BIOSTATISTICS

Second Edition

Authors

Prof. Dr. Farkhanda Manzoor
Dr. Najiya al-Arifa
Usman Bashir

Applied Sciences Research Centre

Biostatistics

A PRACTICAL GUIDE TO STATISTICAL TOOLS

AUTHORS

Prof. Dr. Farkhanda Manzoor
Minhaj University Lahore, Lahore

Dr. Najiya al-Arifa
Lahore College for Women University, Lahore

Usman Bashir
Minhaj University Lahore, Lahore

2nd Edition

Published by
FJ Publishers

PREFACE

Bio-statistical manuals guide students of natural sciences, medical sciences and social sciences to apply statistical methods in their studies. This book contains step by step guide to statistical methods in scientific studies through statistical tools like IBM-SPSS. The use of statistical techniques has increased in biological sciences for analyzing the quantitative and qualitative data. In this manual, the basic concepts of biostatistics have been illustrated along with mathematical interpretation of the scientific problems. Examples have been presented with respective data to address statistical problems. The manual provides guidelines to University students to use IBM-SPSS statistical tool for the analysis of the data. The main goal of this manual is to teach biology students which statistical methods are appropriate for their scientific study and how the statistical test is applied using various types of software. Datasheets have been provided along with solution keys to apply appropriate statistical techniques. This manual consist of 7 chapters which provide important techniques for analysis of scientific data that covers the basic concepts of bio-statistics, types of data presentation and how data is tabulated and expressed. Sampling methods described in this book are used for estimation of sample size for population analysis. Measure of central tendencies has been illustrated in this book to give basic concepts of mean, mode, median, etc. The idea of probability and its distribution has been provided in this book with simplified algebraic equations. Analysis of the significance of variance has been described through one-way and two-way ANOVA. This statistical methodology is most commonly applied in scientific researches to check the level of significance. Study of relation between dependent and independent variable regression and correlation are explained. At the end of every topic/chapter the solution keys have been provided to apply a particular statistical test. The data is explained in tabulated form to help students understand how data is presented. A clear demonstration of statistical tool IBM-SPSS has been provided at end of every chapter along with screen shots for the guidance of University students to use statistical software for easy analysis of biological data.

AUTHORS

CONTENTS

CHAPTER 5: PROBABILITY 65

CHAPTER 6: PROCEDURES FOR HYPOTHESIS TESTING 81

CHAPTER 7: ANALYSIS OF VARIANCE (ANOVA) 105

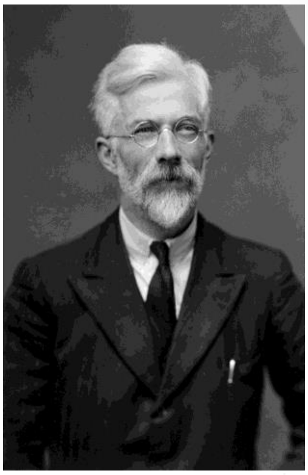
CHAPTER ONE

Introduction to Biostatistics

1.1. INTRODUCTION

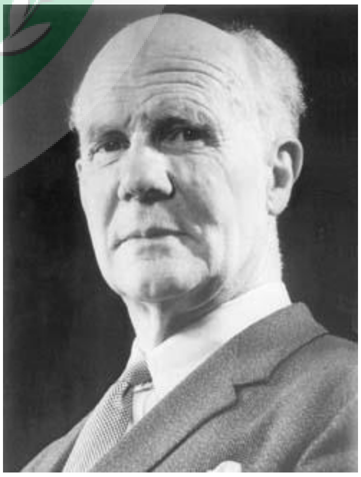
The term statistics is derived from Latin word “Status” or from Italian word “Statist”, both words mean “political state”. **Statistics** is the science which deals with collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of a phenomenon. It is the branch of mathematical sciences which deals with the study and use of the methodologies for the analysis of data. These are the scientific mathematical methods for data collection, presentation, analysis and drawing inferences from data set.

Three scientists, Ronald A. Fisher, Jersey Neyman and Egon S. Pearson introduced the modern concepts of statistics.



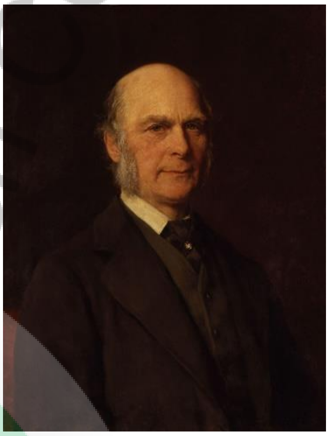
Ronald A. Fisher
(1890-1962)

Jersey Neyman
(1894-1981)



Egon S. Pearson
(1895-1980)

Biostatistics is the branch of biological sciences which deals with the study of statistical methods applied to solve the biological problems. Biostatistics is the collection, compilation, analysis and inferences of the biological data. Basically, it is the study of applying statistical methods to the biological data. While working on a research problem it is necessary to understand and apply statistical methods on the data set because it helps the researcher in understanding the nature of variability and derive general laws from the samples.



Sir Francis Galton is called the “Father of Biostatistics” who introduced the term “correlation”. He was the first person to apply statistical methods for studying patterns of inheritance in populations. He created questionnaires and designed surveys for data collection of human communities. While working on human population he introduced the statistical tools for the first time to study differences in human population. Galton also invented the term eugenics in 1883.

Biometry means “biological measurement” and sometime it is also called biostatistics which is the statistical analysis of the biological data. The term was first used by William Whewell in 1800s.

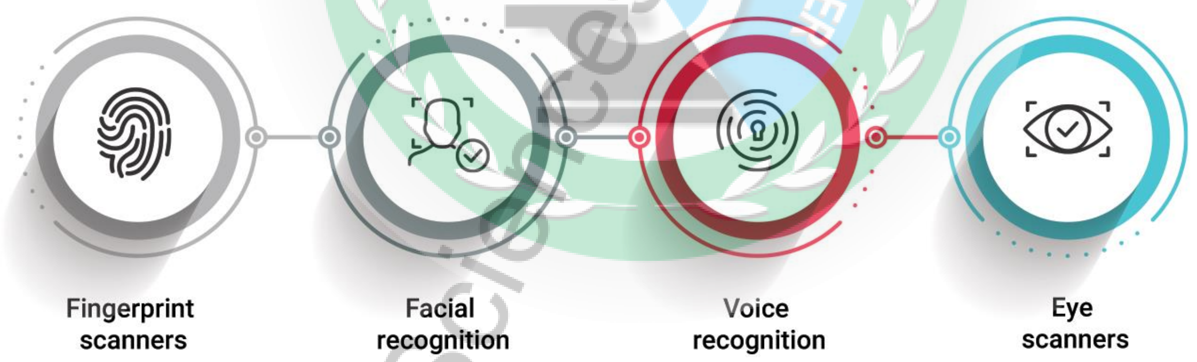


Figure 1.1. Types of biometric authentication

1.2. BRANCHES OF BIOSTATISTICS

Biostatistical applications and methods are employed to solve problems in the biological and health sciences. Bioinformatics is mainly used to extract knowledge from biological data through the development of algorithms and software. Bioinformatics is widely applied in the examination of Genomics, Proteomics, 3D structure modelling of Proteins, Image analysis, Drug designing and a lot more.

There are several branches of biostatistics that are commonly recognized.

1. Descriptive Biostatistics
2. Inferential Biostatistics
3. Survival Analysis
4. Epidemiology
5. Biopharmaceutical Statistics
6. Bioinformatics:
7. Bayesian Biostatistics

1.2.1. DESCRIPTIVE BIOSTATISTICS

This branch of biostatistics deals with summarizing and presenting data collected from a population or sample using statistical measures such as mean, median, mode, range, and standard deviation. Descriptive biostatistics is employed for:

- Producing quantitative summaries of information in biological sciences
- Summarize or describe features of a data set
- Tabulation and graphical presentation

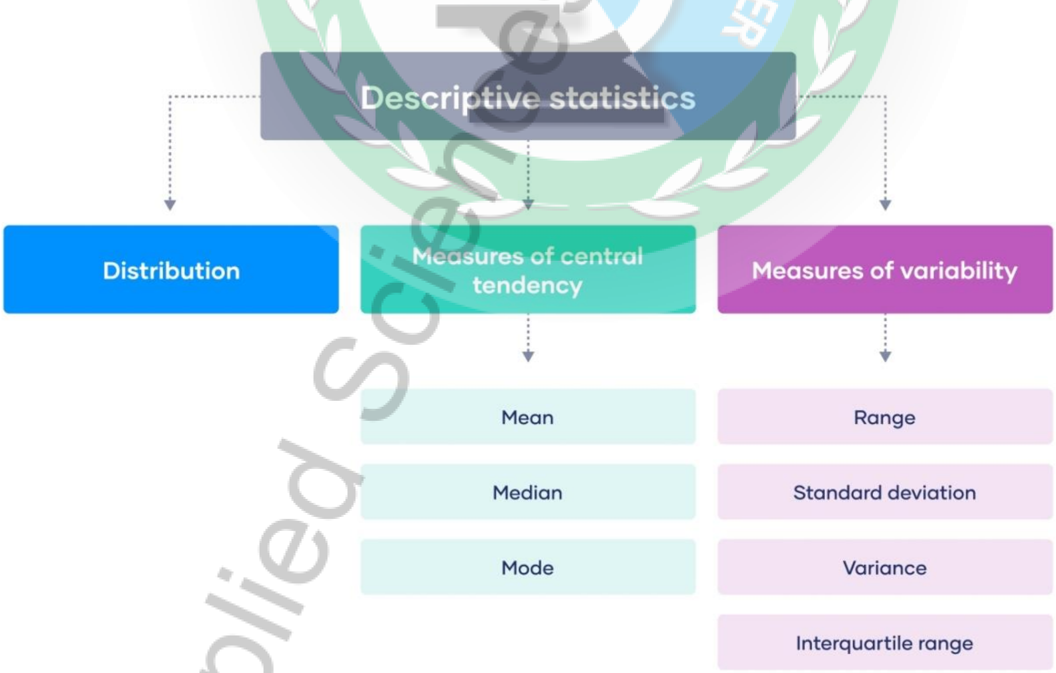


Figure 1.2. Types of descriptive biostatistics

1.2.2. INFERENTIAL BIOSTATISTICS

This branch of biostatistics involves making inferences and drawing conclusions about a population based on data collected from a sample. This includes hypothesis testing and estimation. Inferential biostatistics is useful in making generalizations about a larger group based on information from a sample of that particular group in biological sciences.

Primarily, inferential biostatistics is performed in two ways:

- Estimation
- Testing of hypothesis

Inferential statistics: testing of statements about the population on the basis of sample characteristics.

