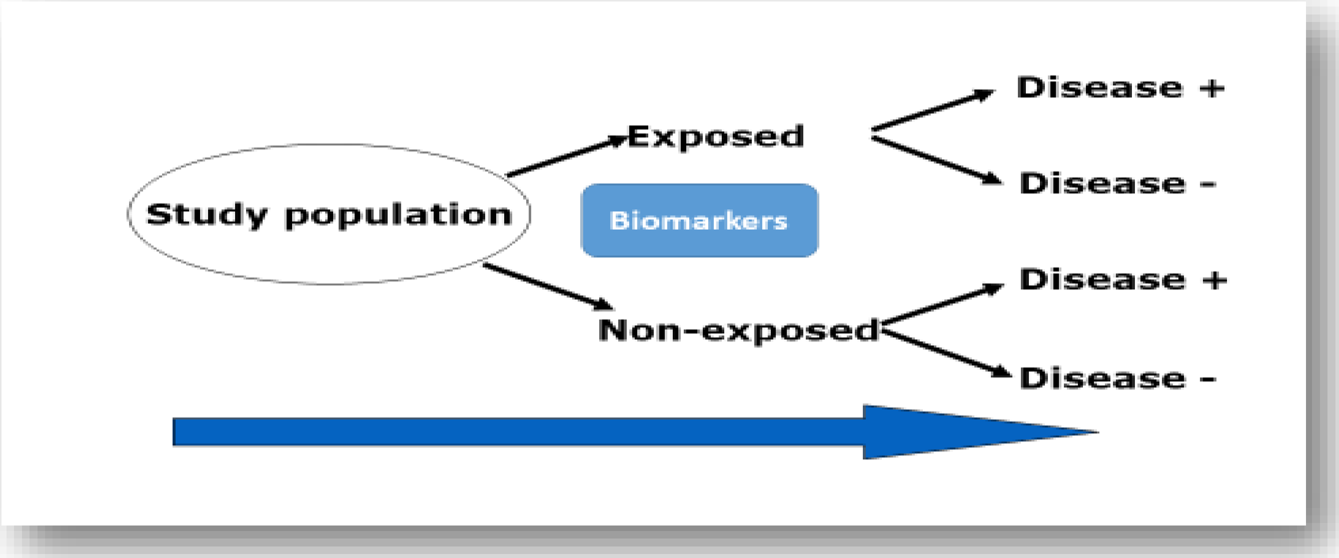


Figure 1.12. A summary of the methodology of a cohort study

Cohort studies are considered a high-quality research design as they confirm/establish a cause-effect relationship between exposure to certain factors in developing a disease or the disease outcome.

The Nurses' Health Study is one of the largest and longest-running cohort studies in medical research. It was initiated in 1976 when over 120,000 female registered nurses in the United States completed a baseline questionnaire regarding their health and lifestyle. Since then, these nurses have been followed up with repeated questionnaires and assessments, providing invaluable data on various health outcomes, including cancer, cardiovascular disease, and reproductive health. The study has led to numerous significant discoveries and has had a major impact on public health policy and medical practice.

Case-control study

A case-control study is an observational study exploring the relationship between an outcome and a potential risk factor or exposure. By default, it is a backward study compared to clinical trial and cohort studies where they go in the future (move forward). In a case-control study, individuals with a particular outcome or disease (cases) are identified and compared with the matched group of people without outcome or disease (controls).

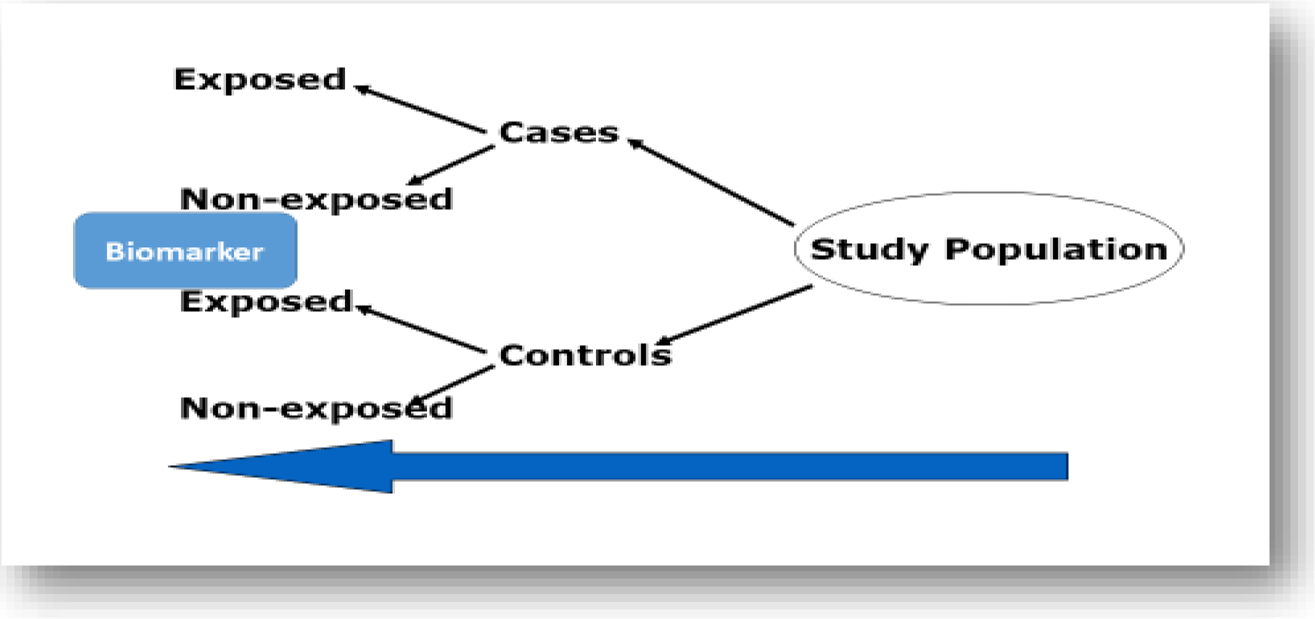


Figure 1.13. A summary of the methodology of a case-control study

The prime objective of this kind of study is to determine any differences between the two groups to identify relative risk. The case-control studies are helpful in rare and complex diseases and explore novel risk associations. The study design has the beauty of exploring multiple risk factors simultaneously. However, when the study enquires about past exposure, such as smoking habits, thus there is always a risk of recall bias.

The best example of a case-control study examined whether individuals who develop lung cancer are more likely to have a smoking history than those who do not. Similarly, a history of betel nut chewing can be explored as a risk of oral cancer using a case-control study design.

The most prominent case-control study in medical research is the UK Biobank study, which includes data from over 500,000 individuals. The study aims to investigate the causes and risk factors for a wide range of diseases, and its large size allows for more accurate estimates of disease risk and more robust conclusions about the associations between various factors and disease outcomes.

Case series

In medical research, a case series refers to collecting and analysing data from patients with specific characteristics or medical conditions. Case series studies are often used to explore the natural history of a disease, investigate potential risk factors, and evaluate the efficacy of the proposed treatments. They may provide valuable insights into rare diseases or conditions that are difficult to study through randomized controlled trials. Case series are also helpful in understanding new diseases such as COVID-19. However, because case series studies lack a control group, the findings may be less reliable than those of other study designs. Nonetheless, case series studies remain an essential tool in medical research and can help to guide clinical decision-making and inform future research efforts.

Animal model study

The findings obtained from animal model studies, although they are not directly applicable to human subjects, are a critical part of medical research, as they allow scientists to understand how diseases and treatments might affect the human body. In these studies, researchers use animals (such as mice, rats, or monkeys) to simulate human physiology to identify potential new treatments or test the efficacy and safety of existing treatments. For example, researchers might use animal models to study the effects of a new cancer drug on tumour growth or to investigate the causes of a particular disease. Animal models can also help in testing new medical devices or surgical procedures, allowing researchers to refine their techniques before applying them to humans.

Although animal models provide valuable information about the human body system due to their physiological or anatomical similarities, animal models are not a perfect substitute for the human body. Thus the interpretation of animal model studies needs confirmation in the human body. The general process for creating an animal model begins with identifying a specific disease or condition of interest and determining which animal species would be the best candidate to model the disease. Once animal species are chosen, researchers may introduce genetic mutations or manipulate existing genes in the animal to create the desired disease phenotype.

Researchers then monitor the animal for signs of the disease or condition, using various methods such as behavioural tests, imaging techniques, and blood analysis. Once the animals display the desired disease phenotype, they can be used to test potential treatments or interventions. Ethical concerns must be taken into account while dealing with animal model studies.

Cell line/ cell culture studies

Cell line studies involve growing and testing cells in a laboratory setting. This type of research is often used to understand how certain diseases develop and progress and to test new drugs and treatments. Many types of cell lines can be studied, ranging from cancer to stem cells. Researchers can manipulate these cells in various ways to learn more about how they function and respond to stimuli. Cell line studies are essential in advancing our understanding of human biology and developing novel treatment options for diseases.

Cell culture studies are specialized methods where particular infrastructure and reagents are required. Firstly, the growth medium is a crucial component of cell culture. This liquid or gel-like substance provides cells with the necessary nutrients, growth factors, and other essential components required for cell growth and maintenance. Different types of cells require different growth media, so it's crucial to determine the optimal medium based on the type of cells to be cultured.

The next step is to select an appropriate cell culture dish or flask to grow cells appropriately. The most commonly used cell culture dish is the petri dish, a flat, circular dish with a lid providing a sterile environment. The next step is seeding the cells, which involves transferring cells from the source (such as a tissue sample or previous culture) into the growth medium in the dish. Several methods for seeding cells, including pipetting, scarping and centrifugation, depending on the type of cells and the experimental design. Once cells settle down, they start growing and multiplying. It's essential to maintain the appropriate conditions for the cells to ensure their survival and proliferation; this includes maintaining appropriate temperature, pH, and humidity and provision of culture medium.

Chapter 2.
Understanding Continuous data

2. Understanding continuous data

Data is the information that has been collected and stored for later use. It can come in many forms, including text, images, numbers, etc. It is used in various ways, from scientific research to marketing, for everyday decision-making. It's vital to ensure accuracy and reliability when working with data, as any errors or inaccuracies can lead to incorrect conclusions or decisions.

Continuous data

Continuous data is quantitative data that can take any value within a specific range and is often used in medical research to measure parameters like blood pressure, heart rate, or other physiological variables. It is simply the data type where decimals such as Hemoglobin 9.4 make sense. One of the advantages of continuous data is that it allows for greater precision in measurements. For example, measuring blood pressure to the nearest millimetre of mercury can provide more detailed information than simply categorizing patients as having high or low blood pressure.

Continuous data can be analyzed using various statistical methods, such as regression analysis, to help researchers identify relationships between variables and predict outcomes. However, it is essential to note that continuous data requires careful handling and interpretation. Researchers must ensure accurate and reliable measurements and consider factors such as variability within the population being studied. It is essential to note that if a variable has data input as continuous, it should be collected as continuous, do not make categories when designing the data collection. The fundamental measures to understand continuous data include evaluating central tendency and dispersion.

Measures of central tendency

Measures of central tendency are a way to describe the "centre" of a set of continuous data. There are three most common measures of central tendency, which need to be evaluated as a foundation of the analysis:

1. Mean
2. Median
3. Mode

The median is the middle value in a data set when the values are arranged in ascending or descending order. It is not affected by outliers. The mode is the most frequently occurring value in a set of data. It can be used when there is a single mode or multiple modes. All three measures can help understand the central tendency of a dataset and can provide different insights depending on the nature of the data.

Mean

The mean, also known as the arithmetic average, is a measure of central tendency representing the average value of a numerical data set. It is calculated by adding up all the values in the data set and dividing the sum by the total number of values. The mean is a useful measure because it provides a single value that can summarize the central tendency of a data set. However, outliers can affect it (extreme values that are much higher or lower than the rest of the data). It may not always be the most representative measure of central tendency for a particular data set, especially when the sample size is small Figure 2.1 presents an example of the calculation of the mean.

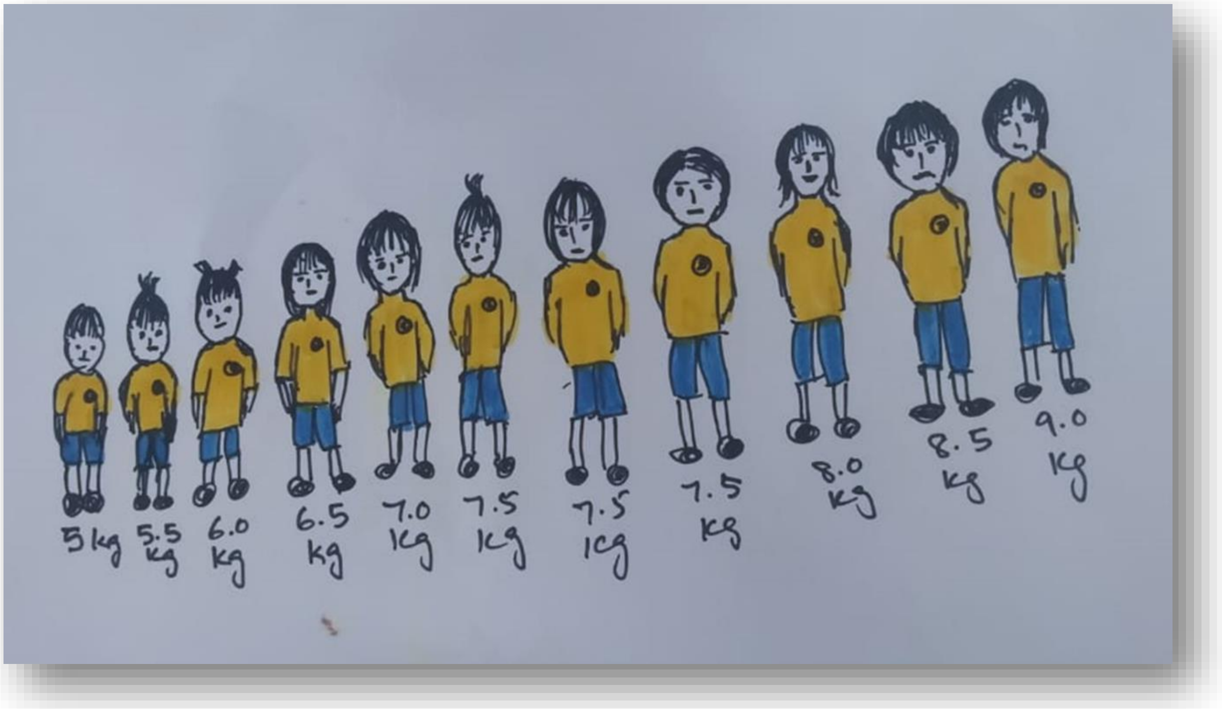


Figure 2.1. Calculation of mean (5+5.5+6.0+6.5+7.0+7.5+7.5+7.5+8.0+8.5+9.0/ 11)

Median

The median is a statistical measure representing the middle value or central value of a data set (dividing the data into two equal parts) when it is arranged in order from lowest to highest (chronological order). In the odd number data set, exactly the middle value is a median; however, if the data set has an even number of values, the median is the average of the two middle values. The median is often used as a measure of central tendency in data sets because it is less sensitive to outliers as compared to the mean. This means that if there are extreme values in the data set, they will not affect the median as much as they would affect the mean. This makes it a more robust measure of central tendency in such cases. Figure 2.2 presents an example of the calculation of the median.

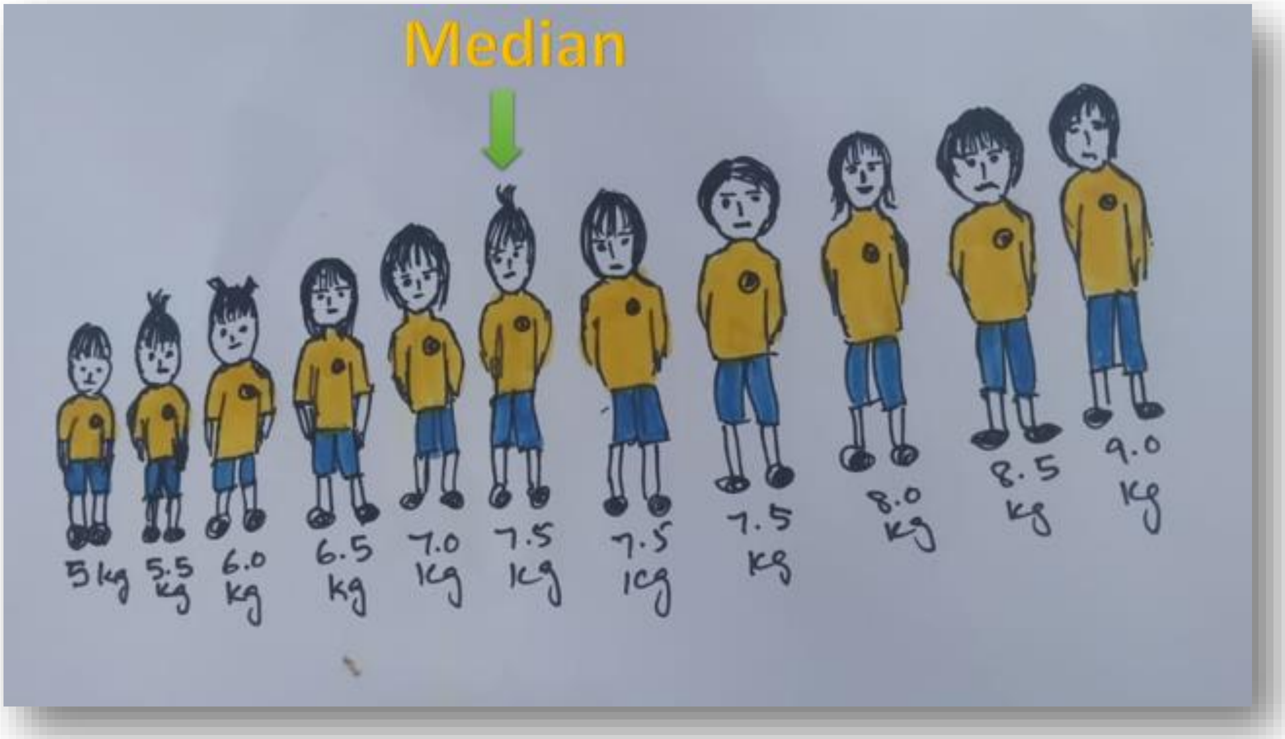


Figure 2.2. Calculation of median (exactly middle observation)

Mode

In statistics, mode is a measure of central tendency that describes the most frequently occurring value in a dataset. Finding the mode can be useful when trying to understand the typical or most common value in a dataset. For example, if you were analyzing the ages of a group of people, the mode would be the age that appears most frequently.

It's also worth noting that a dataset can have more than one mode, which means that there are multiple values that appear with the same frequency. In such cases, the dataset is said to be multimodal. Calculating the mode is relatively simple. To find the mode of a dataset, you simply need to identify which value appears most frequently. This can be done by looking at a frequency distribution table, which displays the number of times each value appears in the dataset. Alternatively, a histogram or bar chart can be created to visualize the frequency of each value, which can make it easier to identify the mode. Figure 2.3 presents calculation of mode in a dataset.