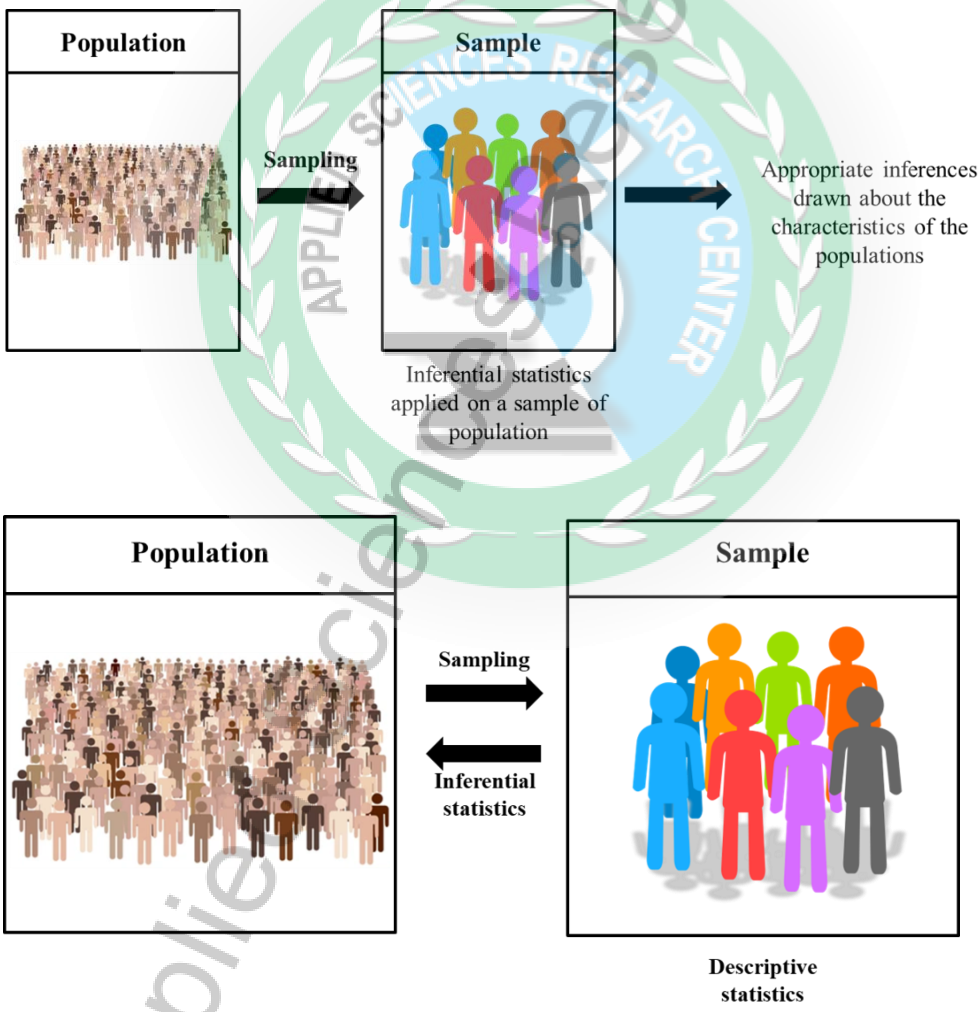
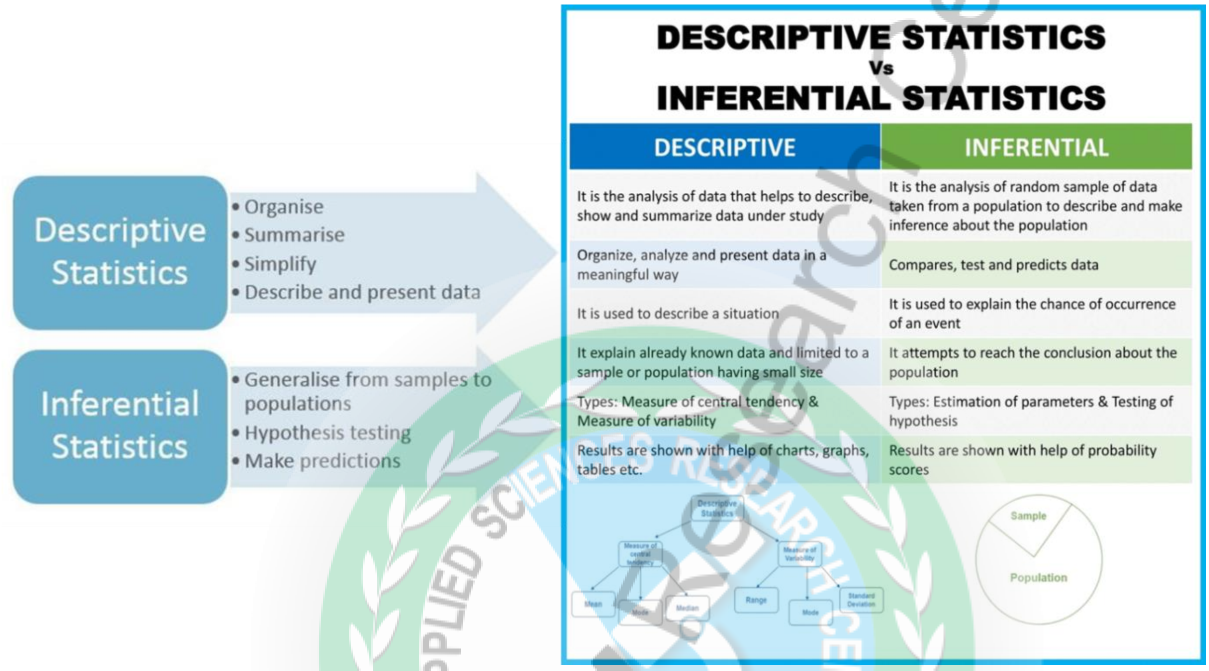

Population
Sampling
Sample
Appropriate inferences drawn about the characteristics of the populations
Inferential statistics applied on a sample of population
Population
Sampling
Inferential statistics
Sample
Descriptive statistics

Figure 1.3. Drawing conclusions about a population based on data collected from a sample

1.2.2.1. Descriptive Biostatistics vs Inferential Biostatistics

Descriptive biostatistics describes the characteristics of data such as population frequency variables whereas; the inferential biostatistics encompasses studies on sample of same data such as grade and percentile.



DESCRIPTIVE STATISTICS Vs INFERENTIAL STATISTICS

DESCRIPTIVE	INFERENTIAL
It is the analysis of data that helps to describe, show and summarize data under study	It is the analysis of random sample of data taken from a population to describe and make inference about the population
Organize, analyze and present data in a meaningful way	Compares, test and predicts data
It is used to describe a situation	It is used to explain the chance of occurrence of an event
It explain already known data and limited to a sample or population having small size	It attempts to reach the conclusion about the population
Types: Measure of central tendency & Measure of variability	Types: Estimation of parameters & Testing of hypothesis
Results are shown with help of charts, graphs, tables etc.	Results are shown with help of probability scores

Figure 1.4. Descriptive biostatistics vs inferential biostatistics

Descriptive Biostatistics		Inferential Biostatistics	
Measures of Central Tendency	**Measures of Dispersion**	**Hypothesis Testing**	**Regression Analysis**
Mean	Range	Z test	Linear Regression
Median	Standard Deviation	F test	
Mode	Variance	T test	
	Absolute Deviation		

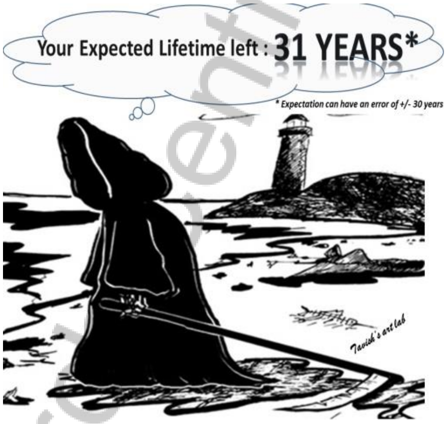
Table 1.1. Testing descriptive and inferential data

1.2.3. SURVIVAL ANALYSIS

Survival analysis is a statistical technique used to analyze time-to-event data. It is widely used in various fields, including medicine, engineering, finance, and social sciences. In survival analysis, the focus is on the time until a particular event of interest occurs, such as death, failure of a machine, or the onset of a disease.

There are several important concepts in survival analysis.

1. Survival function
2. Hazard function
3. Cumulative hazard function


Your Expected Lifetime left : 31 YEARS*
* Expectation can have an error of +/- 30 years
Tavish's art lab

1.2.3.1. Survival Function

The survival function represents the probability that an individual or system survives beyond a certain time point.

1.2.3.2. Hazard Function

The hazard function represents the instantaneous rate of occurrence of the event of interest, given that the individual or system has survived up to that time point.

1.2.3.3. Cumulative Function

The cumulative hazard function represents the cumulative risk of the event of interest up to a certain time point.

1.2.4. EPIDEMIOLOGY

The branch of biostatistics that involves studying the distribution and determinants of health and disease in populations is called epidemiology. It includes the design and analysis of clinical trials and observational studies.

By definition, epidemiology is the study (scientific, systematic, and data-driven) of the distribution (frequency, pattern) and determinants (causes, risk factors) of health-related states and events (not just diseases) in specified populations (neighborhood, school, city, state, country, global).

There are two major types of epidemiologic techniques.

1. **Observational Study**
   a. Descriptive Study
      i. Ecological studies
      ii. Cross-sectional studies
   b. Analytical Study
      i. Case control study

ii. Cohort studies

2. **Interventional Study**

    a. Randomized controlled trials (RCTs)

    b. Quasi-experimental studies

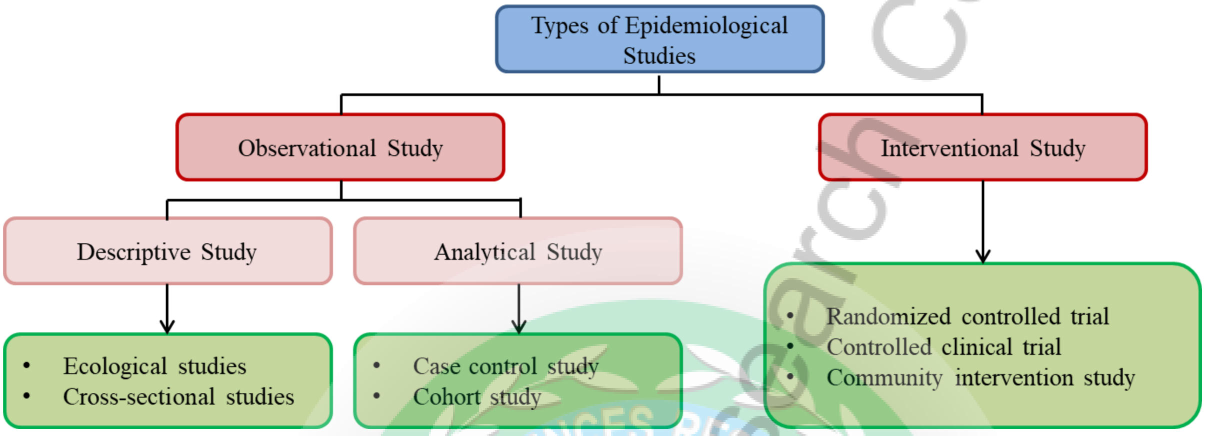

Types of Epidemiological Studies
Observational Study
Interventional Study
Descriptive Study
Analytical Study
• Ecological studies
• Cross-sectional studies
• Case control study
• Cohort study
• Randomized controlled trial
• Controlled clinical trial
• Community intervention study

Figure 1.5. Major types of epidemiologic techniques

1.2.4.1. Observational study

An observational study is where the researcher observes and measures the characteristics and behaviors of a group of individuals without intervening or manipulating any variables. The researcher simply observes without interference or manipulation.

Observational studies are useful in situations where it is not ethical or practical to conduct a randomized controlled trial (RCT), which is a type of study where the researcher randomly assigns participants to different groups to compare the effects of different interventions.

There are two main types of observational studies:

1. **Descriptive Study**
    a. Ecological studies
    b. Cross-sectional studies
2. **Analytical Study**
    a. Case control studies
    b. Cohort studies

DESCRIPTIVE STUDY

Ecological studies: In epidemiology ecological studies are used to investigate the association between population-level exposures and health outcomes. Unlike other types of epidemiological studies that look at individual-level data, ecological studies analyze data that is aggregated at the group or population level. This can include data on environmental exposures, social determinants of health, and other factors that may impact health outcomes.

Cross-sectional studies: In a cross-sectional study, the researcher collects data at a single point in time from a group of individuals to assess the prevalence of a particular disease or health-related behavior. For example, a researcher might conduct a cross-sectional study to assess the prevalence of smoking among college students.

ANALYTICAL STUDY

Case control studies: Case-control studies are commonly to investigate the association between potential risk factors and disease outcomes. In case-control studies, researchers identify individuals with the disease of interest (cases) and individuals without the disease (controls), and then compare the frequency of exposure to potential risk factors between the two groups.

Cohort studies: In a cohort study, the researcher follows a group of individuals over time to assess the incidence of a particular disease or health-related behavior. For example, a researcher might conduct a cohort study to assess the incidence of cardiovascular disease among individuals with high blood pressure.

1.2.4.2. Interventional Study

An interventional study is a type of research study in which the researcher manipulates one or more variables to observe the effect of the manipulation on a particular outcome. This type of study is also known as an experimental study, because the researcher is actively intervening in the study subjects.

In an interventional study, the researcher randomly assigns participants to different groups, where each group receives a different treatment or intervention. The researcher then observes the outcomes of each group and compares them to determine the effect of the intervention.

There are two main types of interventional studies:

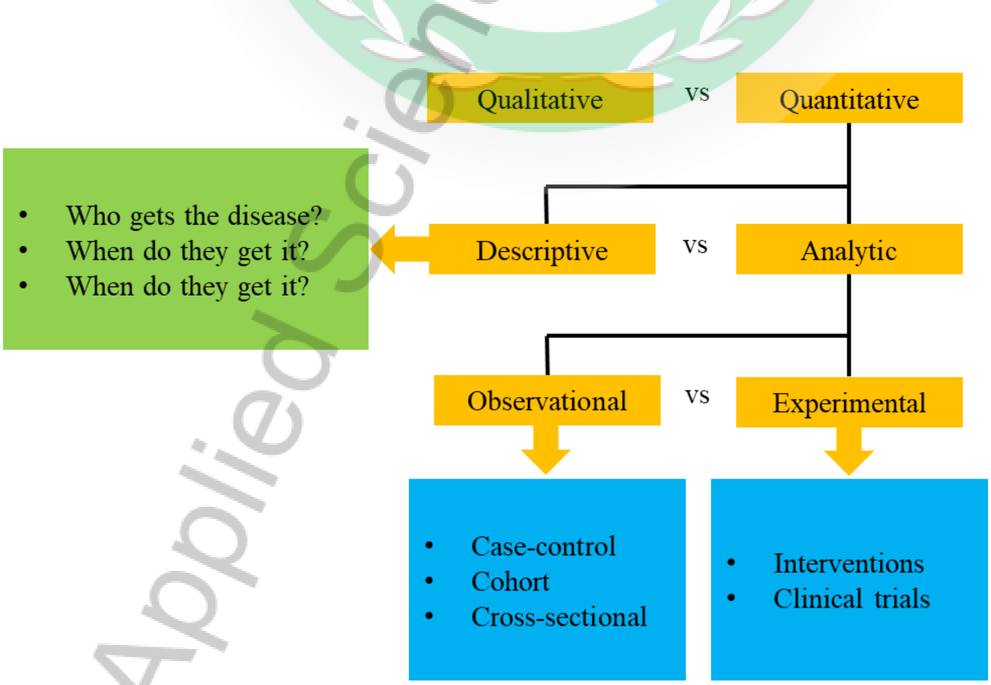
1. Randomized controlled trials (RCTs)
2. Quasi-experimental studies

Randomized controlled trials (RCTs): In an RCT, participants are randomly assigned to either a treatment group or a control group. The treatment group receives the intervention being studied, while the control group does not receive any intervention or receives a placebo. RCTs are considered the gold standard for evaluating the effectiveness of interventions.

Quasi-experimental studies: In a quasi-experimental study, participants are not randomly assigned to groups. Instead, the researcher compares the outcomes of a group that receives the intervention to the outcomes of a group that does not receive the intervention. Quasi-experimental studies are often used when it is not ethical or practical to conduct an RCT.

1.2.4.3. Observational Study vs Interventional Study

Interventional studies have some advantages over observational studies, as they allow researchers to manipulate variables and observe cause-and-effect relationships. However, they can also have some limitations, such as the potential for bias and ethical concerns. Therefore, careful design and implementation of interventional studies are essential to ensure their validity and reliability.



Study Design	Measure of Disease	Measure of Risk	Temporality
Ecological	Prevalence	Prevalence Ratio	Retrospective
Mortality	Proportional Mortality Standardized Mortality	Proportional Mortality Ratio Standardized Mortality Ratio	Retrospective
Cross Sectional	Point Prevalence Period Prevalence	Odds Ratio Prevalence Odds Ratio Prevalence Ratio Prevalence Differences	Retrospective
Case-Control	None	Odds Ratio	Retrospective
Cohort	Point Prevalence Period Prevalence Incidence	Odds Ratio Prevalence Odds Ratio Prevalence Ratio Prevalence Differences Incidence Rate Ratio Relative Risk Risk Ratio Hazard Ratio	Retrospective only Both Retrospective & Prospective Prospective only

Tabl 1.2. Comparison of study design

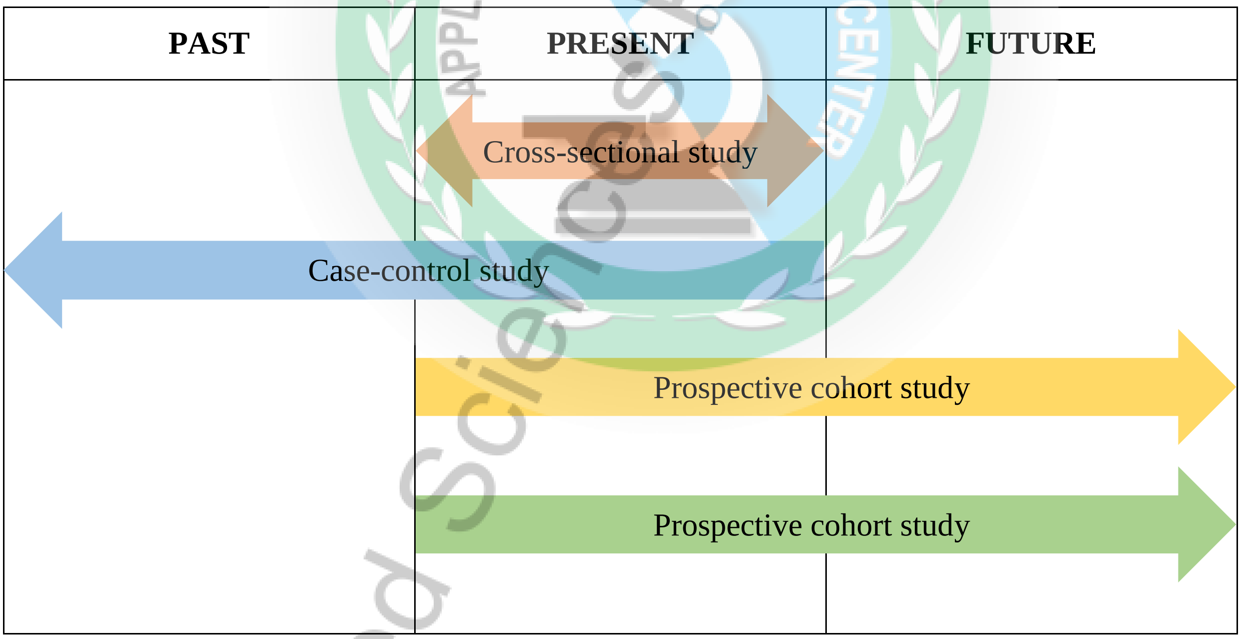


Figure 1.6. Observational study vs interventional study

1.2.4.4. Types of epidemiological studies

There are several types of epidemiological studies.

1. Descriptive epidemiology
2. Analytical epidemiology

3. Clinical epidemiology
4. Molecular epidemiology
5. Social epidemiology
6. Environmental epidemiology
7. Infectious disease epidemiology

Descriptive epidemiology: This type of epidemiology aims to describe the patterns of disease occurrence in a population. It provides information on the distribution of disease by person, place, and time.

Analytical epidemiology: This type of epidemiology aims to identify the causes of disease by studying the relationships between exposure to risk factors and the occurrence of disease. It includes both observational and experimental study designs.

Clinical epidemiology: This type of epidemiology focuses on the application of epidemiological principles and methods to the study of clinical practice, including the diagnosis, treatment, and prognosis of diseases.

Molecular epidemiology: This type of epidemiology uses molecular and genetic techniques to identify the genetic and environmental factors that contribute to the development of diseases.

Social epidemiology: This type of epidemiology focuses on the social determinants of health, such as poverty, education, and social inequality, and how they affect the distribution of disease in populations.

Environmental epidemiology: This type of epidemiology focuses on the effects of environmental exposures, such as pollution and toxins, on the development and distribution of disease.

Infectious disease epidemiology: This type of epidemiology focuses on the study of infectious diseases, including the transmission, distribution, and control of infectious agents in populations.

1.2.5. BIOPHARMACEUTICAL STATISTICS

This branch of biostatistics deals with statistical methods used in drug development, including clinical trial design, data analysis, and regulatory submissions.

1.2.6. BIOINFORMATICS

This branch of biostatistics deals with the analysis of biological data, particularly high-throughput data such as genomics, proteomics, and metabolomics. It involves the development and application of statistical and computational methods for analyzing large and complex datasets.

1.2.7. BAYESIAN BIOSTATISTICS

This branch of biostatistics is concerned with using Bayesian methods for analyzing data in the biological and health sciences. It involves specifying a prior distribution for a parameter of interest and updating this distribution based on observed data.

1.3. BASIS OF BIOSTATISTICS

1.3.1. INTRINSIC VARIATIONS

Intrinsic variations in epidemiology are used as a source of medical uncertainties. **Intrinsic variability** is a relatively well understood aspect of biological models. It arises from the probabilistic nature of the timing of collision events between reacting biological molecules, and its effect is most pronounced when the number of molecules in the system is small.