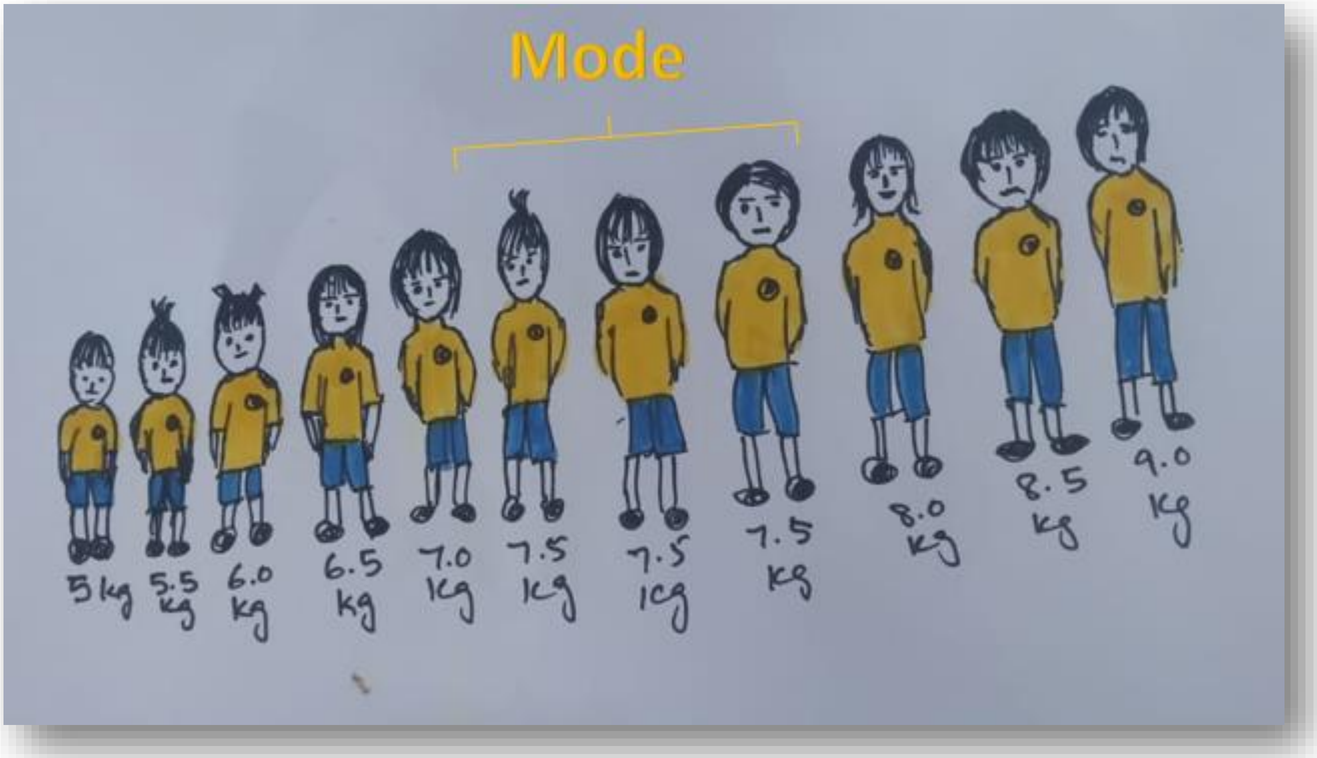


Figure 2.3. calculation of mode (most frequently occurring observation

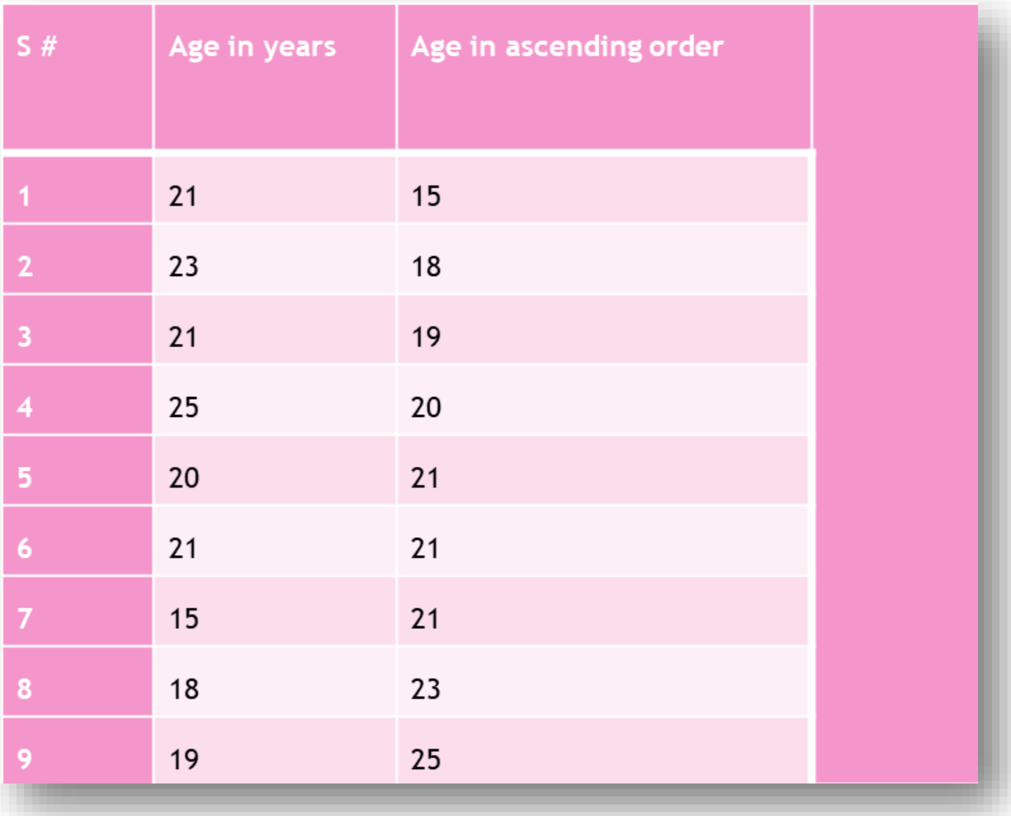
<u>Measures of dispersion</u>

Measures of dispersion are useful in determining the spread of a data set. In the case of continuous data, several measures of dispersion can be used. The most commonly used measures of dispersion include:

1. Range
2. Interquartile range
3. Variance
4. Standard deviation
5. Confidence interval

Range

One of the most common measures of dispersion is the range, which is simply the difference between the largest and smallest values in the data set. While the range is a quick and easy way to get a general sense of the spread of the data, it can be heavily influenced by outliers and is, therefore, not always the most reliable measure of dispersion. In such cases, other measures, such as the interquartile range or standard deviation, may provide a more accurate representation of the spread of the data. However, it is often used with measures of central tendency. An example of the range is given in Figure 2.4. A large range indicates that the data is more spread out or varied, while a small range indicates that the data is more tightly clustered around a specific value or range of values.



S #	Age in years	Age in ascending order
1	21	15
2	23	18
3	21	19
4	25	20
5	20	21
6	21	21
7	15	21
8	18	23
9	19	25

Figure 2.4. Range of a dataset (15-25)

Interquartile range

The interquartile range is a measure of spread or variability in a dataset that helps to identify the dispersion of the central 50% of the data. It is calculated as the difference between a dataset's third quartile (75th percentile- Quartile 3) and the first quartile (25th percentile- Quartile 1). The interquartile range is more robust to outliers than the range and provides a better sense of the spread of the central portion of the data.

First, data needs to be arranged in ascending order to calculate the interquartile range. Then, the median divides the dataset into two equal halves. - the lower half and the upper half. These two halves are further divided into two equal parts making four equal parts of the dataset (Figure 2.5). The first quartile (i.e. Q1) represents the 25th percentile, which is the value that separates the lowest 25% of the data from the highest 75%. To calculate Q1, you need to find the median value of the lower half of the dataset. The third quartile (i.e. Q3) represents the 75th percentile, which is the value that separates the lowest 75% of the data from the highest 25%. To calculate Q3, you need to find the median value of the upper half of the dataset. Finally, the interquartile range is the difference between Q3 and Q1. This range measures the spread of the data that is less sensitive to outliers.

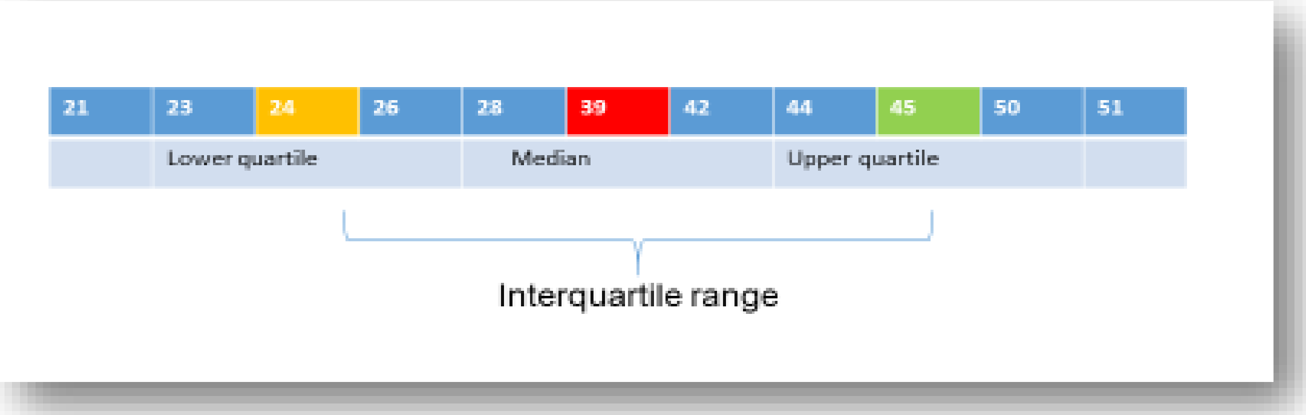


Figure 2.5. Interquartile range

Variance

Variance is a statistical concept that measures how a set of data is spread out. It measures how much the individual data points differ from the mean or average of the entire set. It is used in many fields, including finance, physics, and engineering. It is often used with standard deviation, the square root of variance.

The mean of the data set needs to be calculated to calculate variance. Then, for each data point, square the difference between that data point and the mean. Finally, these squared differences will be added up and divided by the total number of data points minus one.

The formula for calculating variance is:

variance = (sum of (value - mean)^2) / (number of values - 1)

Standard deviation

Standard deviation is a statistical measure used to quantify the variation or dispersion in a set of data. It is a commonly used measure of data spread, as it is easier to interpret than variance.

First, the mean of the dataset must be calculated to calculate the standard deviation. Then, the difference between each data point and the mean, and square these differences. The average of these squared differences is the variance. Finally, the square root of the variance will be calculated to get the standard deviation since standard deviation is useful because it gives an idea about the deviation of each data point from the mean. The data is widely spread and more diverse if the standard deviation is large. If the standard deviation is small, the data is clustered around the mean and less diverse.

The formula for calculating standard deviation is:

standard deviation = square root of [(sum of (value - mean)^2) / number of values]

Confidence interval

A confidence interval is a range of values used to estimate an unknown population parameter with a certain confidence level. It is typically represented as a range of values, with a lower and upper bound and a corresponding confidence level, often expressed as a percentage. For example, a 95% confidence interval for the mean weight of a certain population might be 150-170 lbs, which means that we are 95% confident that the true population mean weight falls within this range.

Confidence intervals are important in statistical analysis because they allow us to estimate population parameters based on sample data while accounting for the inherent uncertainty and variability in the data.

A 95% confidence interval is commonly used in statistics because it provides a range of values likely to include the true population parameter with a 95% confidence level. This means that if we repeated our sampling and estimation process many times, approximately 95% of the time, our confidence interval would contain the valid population parameter. It is a good balance between precision and reliability. If we reduce the confidence interval to 90%, we are willing to accept a more significant margin of error in our statistical analysis. This means that we allow a 10% chance that our results are inaccurate, which is higher than the standard 95% confidence interval.

On the other hand, increasing the confidence interval to 99% means we are increasing our level of certainty in the accuracy of our results. It also means a greater chance for the true population parameter to fall within the confidence interval. However, this comes at the cost of wider confidence intervals, which means a larger range of values could potentially be the true parameter. Overall, increasing the confidence interval to 99% is a trade-off between increasing our confidence in our results' accuracy and decreasing our estimates' precision.

Distribution of data

When we have continuous data, which can take on any value within a certain range, we often represent it using a statistical distribution. The distribution pattern helps decide the test to be applied for hypothesis testing. The most common pattern of naturally occurring data is a normal distribution. Other distributions for continuous data include the exponential distribution, the gamma distribution, and the beta distribution. However, for medical research, basic distribution concepts need to be understood. There are types of distribution patterns observed in continuous data:

a. Normal distribution

b. Skewed data

c. Kurtosis

Normal distribution

Normal distribution is a statistical term referring to the shape of a set of continuous data plotted on a graph. When a data set follows a normal distribution, the data is evenly distributed around the mean. This results in a bell-shaped curve. The normal distribution is important because many natural phenomena follow its shape, such as IQ scores or the heights of individuals in a population.

Parameters of a normal distribution include:

1. A symmetrical bell-shaped curve
2. Measures of central tendency(i.e. mean, median and mode) lie in the centre
3. Normal distribution has a particular internal distribution for the area under the curve:

a) $\mu\pm\sigma$ will always contain 68.26%

b) $\mu\pm2\sigma$ will always contain 95.44%

c) $\mu\pm3\sigma$ will always contain 99.73%

These known parameters make the normal distribution helpful for analyzing data and making predictions. A normal distribution is common in many natural phenomena, such as height, weight, and intelligence. It is also used in many applications, such as finance, biology, and engineering. However, in clinical research, data collected from a selected group, such as hospital setting data, is usually not normally distributed (Figure 2.6).