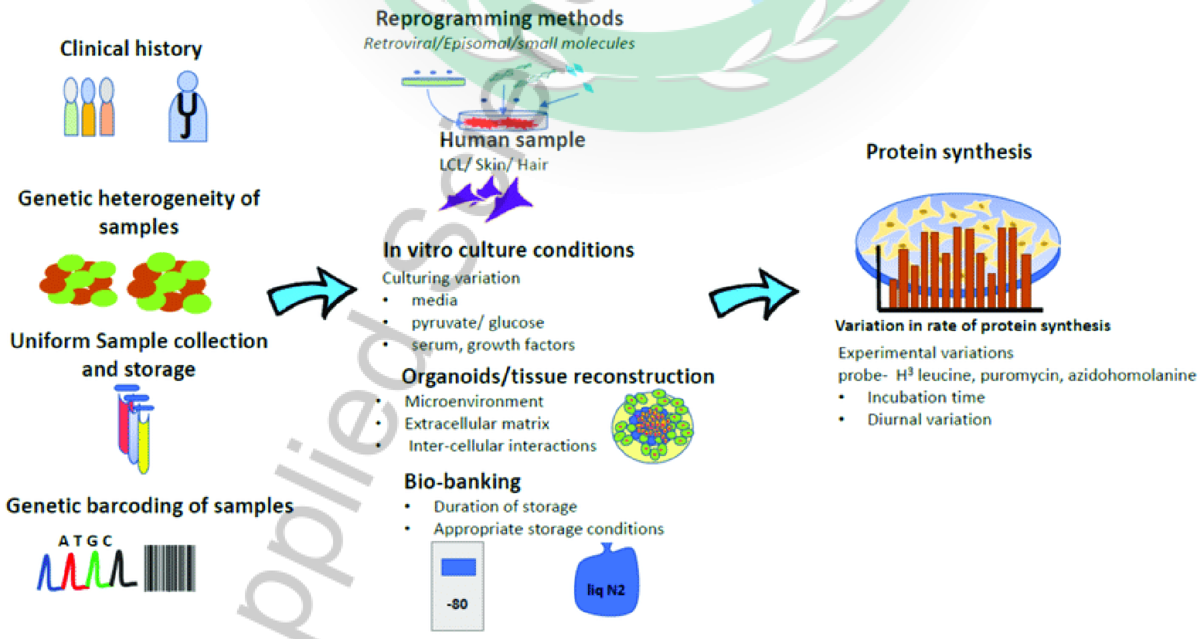

Clinical history
Genetic heterogeneity of samples
Uniform Sample collection and storage
Genetic barcoding of samples
ATGC
Reprogramming methods
Retroviral/Episomal/small molecules
Human sample
LCL/ Skin/ Hair
In vitro culture conditions
Culturing variation
• media
• pyruvate/ glucose
• serum, growth factors
Organoids/tissue reconstruction
• Microenvironment
• Extracellular matrix
• Inter-cellular interactions
Bio-banking
• Duration of storage
• Appropriate storage conditions
-80
liq N2
Protein synthesis
Variation in rate of protein synthesis
Experimental variations
probe- H³ leucine, puromycin, azidohomolanine
• Incubation time
• Diurnal variation

Figure 1.7. Schematic representation of intrinsic variation at multiple levels

Biological variations occur due to age, gender, heredity, height and weight etc. Anatomical, physiological and biomechanical variations are also biological parameters. Environment variations occur due to nutrition, smoking, pollution, facilities of water and sanitation, road traffic, stress and strain etc. Everything in medicine be it research, diagnosis or treatment, depends on counting or measurement. High or low blood pressure has no meaning, unless it is expressed in figures. In nature, blood pressure, pulse rate, action of a drug or any other measurement or counting varies not only from person to person but also from group to group. Variation more than natural limits may be pathological, i.e., abnormal due to the play of certain external factors. Hence biostatistics may also be called a science of variation. The ultimate objective of biostatistics is improve the health of individuals and community

Biostatistics deals with development and application of the most appropriate methods for the:

- Collection of data
- Presentation of the collected data
- Analysis and interpretation of the results
- Making decisions on the basis of such analysis

1.4. SCOPE OF BIOSTATISTICS

Knowledge of the statistical methods is necessary in research and medical field for data analysis and possible outcomes of the research. It enables us to study and compare the significance of treatments in medical sciences. Biostatistics is applied in all branches of life sciences. Following are few applications of it.

1.4.1. PHARMACOLOGY

1.4.1.1. Clinical trials: Biostatisticians help design and analyze clinical trials, which are studies that test the safety and efficacy of new drugs in human subjects. They help determine the appropriate sample size, randomization procedures, and statistical tests to be used. They also analyze the data generated by these trials to determine whether the drug is safe and effective, and to identify any adverse effects.

1.4.1.2. Drug development: Biostatisticians help pharmaceutical companies design and analyze preclinical studies, which are studies that test the safety and efficacy of drugs in animals before they are tested in humans. They also help design and analyze pharmacokinetic and pharmacodynamic studies, which investigate how drugs are absorbed, metabolized, and eliminated in the body, and how they interact with their targets.

1.4.1.3. Pharmacovigilance: Biostatistics is used in pharmacovigilance, which is the monitoring of the safety of drugs that are already on the market. Biostatisticians help to identify and analyze adverse events and to determine whether there is a causal relationship between the drug and the event.

1.4.1.4. Pharmacoeconomics: Biostatistics is also used in pharmacoeconomics, which is the study of the costs and outcomes of drug therapies. Biostatisticians help to design and analyze cost-effectiveness studies, which evaluate the relative costs and benefits of different treatment options.

1.4.2. MEDICINE

1.4.2.1. Study design: Biostatisticians help to design studies that are well-controlled and appropriate for answering specific research questions. They ensure that studies are statistically valid and are able to generate meaningful results.

1.4.2.2. Data analysis: Biostatisticians use statistical methods to analyze the data generated by clinical studies. They help to identify patterns, relationships, and trends in the data and determine whether the results are statistically significant.

1.4.2.3. Clinical trials: Biostatisticians play a crucial role in clinical trials by helping to determine sample sizes, randomization procedures, and statistical tests to be used. They also help to ensure that the results of clinical trials are reliable and valid.

1.4.2.4. Epidemiology: Biostatistics is also used in epidemiological studies to identify risk factors for diseases, determine the prevalence of diseases in specific populations, and evaluate the effectiveness of public health interventions.

1.4.2.5. Predictive modeling: Biostatisticians use statistical methods to develop predictive models that can be used to forecast the likelihood of developing certain diseases or conditions based on a variety of factors, such as age, sex, genetics, and lifestyle.

1.4.3. CLINICAL STUDY

1.4.3.1. Study design: Biostatisticians help to design clinical studies that are well-controlled and appropriate for answering specific research questions. They ensure that studies are statistically valid and are able to generate meaningful results.

1.4.3.2. Sample size determination: Biostatisticians help determine the appropriate sample size for a study to ensure that it has adequate statistical power to detect meaningful differences between treatment groups.

1.4.3.3. Randomization: Biostatisticians help to randomize study participants to treatment groups, which ensures that any differences between treatment groups are due to the treatment itself and not to other factors such as age or gender.

1.4.3.4. Data analysis: Biostatisticians use statistical methods to analyze the data generated by clinical studies. They help to identify patterns, relationships, and trends in the data and determine whether the results are statistically significant.

1.4.3.5. Safety monitoring: Biostatisticians help monitor the safety of study participants, including the incidence of adverse events, and determine whether these events are related to the treatment being studied.

1.4.3.6. Interim analysis: Biostatisticians may perform interim analyses of the data during the course of the study to determine whether the study should be stopped early due to overwhelming evidence of a treatment effect or futility.

1.4.4. PREVENTIVE MEDICINE

1.4.4.1. Risk assessment: Biostatisticians use statistical methods to assess the risk of developing certain diseases or conditions based on a variety of factors, such as age, sex, genetics, and lifestyle. This information can be used to develop targeted prevention strategies for at-risk populations.

1.4.4.2. Screening programs: Biostatistics is used to design and evaluate screening programs that identify individuals at high risk of developing certain diseases, such as breast cancer or colorectal cancer. Biostatisticians help to determine the appropriate screening intervals and the best tests to use.

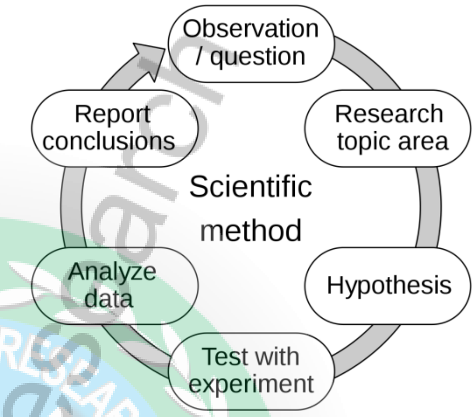
1.4.4.3. Epidemiological studies: Biostatistics is used in epidemiological studies to identify risk factors for diseases, determine the prevalence of diseases in specific populations, and evaluate the effectiveness of preventive interventions.

1.4.4.4. Clinical trials: Biostatistics is used in clinical trials of preventive interventions, such as vaccines or lifestyle interventions, to determine their effectiveness in preventing diseases.

1.4.4.5. Public health policy: Biostatistics is used to inform public health policy decisions, such as the implementation of vaccination programs or the promotion of healthy lifestyle choices.

1.4.5. SCIENTIFIC STUDY

Biostatistics is very important in research because it is required for



- Chromosomal mapping
- Creating logical conclusions of the data
- Derive single values from group of data
- Designing the experiments
- Gene inheritance
- Gene structure of a population
- Mendelian genetics
- Population genetics i.e. behavior of genes in population
- Research work is incomplete without statistical analysis
- Selecting data collection methods
- Summarization and presentation of data
- Validation of scientific results

1.4.6. BIOTECHNOLOGY

1.4.6.1. Experimental design: Biostatisticians can help design experiments that optimize the use of resources and minimize variability in the data generated. This includes selecting appropriate sample sizes, randomization methods, and statistical models for data analysis.

1.4.6.2. Data analysis: Biostatisticians can analyze the data generated from biotech experiments to help identify trends, determine the significance of differences between groups, and develop predictive models. They may also use advanced statistical methods, such as machine learning and artificial intelligence, to analyze complex data sets.

1.4.6.3. Quality control: Biostatisticians can help develop quality control processes to ensure that biotech products meet regulatory standards and are safe and effective for their intended use.

1.4.6.4. Clinical trials: Biostatisticians can design and analyze clinical trials to evaluate the safety and efficacy of biotech products in humans. This includes determining appropriate sample sizes, randomization methods, and statistical models for data analysis.

1.4.6.5. Regulatory compliance: Biostatisticians can help ensure that biotech companies comply with regulatory requirements for the development and testing of new products. This includes designing studies that meet regulatory standards and preparing reports that demonstrate the safety and efficacy of new products.

1.4.7. AGRICULTURE

1.4.7.1. Experimental design: Biostatisticians can help design experiments that optimize the use of resources and minimize variability in the data generated. This includes selecting appropriate sample sizes, randomization methods, and statistical models for data analysis.

1.4.7.2. Data analysis: Biostatisticians can analyze data generated from agricultural experiments to identify trends, determine the significance of differences between groups, and develop predictive models. They may also use advanced statistical methods, such as machine learning and artificial intelligence, to analyze complex data sets.

1.4.7.3. Modeling: Biostatisticians can develop models that predict the growth and yield of crops, the spread of pests and diseases, and the impact of environmental factors on agricultural production.

1.4.7.4. Quality control: Biostatisticians can help develop quality control processes to ensure that agricultural products meet regulatory standards and are safe and effective for their intended use.

1.4.7.5. Decision-making: Biostatisticians can provide decision support for farmers, agribusinesses, and policymakers by analyzing data and developing models that inform decisions related to crop and animal production, resource management, and environmental sustainability.

1.4.7.6. Purpose and objective of agricultural statistics

- To provide comprehensive knowledge of the basic information of agriculture, rural areas and the farmers.

- To provide the scientific basis for the study of the development of economic and social development, planning and decision making.
- To provide statistical information services to the planners, scholars and public
- Is of prime importance for agricultural industry.
- Are important in designing development policies in the agricultural sector and the national economy at large.
- Ascertain the crop production, crop yield, qualities of crop produced.
- Furnish information about different operations and different methods which can be adopted for improving crop output.
- Helps to compare the different yields of crops and quality check of crops.

1.4.8. PHYLOGENETIC ANALYSIS

Statistics is critical in analyzing patterns of genomic variation between populations/species. Phylogenetic analysis provides an in-depth understanding of how species evolve through genetic changes. Using phylogenetics, scientists can evaluate the path that connects a present-day organism with its ancestral origin, as well as can predict the genetic divergence that may occur in the future.

1.4.9. NUTRITION

Nutritionists now have the advanced methodologies for the analysis of DNA, RNA, protein, low-molecular-weight metabolites, as well as access to bioinformatics databases. Appropriate statistical analyses are expected to make an important contribution to solving major nutrition-associated problems in humans and animals (including obesity, diabetes, cardiovascular disease, cancer, ageing, and intrauterine growth retardation).

1.4.10. ENVIONMENTAL SCIENCES

- Thinking about risk.
- Probability of an unwanted outcome.
- Scientists calculate the risk to large communities of people in order to inform policy decisions in government.
    - Involve using data on historical events.

  - Projections for the future.
  - Scientific evidence for adverse outcomes.
  - Example: The risk of getting cancer when exposed to pollution.

1.4.11. COMPUTATIONAL BIOLOGY

Computational biology uses mathematical and informational techniques including statistics to solve biological problems. By using computer programs or mathematical models or creating both. One of the major areas of computational biology is data mining which includes the analysis of the data collected by several genome projects. Genome projects are scientific projects that are utilized to map the genome of a living being.

1.5. LIMITATIONS OF BIOSTATISTICS

- Statistical analysis cannot be performed on single observation, it is true on average.
- It cannot be applied to single data.
- It cannot be applied to heterogeneous data.
- Best applied on quantitative data.
- Expertise is needed in collecting, analyzing and inference of data otherwise the results can get wrong.
- Statistical results are not always accurate.
- Errors can be possible (Type I and II errors).

1.6. ROLE OF BIOSTATISTICIANS

- Identify and develop treatments for disease and estimate their effects.
- Identify risk factors for diseases.
- Design, monitor, analyze, interpret, and report results of clinical studies.
- Develop statistical methodologies to address questions arising from medical/public health data.
- Locate, define and measure extent of disease.
- The ultimate objective: Improve the health of individual & community.

1.7. COMMON TERMS USED IN BIOSTATISTICS

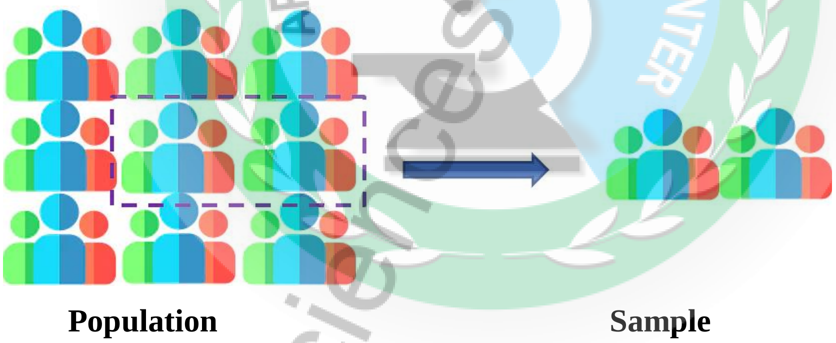
To understand the statistical methods, it is necessary to explain some common terms used in biostatistics.

1.7.1. POPULATION

The term “population” in biostatistics is used for the “group of measurements” under study or it is the “unit” upon which conclusions are drawn. For example, if we want to study the birds’ biodiversity of a specific area then in biostatistics bird biodiversity of that particular area is population. Likewise, if we want to assess the quality of tablets of a batch then all tablets of a specific batch will be the population. The term includes all organisms, objects and events under study. Population can be finite or infinite. The countable population is considered as finite whereas the uncountable population is infinite population and it is comprised of infinite number of elements.

1.7.2. SAMPLE

The term “sample” is applied for the portion of population used for study. In statistics a portion of population is studied instead of working on whole population and then estimates are made about the population. It is described in terms of size, structure, nature and time frame. Scientifically when all the individuals are given equal chances of selection then it is called random sampling which is the most appropriate method of sampling. Larger sample of a population contain more than 30 individuals whereas less than this is considered small sample.



1.7.3. PARAMETER

Parameter is the feature or characteristic of a population and it is a fixed quantity. For example average age of the students of a graduation class. This quantity is always estimated.

1.7.4. STATISTICS

Statistics is the feature or characteristic of a sample and it is not a fixed quantity. It may vary from sample to sample. So statistics is the average sample that varies even if the samples were taken from same population.

1.7.5. DESCRIPTIVE STUDIES

Descriptive studies involve the collection, organization, presentation and summarization of the statistical data. This involves the study of mean, mode, median, standard deviations, graphical representation of the data and no conclusion or inferences have drawn in descriptive studies. For example, birth rate, death rate and development of a chick embryo all these come under descriptive studies and no inferences is involved. This may also include case, report and case study.

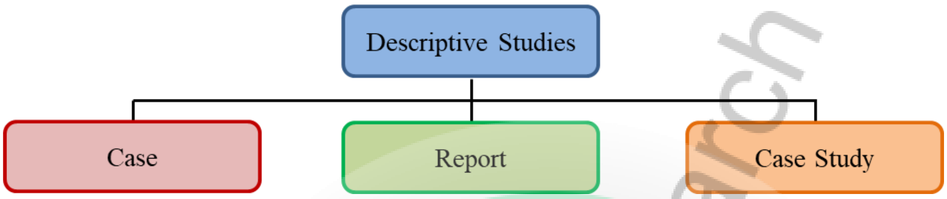


Figure 1.8. An overview of Descriptive Studies

1.7.6. ANALYTICAL OR INFERENTIAL STUDIES

Analytical or inferential studies are conducted to draw conclusions or inferences. The research conducted on dengue is included in analytical studies because it involves causes, spread and prevalence of the dengue fever. This study makes conclusions about population. Random sampling techniques make it more reliable and predictions are made on the basis of analytical studies. Studies based on observations include cohort, case-control and lab trials.

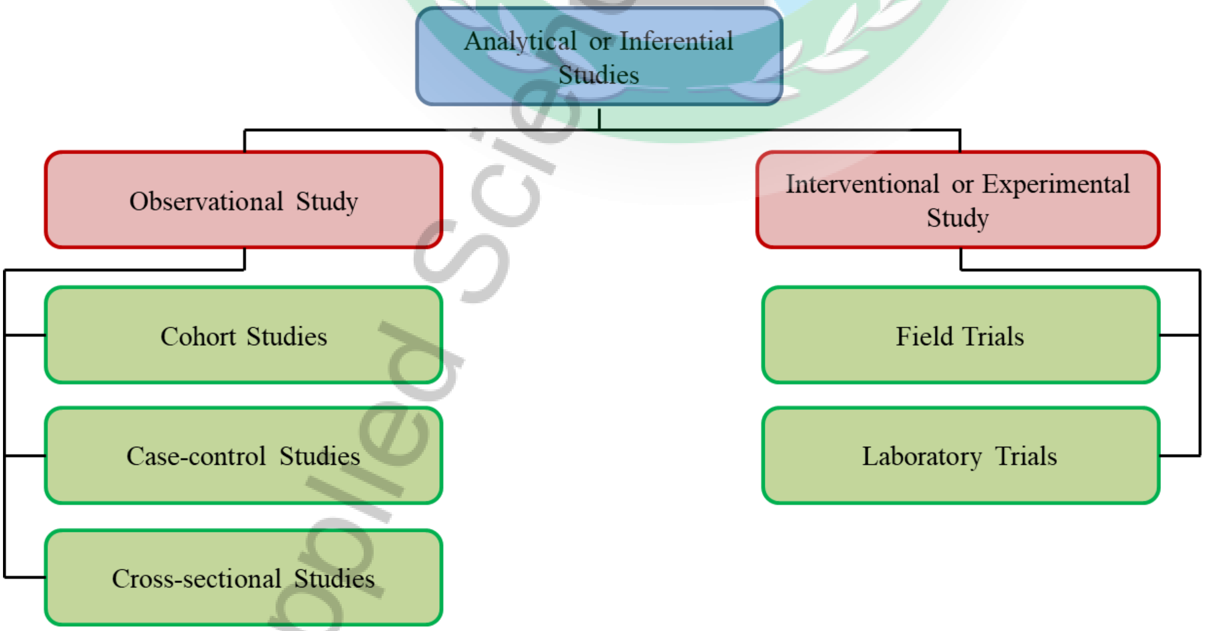


Figure 1.9. An overview of Analytical Studies

1.7.7. COHORT

Cohort is a kind of follow up study in which the study group is analyzed forward in the time. For example in this phenomenon specific organism or case is exposed to certain factor which is related to a specific disease then the object or case is studies for future development and control of the disease. To study the effect of a particular substance the objects are exposed to that substance and some non- exposed groups are also studied for comparison. A long time period is required for this study.

Example: In a cohort of 100 poultry chicken, two groups are made, each group was comprised of 50 chickens. Standard poultry feed was given to Group A, and enriched low cost poultry feed was given to Group B to check the effect of standard and enriched poultry feed on development of the chicken.

Group	Hen	Poultry Feed
A	50	Standard
B	50	Enriched

Table 1.3. A summarization of a cohort study

1.7.8. CASE CONTROL STUDY

Case control study begins with past and it is in backward direction. This usually starts with outcome of a specific disease and then the causes of that disease is studied. This is a short time study and called as retrospective study. It involves data gathering of past selected cases with risk factors and compare it with control to evaluate the differences. **Control** study involves the objects that are free from that factor that is to be studied. This is helpful in comparison. The relation of smoking with lungs cancer is case control study and it is compared with normal cases or unaffected population which is the **control**.

Sample	Lung Cancer (Case)	Normal (Control)
Smoker	40	5
Non-smoker	10	45
Total	100	100

Table 1.4. Summarization of case-control study and control

1.7.9. EXPERIMENTAL TRIALS

Experimental trials are observations are checked in controlled environment and all the conditions are specific which are set by the researcher. For example effect of various commonly used antibiotics are checked in laboratory on multidrug resistant bacteria. Effectiveness of antibiotics can be determined through experimental trials.