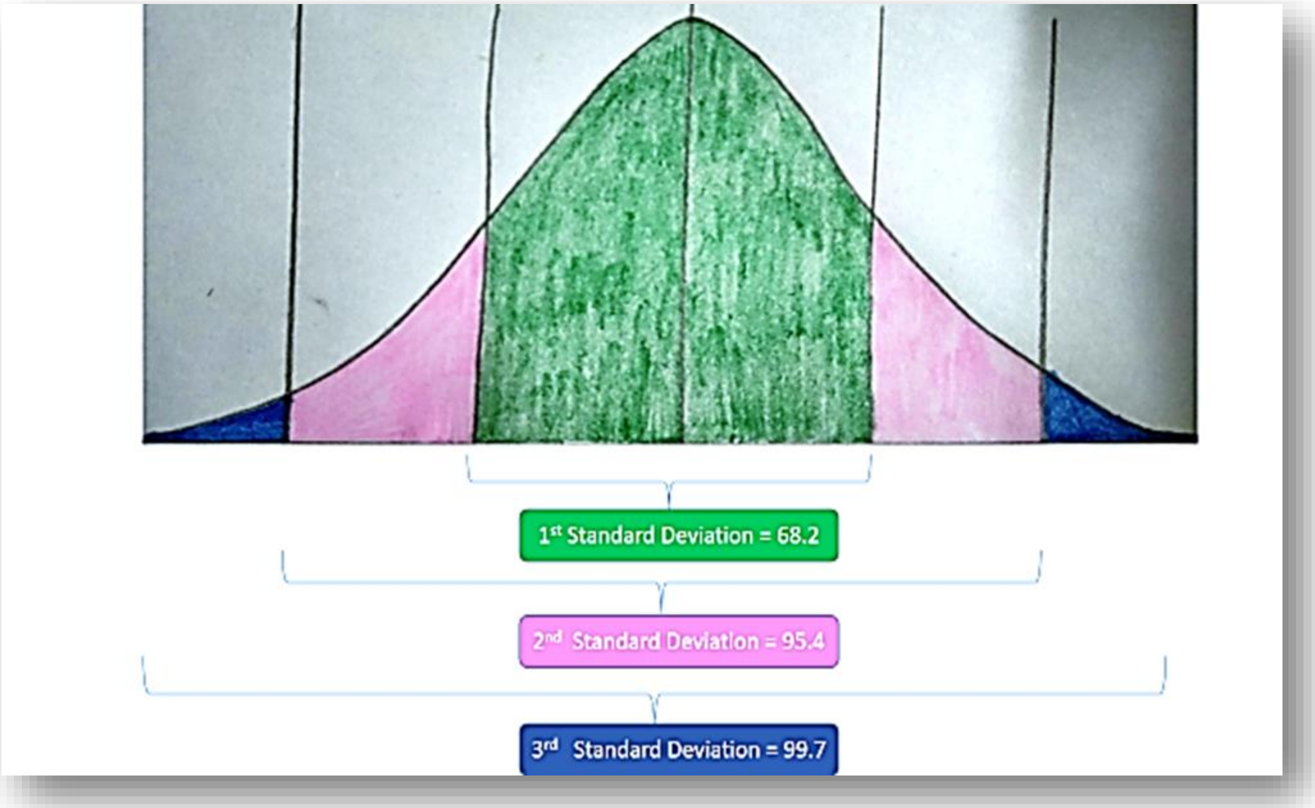

1st Standard Deviation = 68.2
2nd Standard Deviation = 95.4
3rd Standard Deviation = 99.7

Figure 2.6. Graphical presentation of a normal distribution

In order to determine if a set of data follows a normal distribution, statisticians use various tests and measures, such as the Shapiro-Wilk test and the Kolmogorov-Smirnov test. These tests help to determine if the data is normally distributed, and, if not, what type of distribution it may follow.

It is essential to remember that if the data is following normal distribution pattern, all parametric tests will be applied for hypothesis testing; if it is not, then non-parametric tests will be applied.

<u>Skewed data</u>

Skewness refers to the degree of asymmetry in a distribution of data. A perfectly symmetrical distribution has a skewness of zero. In contrast, a distribution with a longer tail on one side than the other will have a positive or negative skew depending on which side the tail is longer (Figure 2.7). Skewness can be an important factor when analyzing data, as it can affect the accuracy of measures such as the mean and standard deviation.

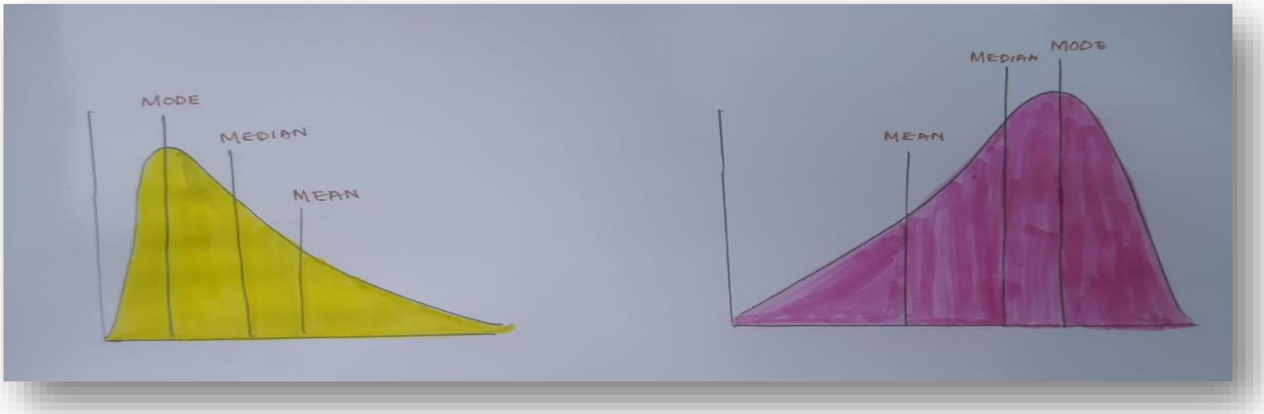


Figure 2.7. Skewed data – Yellow presents positively skewed data and pink presents negatively skewed data

Kurtosis

Kurtosis measures the degree of peakedness or flatness of a distribution of data around its mean. It tells us how much of the variability in the data is due to extreme values. A positive kurtosis value means that the data is more peaked than a normal distribution, and there are more data points in the tails. A negative kurtosis value indicates a flatter distribution with fewer extreme values. It's important to understand the kurtosis of data because it can affect statistical analyses such as hypothesis testing and regression modelling Figure 2.7 explains the pattern of Kurtosis.

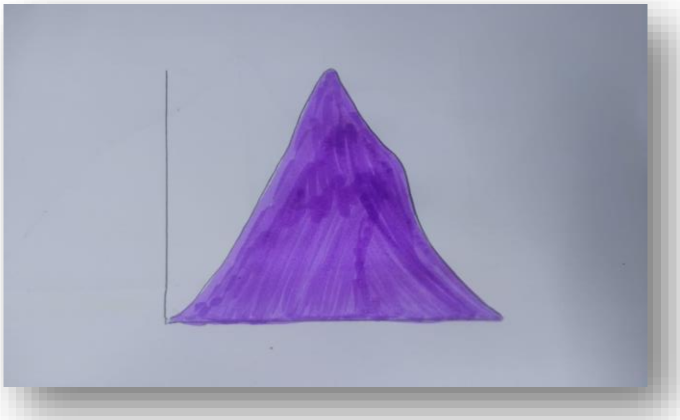


Figure 2.8. Presents Kurtosis in a dataset

Chapter 3.
Understanding categorical data

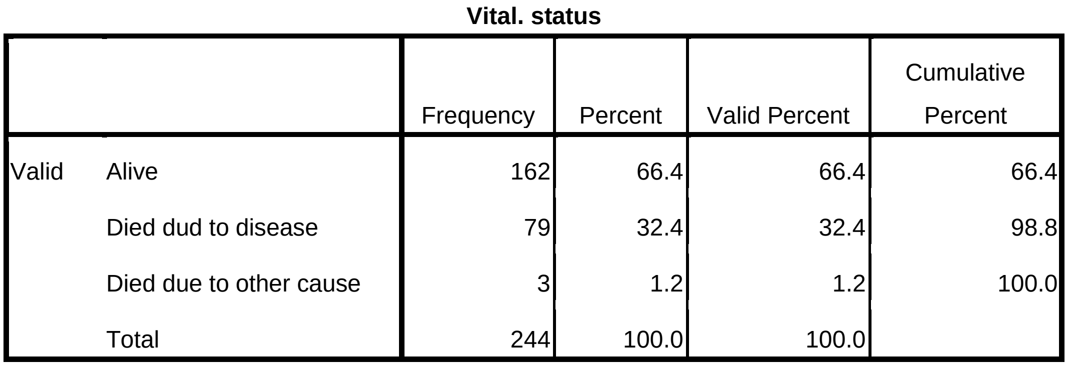
3. Understanding Categorical data

Categorical data is the type of data input with groups or categories pre-designed for entry. These are discrete numbers without decimals. There are several types of categorical data, including nominal, ordinal and binary data.

Nominal data is used to describe categories with no particular order, such as colours or types of fruit. All categories have equal weightage and status. Ordinal data, on the other hand, describes categories with a specific order or ranking, such as grades or socioeconomic status. Though sometimes the difference in the status of the categories is not clearly defined, there is a difference. Binary data refers to any data with only two possible outcomes, such as true or false. Dichotomous data is a special case of binary data that only has two mutually exclusive options, such as yes or no.

Analysis of the categorical variables

A frequency distribution is a table that displays the number of times each category in a categorical variable appears in a dataset. This type of table is very useful for summarizing categorical data and getting a quick overview of the most common categories. Furthermore, charts and graphs can also be created that help visualize the distribution of categories more clearly (Figure 3.1 a and b).



Vital. status

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Alive	162	66.4	66.4	66.4
	Died dud to disease	79	32.4	32.4	98.8
	Died due to other cause	3	1.2	1.2	100.0
	Total	244	100.0	100.0	

Figure 3.1.a presents the frequency distribution table calculated in Statistical Package for Social Sciences (SPSS)

Analysis of categorical variables is relatively easy, especially when using popular software such as statistical package for social science (SPSS).

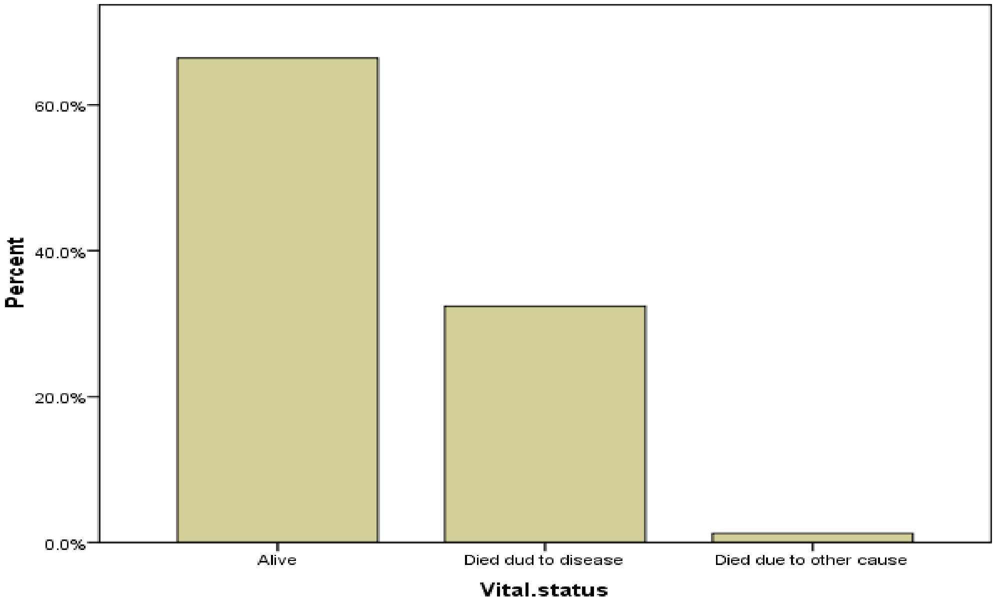


Figure 3.1.b. Bar chart of the same categorical variable presented in Figure 3.1.a (generated in Statistical package for social sciences – SPSS)

Chi square test

The chi-square test is a statistical tool used to measure the relationship between two categorical variables. It determines whether there is a significant difference between the observed and expected frequencies, assuming no relationship exists between the variables.

It is commonly used in social sciences, epidemiology, biology, and psychology to analyze data and test hypotheses. A chi-square test is a powerful tool that helps researchers analyze their data and make informed conclusions about the relationship between categorical variables.

Assumptions of the Chi-square test

The assumptions for a chi-square test depend on the type of test being conducted. However, some common assumptions include:

1. Independence: The observations in each category are independent of one another.

2. Sample size: The sample size is sufficiently large for the expected frequency counts in each cell to be greater than 5 (if it is <5, then Fischer exact test will be applied instead of chi-square).

3. Random sampling: The data is a simple random sample from the population of interest.

4. Expected frequencies: The expected frequencies are greater than 1 in all cells.

5. Non-negative values: All observed and expected frequencies are non-negative.

It's important to note that violating these assumptions may affect the validity of the chi-square test results.

Despite these limitations, the chi-square test is a widely used statistical tool that can provide valuable insights into relationships between categorical variables.