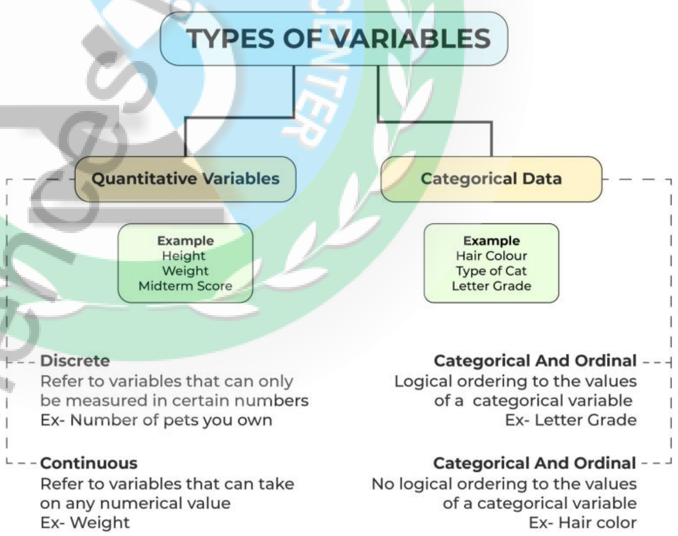
1.7.10. VARIABLES

Variable is the feature of an individual which shows different values at different circumstances for example age, weight and height of individual, bacterial colony shape, color, etc. all these are example of variables. Variables are of two types categorical and numerical. Variable can be dependent and independent as well.

1.7.10.1. Categorical Variable is the characteristic in which observations are noted in form of categories like age group of an individual child, young, middle age and old. There are four categories of age group. Gender is a variable which has two categories male or female. This variable is a qualitative one.

1.7.10.2. Numerical Variables is the quantitative variable in which observations are recorded in the form of numerical values such as age, height, weight of an individual. Numerical variables are further divided into two type continuous and discrete variable.

- Continuous Variables
- Discrete Variables


TYPES OF VARIABLES
Quantitative Variables
Categorical Data
Example
Height
Weight
Midterm Score
Example
Hair Colour
Type of Cat
Letter Grade
Discrete
Refer to variables that can only be measured in certain numbers
Ex- Number of pets you own
Categorical And Ordinal
Logical ordering to the values of a categorical variable
Ex- Letter Grade
Continuous
Refer to variables that can take on any numerical value
Ex- Weight
Categorical And Ordinal
No logical ordering to the values of a categorical variable
Ex- Hair color

- **Continuous Variables:** Variables which take forever to count i.e. they are infinite. For example you can count the time in hours, minutes second, milliseconds but counting in nanoseconds, picoseconds will be infinite.

- **Discrete Variables:** Variables which are countable and finite variable like age, weight, time in hours, minutes and seconds, etc.

1.7.10.3. Independent and Dependent Variables

Independent variables are the causes to which dependent variables respond e.g. if we want to check the effect of different concentrations of fish feed on fish growth then here fish feed is independent variable and growth of fish is dependent variable.

CHAPTER TWO

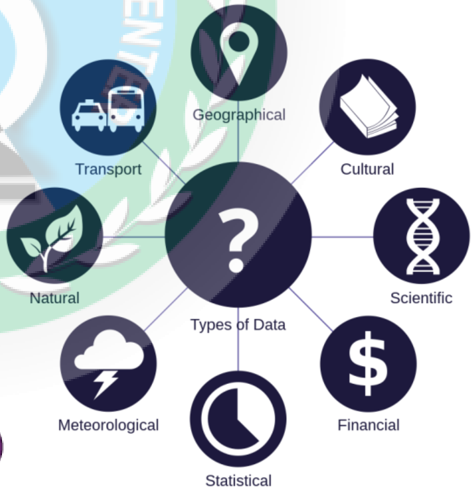
Data Collection, Arrangement and Presentation

2.1. TYPES OF DATA IN BIOSTATISTICS

Data are the specific pieces of information that are gathered and used for analysis. There are various types of data that are collected, presented, analyzed and inferred. Data handling is very important which involves various processes such as tagging, labelling and confidentiality of the data whereas; the analysis of data is presentation and interpretation which results into statistics.

There are two basic types of data which are divided to form 4 different types.

1. **Qualitative or Categorical Data**
   a. Nominal Data
   b. Ordinal Data
2. **Quantitative or Numerical Data**
   a. Discrete Data
   b. Continuous Data



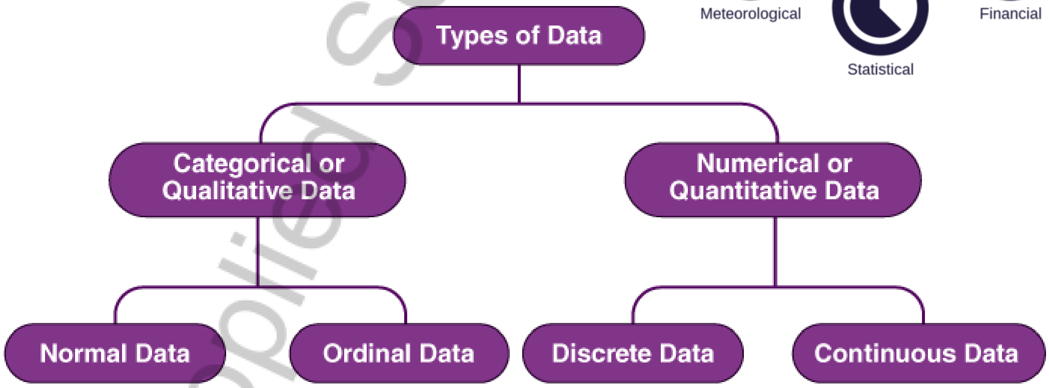


Figure 2.1. Types of statistical data

2.1.1. QUALITATIVE OR CATEGORICAL DATA

Qualitative or categorical data are based on categories that do not contain numerical data. Various variables such as birth place, identity mark, color of individuals, etc.

2.1.1.1. Nominal Data

This is type of qualitative or categorical data which is used to label the variables without signing numerical values. Nominal can be of qualitative or quantitative type but it does not provide numerical values. This data is used in grouping methods in which data is divided into categories and then percentage or frequency is calculated. This data is represented in form of pie chart.

2.1.1.2. Ordinal Data

Ordinal data follows the natural order and in this type differences of data values are not determined. It is mostly used in questionnaire, survey, etc. The data is presented in the form of bar chart and information in data is expressed through tables with each row showing a category.

2.1.2. QUANTITATIVE OR NUMERICAL DATA

It is represented in the form of numerical values and this data gives quantity of a specific thing like how much, how many, etc. Numerical data may include variables like age, weight, length, etc. there are two different types of numerical data.

2.1.1.1. Discrete Data

This data take discrete values and it contains finite number of values. Variables are counted in whole numbers for example number of children in 8th class.

2.1.1.2. Continuous Data

This data has infinite number of values that are selected within range for example pH range, temperature range, etc.

2.2. METHODS OF DATA COLLECTION

Data collection is the first step in research work when we want to analyze a problem then information related to that specific topic is gathered from all possible sources to find the solution. It is very important in finding solution of a research problem and evaluate the possible research outcomes. Mostly information in data is based on assumptions about future trends. There are two types of data collection methods which are further sub-divided.

1. Primary Data Collection Methods
    a. Quantitative Data Collection Methods
    b. Qualitative Data Collection Methods
2. Secondary Data Collection Methods

2.2.1. PRIMARY DATA COLLECTION METHODS

Primary data is collected from the first-hand source via surveys, experiments or observations. It is further sub-divided into two types:

2.2.1.1. Quantitative Data Collection Methods

These are the mathematical calculations that are formed using different formats like mean, mode, median, correlation and regression, close ended questions, etc. This method is cheaper and applied in short duration.

2.2.1.2. Qualitative Data Collection Methods

This method of data collection is through observations, interviews, questionnaires, surveys, case studies, etc. in qualitative data collection mathematical calculations are not involved.

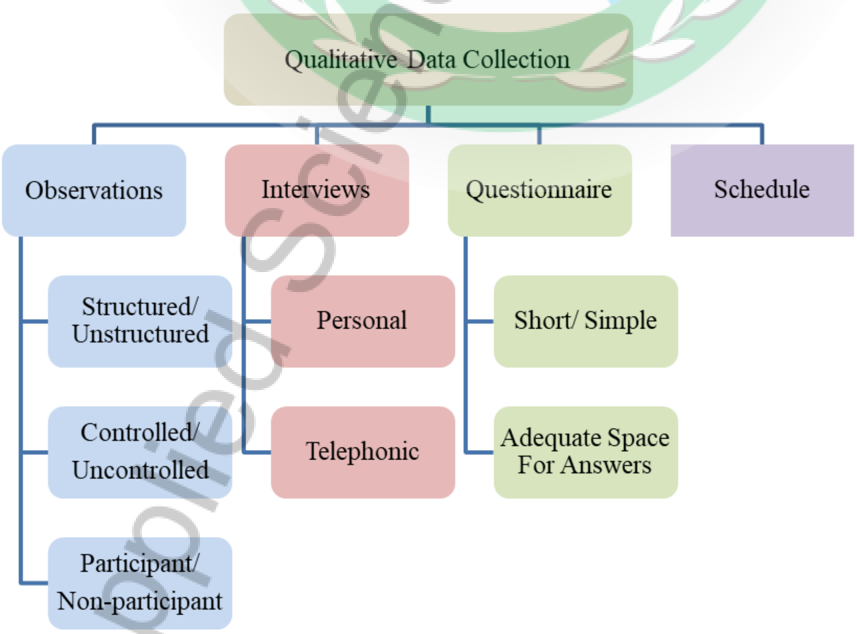


Figure 2.2. Various types of Qualitative Data Collection

2.2.2. SECONDARY DATA COLLECTION

It is collected from sources other than the actual source. Secondary data is already analyzed by someone and it is present in the form of books, journals, magazines, etc. This type of data contain published and unpublished materials.