Interpretation of chi-square test

The interpretation of the chi-square test depends on the p-value, which measures the probability that the observed data could have occurred by chance. In general, if the p-value is less than 0.05 (driven from 95% confidence interval), then it is considered that the results are statistically significant, meaning that the observed differences are unlikely to have occurred by chance alone. If the p-value is less than 0.05, then it is essential to look at the effect size. A commonly used measure of effect size for chi-square tests is Cramer's V. Finally, a contingency table will be looked at, showing each category's counts for the variables of interest. This will then make it clear to understand which categories contribute to the differences resulting in significant p-value.

The formula for the chi-square test is:

$\chi^2 = \Sigma \text{ (Observed - Expected)}^2 / \text{Expected}$

Where: χ^2 is the chi-square statistic, Σ is the sum of all calculations, observed is the actual observed value and expected is the expected value (calculated under the null hypothesis)

Chapter 4.
Population and sample

4. Population and sample

The population is the set of possible observations based on a particular criterion. In the case of a study on Diabetes the population will be all diabetic patients around the globe. However, it is virtually impossible to study the entire set of population for a single research study. For example, if a task was given to measure the weight of fish in a fish form. For the study in an ideal situation, all the fish should be taken out of the pond and weighed, which is time-consuming and can potentially risk the fish's life too. So, to know the weight of the fish in a pond a few fish samples will be taken from different parts of the pond, and the average weight will be taken as the average weight of the fish in that pond.

The other practical example is measuring white blood cells of a patient admitted with infection. A straightforward method should be taking out all the blood, counting WBCs, and returning the blood to the body. This approach is practically impossible because this approach will cost the life of that patient. Therefore, we take a blood sample from a vein, generalize the findings of that 3cc blood, and generalise it for the entire body. The same principle applies to research projects where studying entire population can be extremely time-consuming and expensive. Thus a group from the defined population is selected and studied, called a ***sample***. The findings of a sample are generally considered and applied to the entire population. It must be remembered that the population or the reference population is far bigger and the sample is drawn from the population the researcher is interested in.

A sample should be appropriate in number and follow the entire population's essential characteristics to make it representative so that the findings may be generalized.

Population parameters

Population parameters are defined as the measure computed on a population called a population parameter. For example, the size of the population (denoted as N), the population mean (denoted as μ), variability or dispersion of the population is given as population range or variance (denoted as δ2). Population mean, and standard deviation are the most commonly used parameters that give the population's standard dispersion from its mean.

Why do we sample?

Studying an entire population is time-consuming and expensive; studying an entire population is not practically doable. Thus a sample is studied for the following advantages:

1. Cost-effectiveness: Studying an entire population can be time-consuming, resource-intensive, and expensive. Selecting a representative population sample can be a more cost-effective way to obtain data and make inferences about the entire population.

2. Practicality: In many cases, it may be impractical, if not impossible, to study an entire population. For example, it may be challenging to locate and survey every person with a rare medical condition across a large geographic area.

3. Feasibility: In some cases, studying the entire population may not be feasible due to ethical concerns or logistical limitations. For example, it would not be feasible to conduct a randomized controlled trial of a new cancer treatment on the entire population of cancer patients.

4. Precision: In some cases, studying an entire population may not be necessary to obtain precise estimates of a particular variable. A well-designed sample can provide precise estimates of the variable of interest with a sufficient confidence level.

Sampling

Sampling is the process of selecting a subset of individuals or items from a larger population to make inferences about the characteristics of the population as a whole. While sampling can be a very useful tool, there are some limitations to keep in mind. The primary purpose of selecting a sample is to obtain data that can be generalized to the larger population. The sample must be carefully selected to ensure that it is representative of the population in terms of fundamental characteristics or variables of interest. This helps minimize the potential for bias and increases the study results' validity.

Sampling techniques

Sampling techniques are procedures to select a subset of individuals or samples from a larger population. There are two main types of sampling techniques:

a. Probability sampling

b. Non-probability sampling

The choice of sampling technique will depend on a number of factors, including the size of the population, the resources available, and the research question being asked. It is essential to carefully consider the most appropriate technique for a given study to obtain accurate and reliable results.

Probability sampling

Probability sampling involves selecting a sample so that each member of the population has an equal chance of being selected. This helps to ensure that the sample is representative of the entire population.

By using probability sampling techniques, researchers can increase the accuracy and reliability of their findings, helping to ensure that their conclusions are based on solid data.

Probability sampling techniques have many strengths, as given below:

1. Representative sample: Probability sampling techniques ensure that each member of the population has an equal chance of being selected, which helps to create a sample that is likely to be representative of the entire population.

2. Accurate estimates: Probability sampling techniques provide accurate estimates of the population parameters, such as mean and standard deviation.

3. Randomness: Probability sampling techniques use random sampling methods, which help to eliminate or at least minimize bias and ensure that each member of the population has an equal chance of being selected.

4. Statistical validity: Probability sampling techniques provide a high level of statistical validity, which means that the results obtained from the sample can be generalized to the entire population with a high degree of confidence.

Although probability sampling techniques, while widely used in research, and these mentioned strengths have some limitations. One of the main limitations is that it can be difficult to accurately represent the population being studied, particularly if the population is very diverse or difficult to access. Another limitation is that probability sampling techniques require a lot of time and resources to analyze and interpret the collected data properly. There is also a risk of sampling error, which can occur when the sample size is too small, or the sample is not truly representative of the studied population. However, despite these limitations, probability sampling techniques remain valuable for researchers, particularly in cases where a representative sample is necessary to draw meaningful conclusions.

Types of probability sampling technique

Some common probability sampling types include simple random, systematic, stratified, and cluster (Figure 4.1).

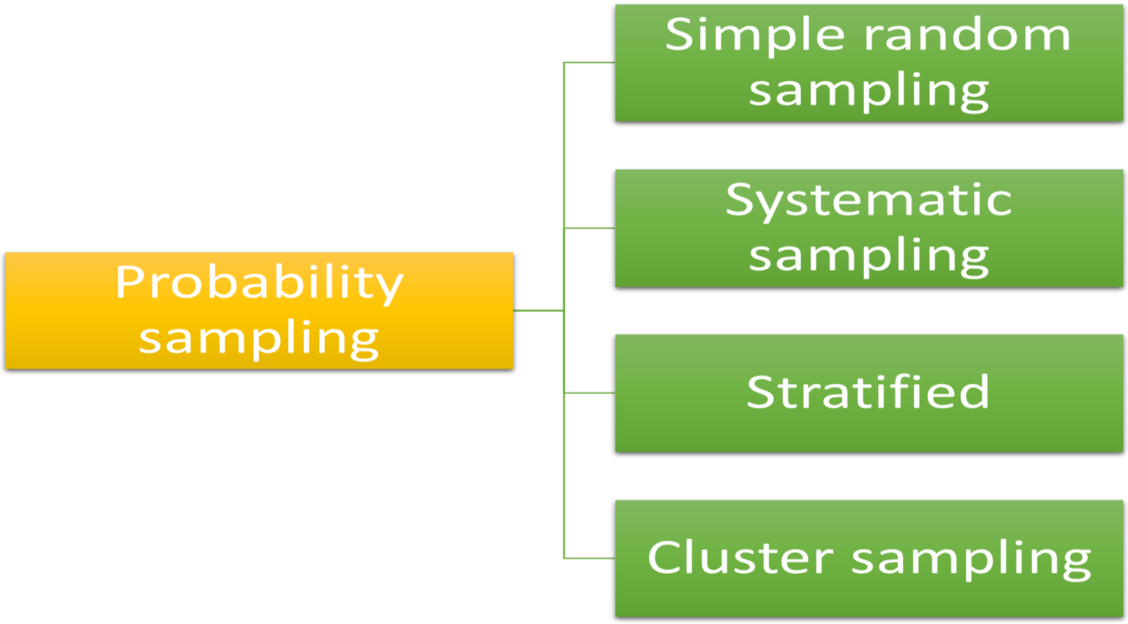


Figure 4.1. Types of probability sampling

Simple random sampling involves randomly selecting individuals from the population, with each individual having an equal chance of being selected. This helps to eliminate bias in the selection process. Simple random sampling is a popular statistical technique used for selecting a sample of individuals or items from a larger population. To carry out simple random sampling, defining the population intended to be stud is essential. Then, a randomization method be adopted to select a sample from that population. This could involve using a computer program to generate a list of random numbers, drawing names or items out of a hat, or using some other process to ensure that the sample is selected entirely at random.

Systematic sampling involves selecting individuals at fixed intervals. This can be useful, but obtaining a complete list of the population being studied is difficult. Systematic sampling is one of the most common research and statistical analysis methods. It involves selecting a sample of data from a larger population by selecting every n^{th} element in the population list.

For example, if you had a population list of 1000 individuals and wanted a sample of 100, you would select every 10th person. This method is useful because it ensures that the sample is representative of the population as long as the list is random and the sampling interval is chosen carefully.

Although it is a powerful sampling method giving an equal chance to all the population members in the order they are listed, it may not be feasible when the population is huge; also this selection at fixed intervals may not select a representative sample as the normal population do not follow sequence pattern but rather random pattern.

Stratified sampling involves dividing the population into subgroups, or strata, and then randomly selecting individuals from each group. This can help to ensure that the sample accurately reflects the characteristics of the population as a whole. The purpose of stratified sampling is to increase the representativeness of the sample by ensuring that each stratum is well-represented in the sample. This is especially important when there is significant variability in the population with respect to the characteristic in question. By dividing the population into strata based on this characteristic, it can be ensured that variability within each stratum has been captured, leading to more accurate estimates of population parameters. It is often used when the population is heterogeneous concerning a certain characteristic, such as age, income level, or education level.

Cluster sampling is a method that involves dividing a larger population into smaller groups, or clusters, and then randomly selecting some of those clusters to be included in the sample. Once the clusters have been selected, researchers can choose to include all of the selected clusters in the study or use another sampling method to select a smaller subset of individuals within each cluster. The cluster is actually representative of the entire population. One of the key advantages of cluster sampling is that it can be more cost-effective than other types of sampling methods, such as simple random sampling or stratified sampling. Since researchers only need to collect data from a subset of the population, they can save time and resources that would otherwise be required to survey or otherwise collect data from every member of the population. However, cluster sampling can also have some potential drawbacks. For example, if the selected clusters are not truly representative of the overall population, then the results of the study may be biased or inaccurate. Additionally, if there is significant variation within the selected clusters, this can lead to increased variability in the results of the study.

Non-probability sampling technique

Non-probability sampling is a technique of sampling that does not involve a random selection of participants. This sampling method selects individuals based on certain characteristics or criteria, such as availability, convenience, or subjective judgment. While non-probability sampling can be useful in some situations, it is important to note that it can introduce bias into research findings. The results obtained from non-probability samples may not represent the larger population or be generalizable. Non-probability sampling can offer several strengths, including flexibility in selection criteria and the ability to gather data quickly and inexpensively. It is also useful when studying hard-to-reach populations, such as those with rare characteristics or geographically dispersed. Finally, non-probability sampling can be used when a researcher is interested in exploring a specific phenomenon or wants to test a particular hypothesis without the need for a representative sample.

Types of non-probability sampling

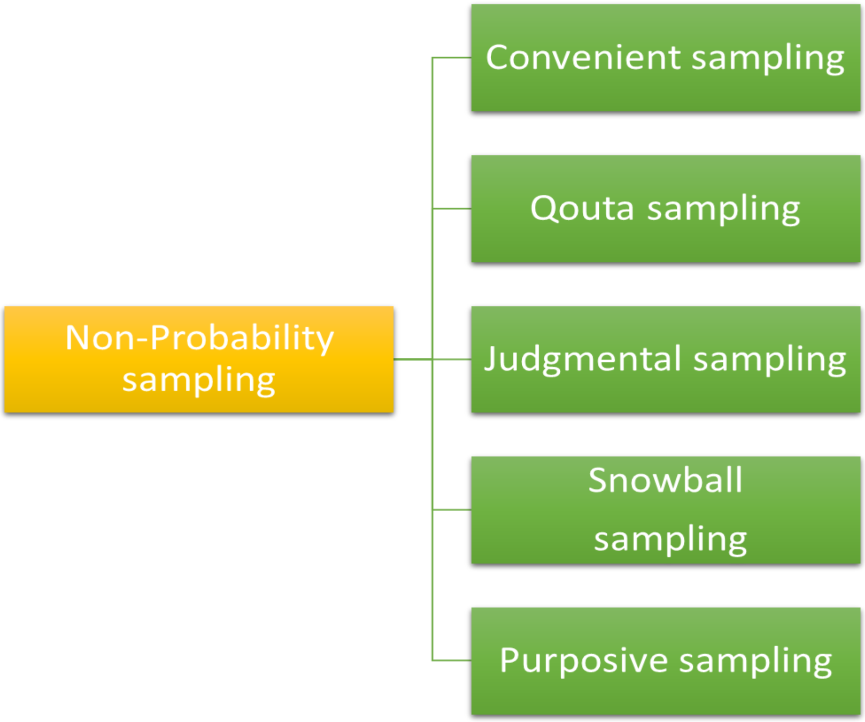


Figure 4.2. Types of non-probability sampling

There are several types of non-probability sampling techniques (Figure 4.2). Some of the most common ones are:

Convenience sampling involves selecting individuals who are readily available or easy to access. Convenient sampling is a type of sampling in which participants are chosen based on their accessibility and ease of participation. This sampling method is often used when the researcher has limited time, resources, or access to potential participants. While convenient sampling can be a quick and easy way to gather data, it can also introduce bias into the study. This is because participants may not represent the population as a whole, which can limit the generalizability of the results.

Quota sampling involves selecting a specific number of individuals from different categories to ensure a diverse sample. It is a method of selecting participants for a study that involves setting quotas or targets for specific characteristics or demographics, such as age, gender, or socioeconomic status, to ensure that the sample represents the population being studied. This method is cost-effective, easy to implement, and there is relatively less risk of bias. There are certain limitations that this method may not be representative of the entire population as the research put some restrictions which might not be appropriate according to the pattern of the reference population.

Judgmental sampling involves selecting individuals based on the judgment of the researcher or someone with expertise in the field. It is a non-probability sampling technique used in research where the researcher selects samples based on their knowledge and judgment of the population being studied. This technique is often used when it is difficult to obtain a representative sample or when the research objective requires a specific sample characteristic.

In judgmental sampling, the researcher selects individuals or units they believe represent the population being studied. This can be done through various methods such as purposive sampling, quota sampling or snowball sampling. The researcher's judgment plays a crucial role in ensuring that the sample is a valid representation of the population being studied.

While judgmental sampling is quick and cost-effective, it has its limitations. The researcher's biases and beliefs can influence the selection process, leading to a non-representative sample. Therefore, it is essential to use judgmental sampling only when other sampling techniques are not feasible or when the research objective specifically requires this technique.

Snowball sampling involves selecting individuals who know others who fit the criteria for the study, creating a "snowball" effect. It is a technique used in research to identify and recruit participants who may be challenging to reach or locate. This method involves starting with a small group of individuals who fit the criteria for the study and then asking them to refer others they know who may also fit the criteria.

The process continues in a "snowball" fashion, with each new contact providing referrals to additional participants. This approach can be instrumental when studying hard-to-reach populations, such as those with rare conditions or stigmatized behaviours.

Purposive sampling: This involves selecting individuals who have a specific characteristic or experience that is of interest to the researcher. It is a type of non-random sampling in which individuals or objects are selected for inclusion in a sample based on specific criteria or characteristics that interest the researcher. This type of sampling is often used in qualitative research, where the focus is on gaining an in-depth understanding of a particular phenomenon or group of people.

The criteria for selecting participants in purposive sampling can vary depending on the research question and the goals of the study. Criteria might include age, gender, occupation, educational level, or prior experiences with a particular issue or topic.

Problems to be considered for sampling

1. Introduction of Bias (non-representative sample)

Sampling bias is a type of error that can occur in statistical analysis when a sample does not represent the population from which it is drawn. This can happen due to various reasons, including inadequate sample size, non-random sampling procedures, or a lack of diversity in the sample. There are many types of bias that can affect decision-making and analysis. Some common types of bias include confirmation bias, where we tend to interpret information in a way that confirms our existing beliefs or hypotheses; availability bias, where we rely too heavily on information that comes to mind easily rather than seeking out more complete or balanced perspectives; and selection bias, where we inadvertently choose or exclude certain data or subjects in a way that skews our results. Other types of bias include anchoring bias, where we rely too heavily on a single piece of information or reference point, and attribution bias, where we tend to attribute the behaviour of others to internal factors rather than external circumstances. It's important to be aware of these biases to avoid or correct them in our decision-making and analysis.

2. Inappropriate number of samples

Sample size calculations are crucial in designing a research study, as they help ensure that the study is adequately powered to detect meaningful effects. The sample size depends on various factors, including the desired level of statistical power, the expected effect size and variability, and the type of statistical analysis that will be performed.

For the calculation of the appropriate sample size for the study, a number of factors need to be considered such as the type of study design, the level of significance, the desired power, the expected effect size, and the anticipated variability in the data. Various statistical formulas and tools are available to help you calculate sample sizes for different types of studies.

When it comes to sample size in studies, it's important to ensure that it's appropriate for the research question being asked. If the sample size is too small, the study may not have enough statistical power to detect important effects or relationships, potentially leading to false negative results. On the other hand, if the sample size is too large, it can lead to statistical significance being achieved for very small effects that may not be practically meaningful.

In summary, inappropriate sample size can compromise the validity and reliability of study results, so it's essential to carefully consider the sample size needed to answer the research question being investigated.

Chapter 5.
Hypothesis testing

5. Hypothesis testing

A research hypothesis is a proposed explanation or prediction for a phenomenon that a researcher intends to investigate in a research project. In simple terms, ***it's a prediction about the relationship of one or more variables in a project***. It is the starting point for any research study, as it guides the formulation of research questions, the design of experiments, and the collection and interpretation of data. A good research hypothesis should be testable, specific, and relevant to the research problem. It should also be based on existing knowledge and supported by evidence. It must be borne in mind that a research hypothesis is not a final conclusion or a proven fact but rather a tentative idea that can be confirmed or refuted by empirical evidence.

Null hypothesis versus alternate hypothesis

Students often get confused with a null and an alternate hypothesis. A null hypothesis is a statement that suggests there is no significant difference between two groups or no significant relationship between two variables. It is often used in statistical hypothesis testing to compare data and determine whether there is a statistically significant effect. Essentially, it is the opposite of an alternative hypothesis, which suggests that there is a significant difference or relationship.

While an alternate hypothesis (also termed an alternative hypothesis) is a statement that suggests a relationship between two or more variables. This hypothesis is used in statistical inference, where it is used to test the validity of a null hypothesis. The null hypothesis assumes no relationship exists between the variables being studied, while the alternate hypothesis suggests a relationship. By testing the alternate hypothesis against the null hypothesis, researchers can determine whether the relationship between the variables is statistically significant.

Theory versus hypothesis versus statement

As mentioned above, a hypothesis is a proposed explanation for a phenomenon that has not been thoroughly tested. It is generally a tentative assumption subject to further examination and experimentation. Hypotheses are based on observations or prior knowledge and are often framed as statements that can be tested through experimentation or data analysis.

Conversely, a theory is a well-established explanation for a phenomenon that has been extensively tested through experimentations and observations. A large body of evidence supports theories and has been repeatedly confirmed by many independent researchers. They are often used to predict future observations or guide further research.

While a statement is a declarative sentence that expresses a fact, opinion or belief, statements may be true or false. Still, they don't necessarily have to be supported by evidence or testing. They can be subjective and influenced by personal perspectives.

Hypothesis testing

Hypothesis testing is a statistical technique used to determine whether a particular hypothesis about a population is likely to be true or false based on sample data. The process of hypothesis testing involves several steps. First, a null hypothesis is formed, representing the status quo or the assumption that there is no significant difference between two groups or variables. An alternative hypothesis is also formed, representing the opposite of the null hypothesis. Additionally, an appropriate significance level needs to be selected. Choosing a test statistic appropriate for the data and hypothesis is also required. Next, a sample is taken from the population, and the relevant statistics are performed to calculate a p-value to determine the level of evidence against the null hypothesis. These statistics are then compared to the expected values under the null hypothesis. If the difference between the sample statistics and the expected values is large enough, the null hypothesis is rejected, and the alternative hypothesis is accepted.

However, it's important to note that hypothesis testing is not foolproof. There is always a chance that we reject the null hypothesis even though it is true or fails to reject the null hypothesis even though it is false.

Here are some caveats of hypothesis testing:

1. False positives: Sometimes, the hypothesis test may lead to false positives, where it appears that there is a significant difference between the groups, but in reality, there isn't.

2. Sample size: The hypothesis test requires a sufficient sample size to achieve statistical significance. If the sample size is too small, the results may not be reliable.

3. Assumptions: Hypothesis testing relies on certain assumptions, such as normality and independence of the data. Violations of these assumptions can lead to incorrect results.

4. Type I and Type II errors: Hypothesis testing can result in Type I errors (false positives) and Type II errors (false negatives). These errors can occur due to factors such as the level of significance chosen and the test's statistical power.

5. Causation: Hypothesis testing can only establish correlation, not causation between variables. It is essential to be cautious when interpreting the results of hypothesis testing as causal relationships.

Types of errors in hypothesis testing

Hypothesis testing is crucial in medical research, as it helps confirm or reject the proposed hypothesis based on the available data. Essentially, it's a way to determine whether the results observed in a study are due to chance or if there is a real effect. As mentioned above it is not foolproof since we are drawing inferences from samples thus, there is always a chance of error. There are two types of errors always kept in mind while interpreting the results of hypothesis testing.

Type I error

It is also known as a false positive, which occurs when a statistical test mistakenly rejects a null hypothesis that is true. This means that the test concludes that there is a significant effect or difference between groups when there is no difference in reality. Type I errors can happen for various reasons, such as using an incorrect significance level or testing a large number of hypotheses without adjusting for multiple comparisons or the sample was not appropriate.

Type II error

It occurs when a hypothesis test incorrectly fails to reject a false null hypothesis. This means that the test concludes that there is no statistical significance, even though the difference actually exists. A number of factors, such as a small sample size, a weak effect size, or a high level of variability in the data can cause type II errors. Additionally, if the significance level is set too high, it can increase the likelihood of a Type II error.

Interpretation of hypothesis testing

The hypothesis testing is based on the p-value if it is less than the level of significance (alpha), we reject the null hypothesis. It's important to note that the results of hypothesis testing are always subject to some degree of uncertainty. This is because statistical tests can never provide absolute certainty but rather a level of confidence in the results. The confidence level, or significance level, is typically set at 0.05 or 0.01, meaning there is a 5% or 1% chance that the observed effect is due to chance.

One tailed or two tailed hypothesis testing

A one-tailed test is used in statistical hypothesis testing when the null hypothesis is tested only in one direction. This means that the alternative hypothesis is directional and predicts that there will be a change in a specific direction (either an increase or a decrease). For example, let's say we're testing whether a new drug improves cognitive function. The one-tailed null hypothesis would be that the drug does not affect cognitive function, while the alternative hypothesis would be that the drug improves cognitive function.