Published data may contain

- Books, Journals, Public records, Historical documents, Government documents

Unpublished data my contain

- Unpublished biographies, Letters, Diaries

2.2.3. QUALITATIVE vs QUANTITATIVE DATA

	QUALITATIVE DATA	QUANTITATIVE DATA
1.	My friend is 5 feet and 7 inches tall	My friend has curly brown hair
2.	They have size 6 shoe size	They have large feet
3.	She weighs 62 kilograms	She has green eyes
4.	He has one older and two younger siblings	He is funny, loud and compassionate
5.	My best friend has two cats	My best friend is impatient and impulsive
6.	My aunt lives 20 miles from me	My aunt drives a red car
7.	We go swimming 4 times a week	We have friendly faces and good laughs

2.2.4. PRIMARY vs SECONDARY DATA

	PRIMARY DATA	SECONDARY DATA
1.	Needs to be generated	Is readily available
2.	First-hand information	Second hand information
3.	Questionnaire	No need for questionnaires
4.	Analysis as per purpose	Descriptive
5.	Require more time and money	Less expensive
6.	Example:	Example:
	Surveys, focus group and interviews	Publications, websites and reports

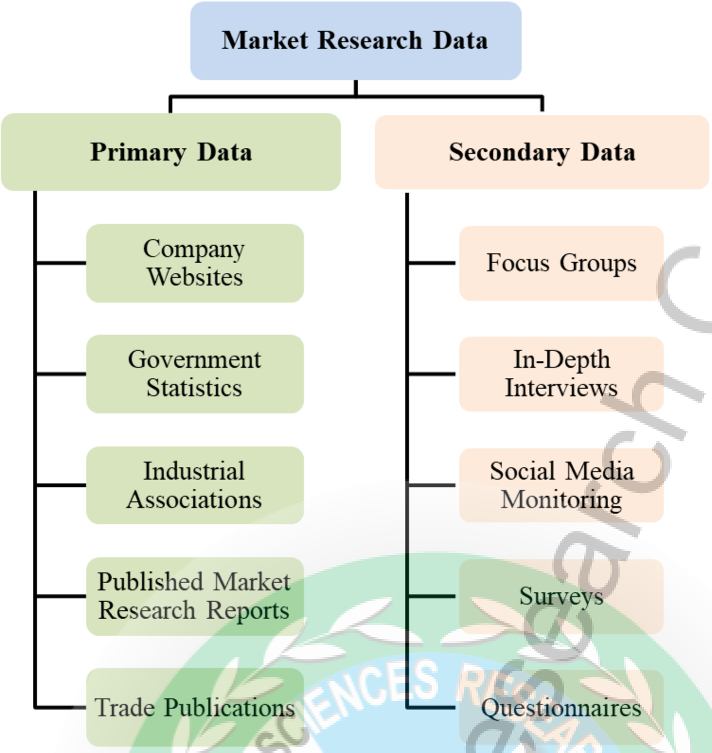
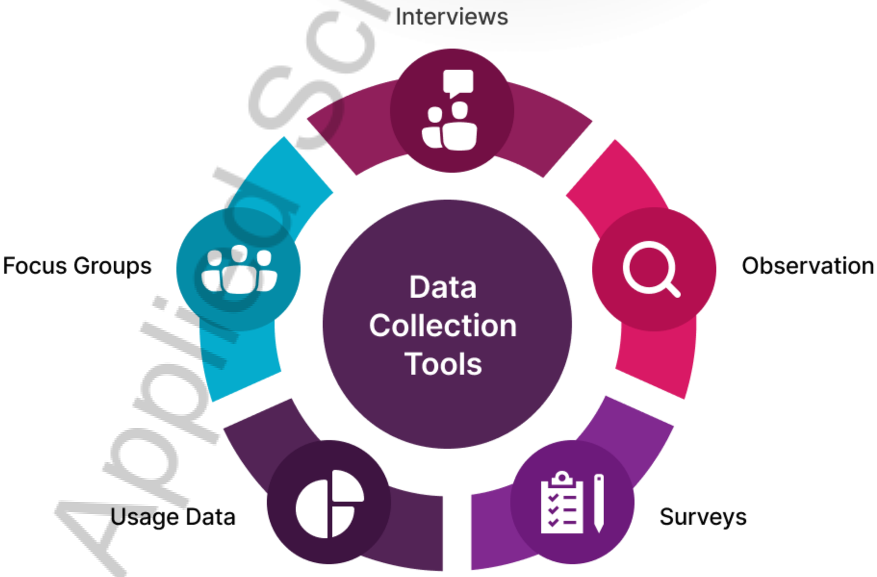


Figure 2.3. Primary vs secondary data collection for market research

2.2.5. DATA COLLECTION TOOLS

Instruments used to collect data are called data collection tools like papers for questionnaires or virtual systems for checklists, interviews, etc.



2.2.6. DATA ORGANIZATION AND PRESENTATION

Data is the collection of information, figures, numbers which is collected on the base of observations and research. Organization of data is in the form of tables, charts, pictures, graphs, etc. Data organization is the classification and categorization of the data to make it more understandable and useable. Documents or data are arranged in order and logical manner that anyone can easily takes the information from it. Data arrangement is important because you can easily access and approach the data. This plays important role in successful research work. It organizes the raw data by categorizing and classifying them into an understandable manner. Following points are important in data organization

- Name the data files in clear and consistent ways.
- Descriptive naming is very helpful in data organization.
- Title should be short and date format should be consistent.

2.2.6.1. Table

Tables are the most convenient method for data analysis in which larger data is organized in coordinated way. The data in tables are arranged in columns and rows which are easy to read and understand. Organization of data includes tables which are analyzed in four following forms:

- Chronological data which is arranged according to time i.e. days, weeks, months, years, etc.
- Spatial data is arranged according to space or geographical locations like towns, cities, districts, countries, etc.
- Qualitative data is categorized according to various features like gender, marital status, nationality, age group, language, social status, etc. these attributes cannot be measured but only classified.
- Quantitative data is measured in numbers like age, weight, height, etc.

2.2.6.2. Tabulation

Tabulation is a systematic and logical representation of numeric data in rows and columns to facilitate comparison and statistical analysis. It facilitates comparison by bringing related information close to each other and helps in statistical analysis and interpretation. The objectives of tabulation are

- Simplify complex data

- Facilitate comparison
- Save space
- Helpful for statistical analysis
- Highlights key features of data

While drawing table few points, one must be keep in mind

- Units should be written in cell heading and avoid writing in body of table
- Independent variables must be placed in 1st column
- Subsequent column must have information of dependent variable
- Decimal places must have a uniform pattern throughout column
- Error of instrument or accuracy must appear in the heading of cell

Title of the table must be explanatory for example the amino acid producing bacteria are isolated from different industrial areas of Lahore. The following table clearly described the number of bacterial strains isolated from various sources.

Location	Source	Total no. of bacterial isolates	Producers	Non-producers
Industrial area of Kot Lakhpat	Sewage	21	15	06
	Soil	35	26	09
Lahore Kasur Road Industrial area	Sewage	10	7	03
	Soil	23	18	05

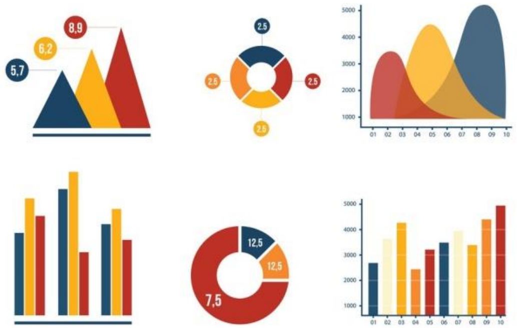
Table 2.1. Amino Acid producing and non-producing bacterial strains isolated from different sources

2.2.6.3. Graphs

The tabulation process makes it easy to analyze the data. The well tabulated data is presented in the form of pictures and graphs that are more eye catching and give comparison through better visual impressions. The presentation of quantitative data in form of charts or figures is the graphical representation. Graphs are more effective for analysis and comparison of the statistical data. The diagrammatic representation helps the individuals to conclude the meaning of data. Different types of graphs represent the data in form of charts, figures and diagrams.

Examples of graphs include;

- Pie chart



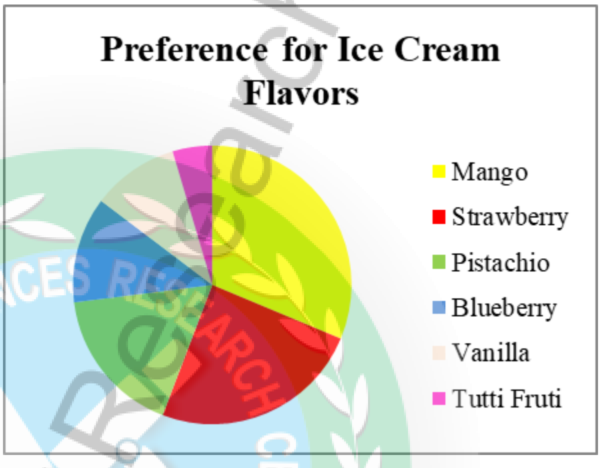
- Bar graph
- Histogram
- Frequency table
- Line graph
- Line plot

Example:

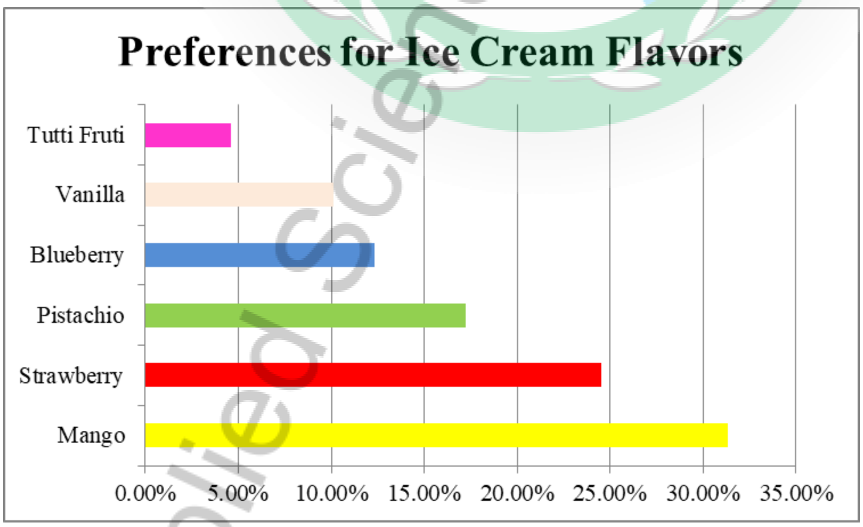
A survey was conducted on the preference of ice cream flavors in boys aged 13-18 years old. The different types of graphical data are represented below.