On the other hand, a two-tailed test is used when the null hypothesis is tested in both directions. This means that the alternative hypothesis is non-directional and predicts that there will be a change of some sort without specifying the direction of the change. For example, if we're testing whether there is a difference in height between men and women, the two-tailed null hypothesis would be that there is no difference in height, while the alternative hypothesis would be that there is a difference in height (without specifying whether men are taller or women are taller).

Thus, if the hypothesis is, directional one-tailed test results will be considered, while non-directional two-tailed test will be considered.

p-value

In statistics, the p-value measures the strength of evidence against a null hypothesis. It is a probability value that indicates how likely it is to obtain a result as extreme or more extreme than the one observed, assuming that the null hypothesis is true. A p-value is typically used in hypothesis testing and is often used to determine the statistical significance of a result. It is important to note that the p-value does not indicate the size or importance of the effect, only the strength of the evidence against the null hypothesis. Additionally, the p-value is influenced by sample size, so a larger sample size can result in a smaller p-value, even if the effect size is small.

The p-value is driven from a 95% confidence level with a 5% probability of a chance finding. 5% probability was then converted to 0.05, described as a p-value by Sir Ronald A. Fisher and later confirmed by Karl Pearson. The effect/relation in the variable has <5% probability of a chance finding, then the p-value comes up <0.05, and we accept the scientific hypothesis.

Chapter 6.
Probability

6. Probability

Probability is a branch of mathematics that deals with studying random events and their likelihood of occurrence. It measures the likelihood of an event occurring, expressed as a fraction or decimal between 0 and 1, with 0 indicating that the event will not occur and 1 indicating that the event is certain to occur. The basic concept of probability is based on the simple premise that all outcomes of a random event are equally likely. For example, when flipping a fair coin, the probability of getting a head or a tail is 1/2 or 0.5.

There are two types of probability:

1. Theoretical probability (it is based on a mathematical model or a theoretical calculation)
2. Empirical probability (it is based on observed data or experimental results)

Various tools and techniques, such as probability distributions, Bayes' theorem, and random variables, are used to model and compute probabilities.

Applications of probability theory

Probability theory is widely used in various fields, including science, engineering, finance, economics, and social sciences, to analyze and predict the likelihood of events. Important applications of probability theory include:

1. **Statistical inference**: Probability theory is used to make inferences about populations based on sample data.

2. **Risk analysis**: Probability theory assesses the likelihood of risk events and develops risk management strategies.

3. **Decision-making:** Probability theory is used to evaluate the likelihood of different outcomes and to make informed decisions based on the available information.

4. **Gaming and gambling:** Probability theory calculates the odds of winning or losing in games of chance and gambling.

Probability theory has numerous applications in medical research. Some of the uses of probability in medical research include:

1. **Clinical trials:** Probability theory is used in the design and analysis of clinical trials to determine the results' sample size, power, and statistical significance.
2. **Disease diagnosis and prognosis:** Probability theory is used to develop predictive models for the diagnosis and prognosis of diseases. For example, probabilistic

models can predict the likelihood of a patient developing a particular disease based on age, gender, family history, and other risk factors.

3. **Risk assessment:** Probability theory is used to assess the likelihood of developing a disease or a health condition based on various risk factors. For example, probabilistic models can estimate the risk of developing cardiovascular disease based on factors such as age, gender, blood pressure, and cholesterol levels.
4. **Genetics:** Probability theory is used to analyze and interpret genetic data, such as DNA sequencing data, and to identify genetic risk factors for diseases.
5. **Survival analysis:** Probability theory is used to model and analyze survival data, such as time-to-event data, to estimate survival probabilities and study the effect of different factors on survival.

In summary, probability theory is a crucial tool in medical research that allows researchers to quantify and analyze the likelihood of different health outcomes and to develop models and strategies for disease prevention, diagnosis, and treatment.

Chapter 7.
Correlational analysis

7. Correlational analysis

The correlational analysis is a statistical method to investigate the relationship between two or more continuous variables in a dataset. It measures the strength and direction of the association between two variables and can be used to identify patterns and trends in the data. The strength of the relationship between the variables is measured by the correlation coefficient, which ranges from -1 to +1. A coefficient of -1 indicates a perfect negative correlation, a coefficient of +1 indicates a perfect positive correlation, and a 0 indicates no correlation between the variables.

Correlational analysis can be used for a variety of purposes, including:

1. **Identifying relationships between variables:** Correlational analysis can identify relationships between variables that may not be immediately apparent. For example, a researcher may use correlational analysis to determine whether there is a relationship between smoking and lung cancer.

2. **Predicting future outcomes:** Correlational analysis can predict future outcomes based on the relationship between variables. For example, a researcher may use correlational analysis to predict the likelihood of a patient developing heart disease based on their age, gender, blood pressure, and cholesterol levels.

3. **Identifying confounding variables:** Correlational analysis can be used to identify confounding variables that may affect the relationship between two variables. For example, a researcher may use correlational analysis to determine whether there is a relationship between exercise and weight loss, while controlling for factors such as diet and genetics.

Scatter plots for determination of correlation

Scatter plots are graphs used to display the relationship between two continuous variables. They are used to visually inspect the relationship between two variables and can be used to determine the strength and direction of the correlation between the variables. Scatter plots consist of a horizontal x-axis and a vertical y-axis, with each point on the graph representing the value of the two variables for a single observation. A positive correlation is when two variables increase together. This is shown by a scatter plot with a positive slope, where the points on the graph move from the lower left corner to the upper right corner. For example, a person's height and weight may have a positive correlation.

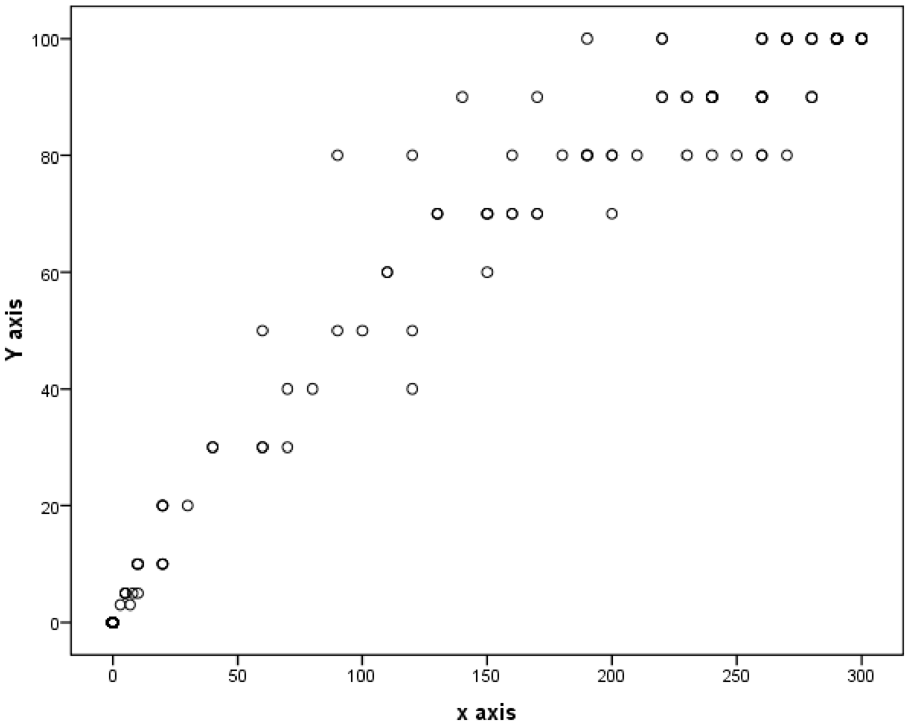


Figure 7.1. presenting scatter plot showing a positive correlation

Scatter plots can also be used to identify outliers, which are observations that do not follow the general pattern of the data. Outliers can significantly affect the correlation between the variables and may need to be removed or investigated further.

Parameters of measuring correlation:

1. Covariance

Covariance is a statistical measure that describes the relationship between two variables. It measures the degree of the change in one variable in relation to the change in second variable. More specifically, covariance measures the degree to which one variable's values correspond with another variable's values.

Covariance can be used to measure the strength and direction of the relationship between two variables. If the covariance is positive, the variables tend to increase or decrease together. If the covariance is negative, the variables tend to move in opposite directions. If the covariance is zero, it indicates no relationship between the variables.

Covariance has some limitations. One of the limitations is that it is not standardized, which means that its value depends on the units of the variables. Another limitation is that it does not consider the variables' scale, which can make it difficult to compare covariances between different sets of data. The unit of covariance is the product of the units of the two variables being measured. Since covariance measures how much two variables change together, the covariance unit reflects the two variables' units. For example, if the two variables being measured are weight (in kilograms) and height (in centimetres), then the unit of covariance would be (kilograms x centimetres). This means that the covariance would be expressed in units such as kilogram-centimetres, which is not a commonly used unit of measurement.

Due to the problem of mixed units, comparing covariances between different data sets can be challenging. As a result, researchers often use a standardized measure of the strength of the relationship between two variables, known as the ***correlation coefficient.***

Correlation coefficient

The correlation coefficient is a statistical measure that describes the strength and direction of the linear relationship between two continuous variables. The correlation coefficient itself is a dimensionless measure that ranges from -1 to +1, making it easier to compare between different datasets, where a coefficient of -1 indicates a perfect negative correlation, a coefficient of +1 indicates a perfect positive correlation, and a coefficient of 0 indicating no correlation between the variables.

Type of correlation coefficient

Several types of correlation coefficients are commonly used in statistical analysis, including:

1. **Pearson correlation coefficient:** This is the most widely used correlation coefficient to measure the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges from -1 to +1. A correlation coefficient of -1 indicates a perfect negative correlation, a coefficient of +1 indicates a perfect positive correlation, and a 0 indicates no correlation between the variables.

2. **Spearman correlation coefficient:** This correlation coefficient measures the strength and direction of the relationship between two variables when the data is not normally distributed or there are outliers present. It is based on the rank order of the data and is denoted by the symbol "r". The Spearman correlation coefficient ranges from -1 to +1, with a coefficient of -1 indicating a perfect negative correlation, a coefficient of +1 indicating a perfect positive correlation, and a coefficient of 0 indicating no correlation between the variables.

3. **Kendall correlation coefficient:** This is another rank-based correlation coefficient used to measure the strength and direction of the relationship between two variables. It is denoted by the symbol "τ". It ranges from -1 to +1, with a coefficient of -1 indicating a perfect negative correlation, a coefficient of +1 indicating a perfect positive correlation, and a coefficient of 0 indicating no correlation between the variables.

4. **Point-biserial correlation coefficient:** This correlation coefficient measures the strength and direction of the relationship between a continuous and binary variable. It is denoted by the symbol "rpb". It ranges from -1 to +1, with a coefficient of -1 indicating a perfect negative correlation, a coefficient of +1 indicating a perfect positive correlation, and a coefficient of 0 indicating no correlation between the variables.

5. **Phi coefficient:** This correlation coefficient measures the strength and direction of the relationship between two binary variables. It is denoted by the symbol "φ". It ranges from -1 to +1, with a coefficient of -1 indicating a perfect negative correlation, a coefficient of +1 indicating a perfect positive correlation, and a coefficient of 0 indicating no correlation between the variables.

Partial versus semi-partial correlation

1. Partial correlation: It is a statistical technique used to measure the relationship between two variables while controlling for the effects of one or more additional variables. It is used to examine the relationship between two variables while holding other variables constant, thereby allowing researchers to determine whether the relationship between the two variables is robust and independent of other variables. Partial correlation is typically used when a third variable is known or suspected to influence the relationship between the two variables of interest. By controlling for this third variable, researchers can determine whether the relationship between the two variables of interest is significant and meaningful.

Partial correlation is used in a variety of fields, including psychology, economics, and social sciences. For example, in a study examining the relationship between exercise and weight loss, researchers may want to control for factors such as age, gender, and diet. By controlling for these variables, the researchers can determine whether the relationship between exercise and weight loss is robust and independent of other variables.

Two main types of partial correlation exist standard partial correlation and multiple partial correlation. These are used to analyze the relationship between two variables while controlling for the effect of one or more additional variables.

1. **Standard Partial Correlation:** This type of partial correlation is used to examine the relationship between two variables while controlling for the effect of one-third

variable. It involves calculating the correlation between two variables while holding a third variable constant.

2. **Multiple Partial Correlation:** This type of partial correlation is used to examine the relationship between two variables while controlling for the effects of multiple additional variables. It involves calculating the correlation between two variables while holding two or more other variables constant.

Semi-partial correlation

Semi-partial correlation, also known as the part correlation, is a statistical technique used to measure the unique relationship between two variables while controlling for the effects of a third variable. It is similar to partial correlation but differs in that semi-partial correlation controls for the effects of only one variable, while partial correlation controls for the effects of multiple variables.

Semi-partial correlation is calculated by examining the correlation between two variables holding a third variable constant and removing the shared variance between the third and dependent variables from the calculation. The resulting correlation coefficient reflects the unique relationship between the two variables that is independent of the third variable.

Semi-partial correlation is useful when a researcher wants to determine the unique relationship between two variables while controlling for the effects of a third variable. For example, a researcher may want to examine the relationship between employee productivity and job satisfaction while controlling for the effects of the number of hours worked per week. By using semi-partial correlation, the researcher can determine the unique relationship between productivity and job satisfaction independent of the number of hours worked.

Chapter 8.
Multivariate analysis

8. Multivariate analysis

Multivariate analysis is a statistical technique used to analyze data that involves multiple variables. It involves examining the relationship between multiple variables simultaneously, allowing researchers to identify patterns, trends, and associations that may not be apparent when examining variables individually.

There are several types of multivariate analysis techniques that are commonly used in research, including:

1. **Factor analysis:** This technique is used to identify underlying factors that contribute to a set of observed variables. It is used to simplify complex data sets and identify the underlying structure of the data.
2. **Cluster analysis:** This technique is used to group similar observations or cases together based on their characteristics or attributes. It is used to identify patterns or groups within a data set.
3. **Discriminant analysis:** This technique predicts group membership based on a set of predictor variables. It is used to identify the most useful variables for predicting group membership.
4. **Structural equation modelling**: This technique is used to examine complex relationships between multiple variables. It is used to test hypotheses about causal relationships and identify a data set's underlying structure.
5. **Canonical correlation analysis:** This technique is used to examine the relationship between two sets of variables. It is used to identify the variables that are most strongly associated with one another.

Multivariate analysis is used in many fields, including psychology, economics, and social sciences. It is particularly useful in situations where multiple variables may influence an outcome or variable of interest. By analyzing these variables simultaneously, researchers can identify the unique contribution of each variable and how they interact with one another.

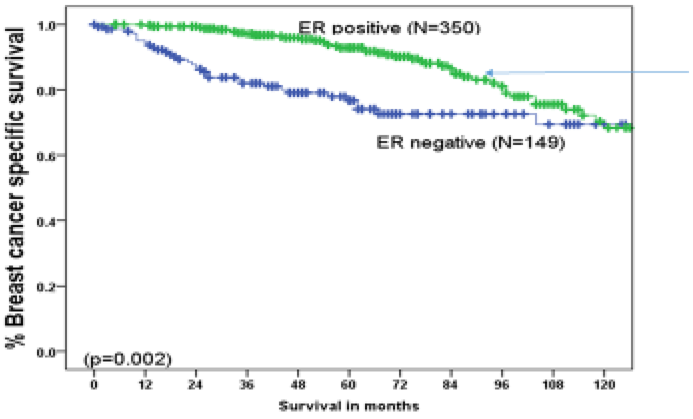
However, it is required that the individual standing of each variable may be evaluated before starting the multivariate analysis. In medical research, multiple risk factors or prognostic factors can be evaluated together to see which one is the independent and strongest factor.

Chapter 9.
Survival analysis

9. Survival analysis

Survival analysis is a statistical method used to analyze the time it takes for an event of interest to occur, such as the onset of a disease, death, or day of recovery. It is used to estimate the probability of an event occurring over time, and to identify factors that may influence the timing of that event. The survival analysis is a useful tool for cohort studies and clinical trials, where outcome variables of interest correlate with different factors.

Survival analysis is based on the concept of "survival function," which is a probability function that describes the proportion of individuals who have not experienced the event of interest over time. The survival function is often represented graphically as a survival curve, which shows the probability of survival over time. The Kaplan-Meier method is useful for estimating survival probabilities when data is censored, meaning that the event of interest has not occurred for all subjects in the study. Censoring can occur when subjects are lost to follow-up, withdraw from the study, or when the study ends before all subjects have experienced the event of interest. The Kaplan-Meier method considers censoring when calculating the survival probability over time.



Survival curve showing vertical line of alive patients and curve dropping with each event.

P-value suggests significant difference in the status of factor of interest

Survival analysis is instrumental in medical research, where it is used to study the survival rates of patients with different diseases or conditions. It can also be used in other fields, such as economics, engineering, and social sciences, to study the time it takes for an event to occur or the duration of an event.

There are several methods used in survival analysis, including:

1. **Kaplan-Meier method:** This is a non-parametric method used to estimate the survival function or happening of the event of interest (outcome variable) without any assumptions about the distribution of survival times. It is used in survival analysis to calculate the probability of an event, such as death, discharge from hospital, etc., based on data collected over time. The method assumes no prior knowledge of the distribution of the data.

The Kaplan-Meier method involves the following steps:
Data collection: Data is collected on the time it takes for an event of interest, such as the onset of a disease or death. The data is typically collected at fixed intervals, such as daily, weekly, or monthly. For this, as a general rule, daily or weekly data may be useful in a situation of short-term duration such as viral diseases. However, for chronic illnesses such as cancer, years of follow-up are required and calculated accordingly.
Identifying subjects at risk: For each interval, the subjects at risk of experiencing the event of interest are identified. This includes subjects who have not yet experienced the event and have not been lost to follow-up. As shown in Figure 9.1 ER is a factor of interest and a comparison was made between the positive and negative groups, and the results suggest that the presence of ER reduces the risk of the event of interest.

Uses of Survival analysis

Calculating the probability of survival: The probability of survival is calculated for each interval by dividing the number of subjects who have not yet experienced the event by the total number of subjects at risk.
Estimating the survival function: The survival function is estimated by multiplying the probabilities of survival for each interval. This gives an estimate of the probability of survival over time.
Plotting the survival curve: The estimated survival function is plotted on a graph, with time on the X-axis and probability of survival on the Y-axis. The survival curve shows the probability of survival over time (Figure 9.1).

2. **Cox proportional hazards model:** This is a semiparametric method used to estimate the effects of multiple covariates on survival time. It assumes that the hazard rate is proportional over time. The Cox proportional hazards model is a statistical method for analysing survival data in medical research and other fields. It is a type of regression model used to estimate the effect of covariates on survival time while assuming that the hazard rate (the probability of an event occurring at a given time) is proportional.
Similar to Kaplan Meier it makes no assumptions about the distribution of the survival time data. Instead, it assumes that the hazard rate is proportional for all subjects in the study. This means the hazard rate for any two subjects is proportional at any given time, regardless of the baseline hazard rate.

The Cox proportional hazards model involves the following steps:
Data collection: Data on survival time and covariates is collected for each subject in the study.
Model fitting: The Cox proportional hazards model is used to fit the data, estimating the effects of covariates on the hazard rate.

Model assessment: The model is assessed to determine how well it fits the data. This can be done using various statistical tests, such as the log-rank or likelihood ratio tests.
Model interpretation: The model results are interpreted to determine the effect of each covariate on the hazard rate. This can be done by examining the hazard ratio for each covariate, which is the ratio of the hazard rates between two groups of subjects with different covariate levels.

The Cox proportional hazards model is particularly useful when multiple covariates may influence survival time. By estimating the effects of each covariate while assuming proportional hazards over time, the model can identify which covariates are most strongly associated with survival time.

One limitation of the Cox proportional hazards model is that it assumes the hazard rate is proportional over time. This may not be true in all cases, and violations of this assumption can lead to biased estimates of the effects of covariates on survival time.

Overall, the Cox proportional hazards model is a powerful statistical tool used in survival analysis to estimate the effect of covariates on survival time while assuming proportional hazards over time. It is widely used in medical research and other fields where survival data is collected.

3. **Accelerated failure time model:** This parametric method assumes a specific distribution for the survival times, such as the Weibull distribution or the log-normal distribution. This model is also used to estimate the effect of covariates on the time to an event of interest, such as the onset of a disease or death.

The AFT model assumes that the logarithm of the survival time follows a linear model with covariates. The AFT model is particularly useful when the survival times are not proportional or when the proportional hazards assumption of the Cox proportional hazards model is violated. Additionally, the AFT model allows for direct interpretation of the coefficients, as the coefficients represent the logarithm of the survival time for a one-unit increase in the covariate.

One limitation of the AFT model is that it assumes a specific distribution for survival times. If the assumed distribution is not appropriate for the data, the estimates may be biased. Additionally, the AFT model is sensitive to outliers in the data. Therefore, in case of the presence of outliers in the data AFT results will be misleading.

Overall, the AFT model is a powerful statistical tool used in survival analysis to estimate the effect of covariates on survival time while assuming a specific distribution for survival times. It is useful when the proportional hazards assumption of the Cox proportional hazards model is violated or when the survival times are not proportional.

4. **Competing risks analysis:** This method is used to analyze situations with multiple possible outcomes, such as death from different causes. Competing risks analysis is a statistical method used to analyze survival data when there are multiple possible outcomes, each of which is a competing risk for the event of interest. For example, in medical research, a patient may be at risk of dying from a primary disease, a secondary disease, or an unrelated cause of death. Competing risks analysis allows researchers to estimate the probability of each type of event occurring and to identify factors that may influence the timing and likelihood of each event.

Competing risks analysis involves several steps:

Define the competing risks: The competing risks must be defined and categorized. This can involve identifying different types of events that are competing with each other, such as different causes of death (ie. In chronic illnesses like cancer a patient may die of disease or some other illness of accident) or different treatment outcomes such as complete recovery, death, recurrence etc.

Collect data: The data on the event of interest and the competing risks are collected for each subject in the study.

Estimation of cause-specific hazard function: The cause-specific hazard function is estimated for each competing risk. The hazard function estimates the instantaneous rate at which each event occurs over time.

Calculation of cumulative incidence function: The cumulative incidence function is calculated for each competing risk. The cumulative incidence function estimates the probability of the event occurring over time, taking into account the other competing risks.

Model fitting: The factors that influence the timing and likelihood of each event are identified using statistical models. The models can be used to estimate the effects of covariates on the cause-specific hazard function or the cumulative incidence function.

One limitation of competing risks analysis is that it assumes that the competing events are independent. This may not always be the case, and violations of this assumption can lead to biased estimates.

Overall, competing risks analysis is a powerful statistical tool used in survival analysis to analyze data when there are multiple possible outcomes. It allows researchers to estimate the probability of each type of event occurring and to identify factors that may influence each event. Competing risks analysis is particularly useful in medical research, but it can also be applied in other fields where there are multiple competing events that can occur.

LUMHS Publishers
Liaquat University of Medical & Health Sciences,
Jamshoro, Pakistan
www.lumhs.edu.pk/publishers

ISBN:978-627-7526-00-9