Ice Cream Flavor	Preference
Mango	31.3
Strawberry	24.5
Pistachio	17.2
Blueberry	12.3
Vanilla	10.1
Tutti Fruti	4.6

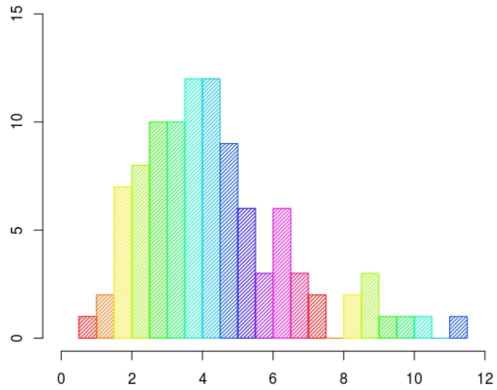
Data tabulation for
ice cream flavors preferences



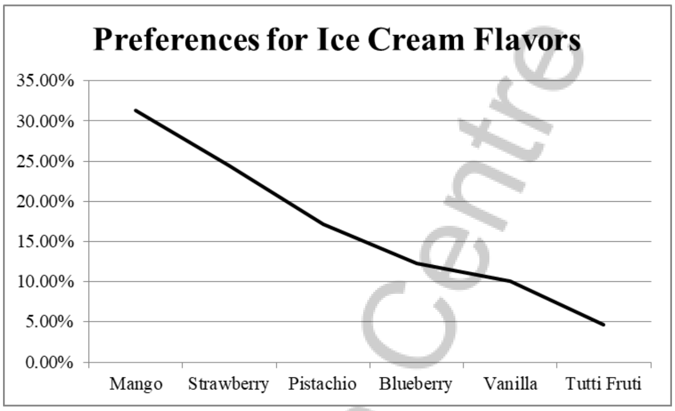
Pie chart illustrating
ice cream flavors preferences



Bar graph illustrating ice cream flavors preferences



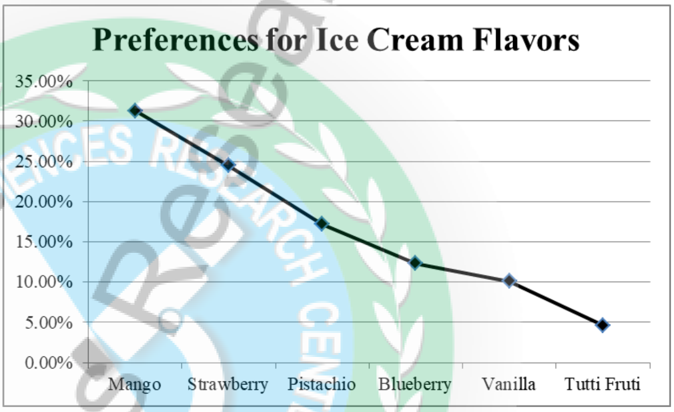
Histogram illustrating ice cream flavors preferences



Line graph illustrating ice cream flavors preferences

Ice Cream Flavors	Tally	Frequency
Mango	~~IIII~~ ~~IIII~~	10
Strawberry	~~IIII~~	5
Pistachio	IIII	4
Blueberry	III	3
Vanilla	II	2
Tutti Fruti	I	1

Frequency table illustrating ice cream flavors preferences



Line plot illustrating ice cream flavors preferences

2.2.6.4. Pie Chart

Also known as circle graph in which the circle is supposed with 100% and the variables occupied with specific percentage. In pie chart circle is divided into areas representing the relative frequencies of a category or variable. For example a class of genetics contain total 75 students out of which 25 are males and 75 are females. By calculating the relative frequencies we can draw a pie chart.

Genetics Class	Frequency	Relative Frequency
Females	50	50x100= 67% 75
Males	25	25x100=33% 75
Total	75	100%

Table 2.2. Strength of students in genetics class

2.2.7. PRACTICAL GUIDE TO DATA PRESENTATION

2.2.7.1. Key to make pie chart in Excel

Enter data in excel sheet → on top of menu click on insert → select pie chart → select design→ graph appear →label axis

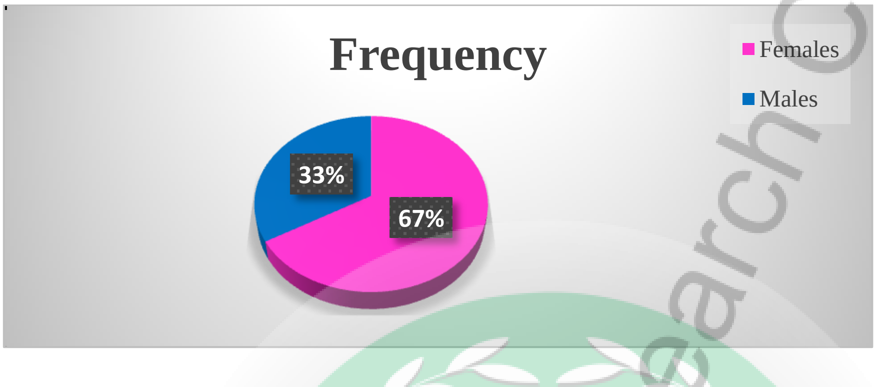
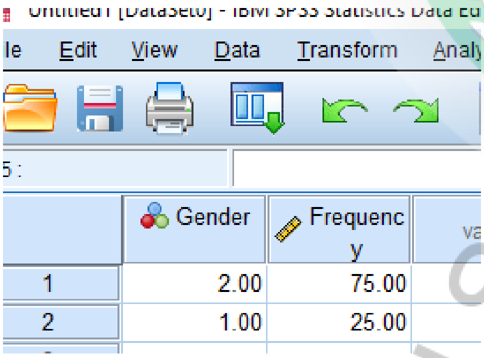


Figure 2.4. Pie chart depicting relative frequencies of male and female students

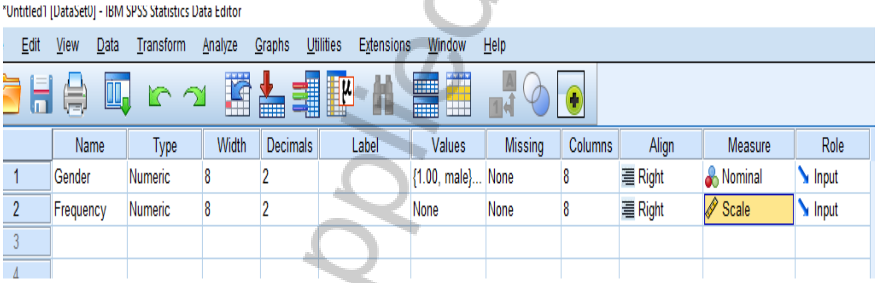
2.2.7.2. Key to make pie chart in IBM-SPSS

Open data sheet → enter data in data view→ open variable view →label the variables →label gender→ go the values and add values 1= male and 2= female → go to chart builders → select pie chart →drag pie chart → drag frequency and gender to pie chart → label → click ok.



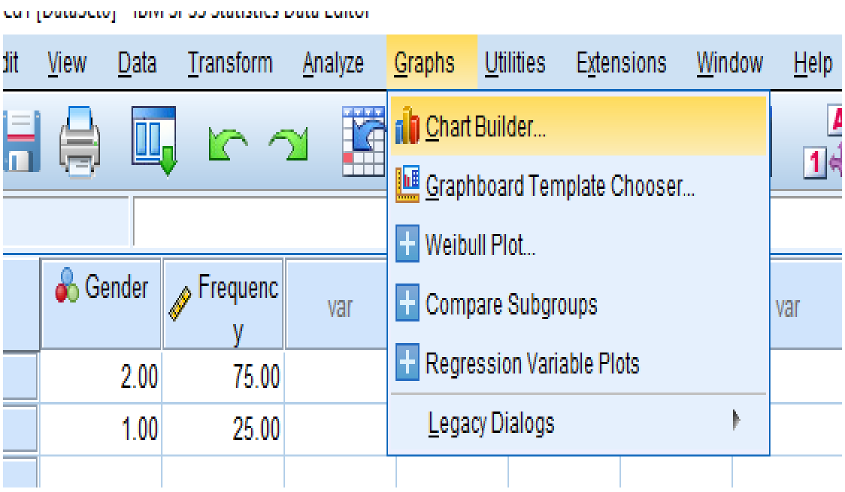
	Gender	Frequency
1	2.00	75.00
2	1.00	25.00

Step 1:
Data Entry



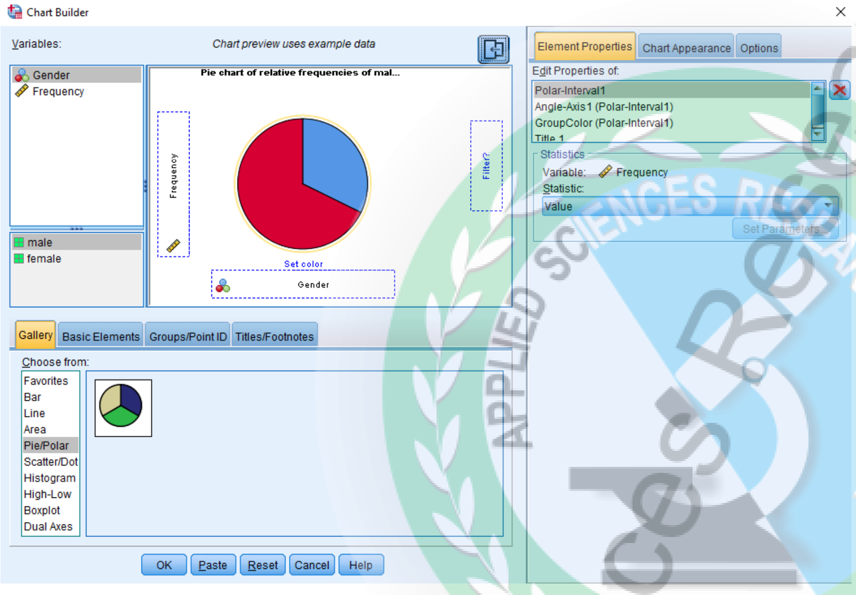
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	8	2		{1.00, male}...	None	8	Right	Nominal	Input
2	Frequency	Numeric	8	2		None	None	8	Right	Scale	Input
3											

Step 2:
Data labeling in variable view

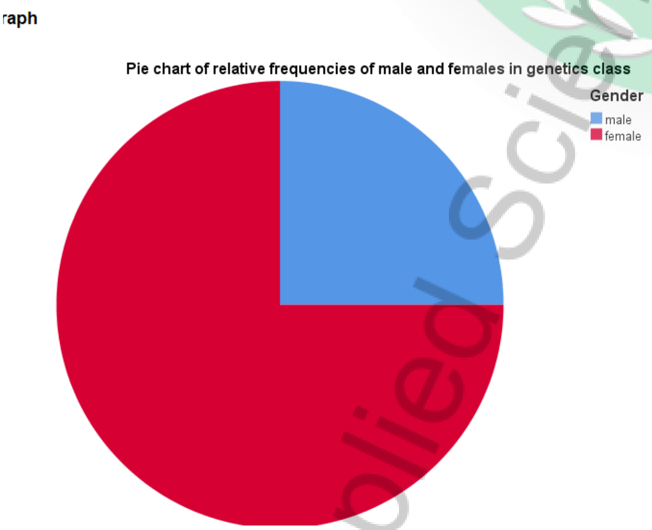

dit View Data Transform Analyze Graphs Utilities Extensions Window Help
Chart Builder...
Graphboard Template Chooser...
Weibull Plot...
Compare Subgroups
Regression Variable Plots
Legacy Dialogs
Gender
Frequency
var
var

Gender	Frequency
2.00	75.00
1.00	25.00

Step 3:
Select Graph and click on chart builder



Step 4:
Select pie chart and label



Step 5:
Chart appears on screen

2.2.7.3. Bar Graph

Bar Graph or columnar graph is the simplest form of representing the categorical data. On X-axis the variables are present whereas Y-axis represents the frequencies. Width of all columns are same and the distance or intervals between two columns must be the same. Bar graphs can be simple or double. In simple bar graph the number of students during year 2020 can be tabulated and simple bar graph is constructed showing student frequencies on y-axis and month on x-axis.

Months	No. of Students in Biology Coaching Classes
January	30
February	50
March	43
April	33
May	41
June	52
July	40
August	39
September	38
October	44
November	37
December	35

Table 2.3. No. of students who joined biology coaching class in the year 2020

2.2.7.4. Key to make simple bar graph in Excel

Enter data in excel sheet → on top of menu click on insert → select bar cart → select design→ graph appear →label axis

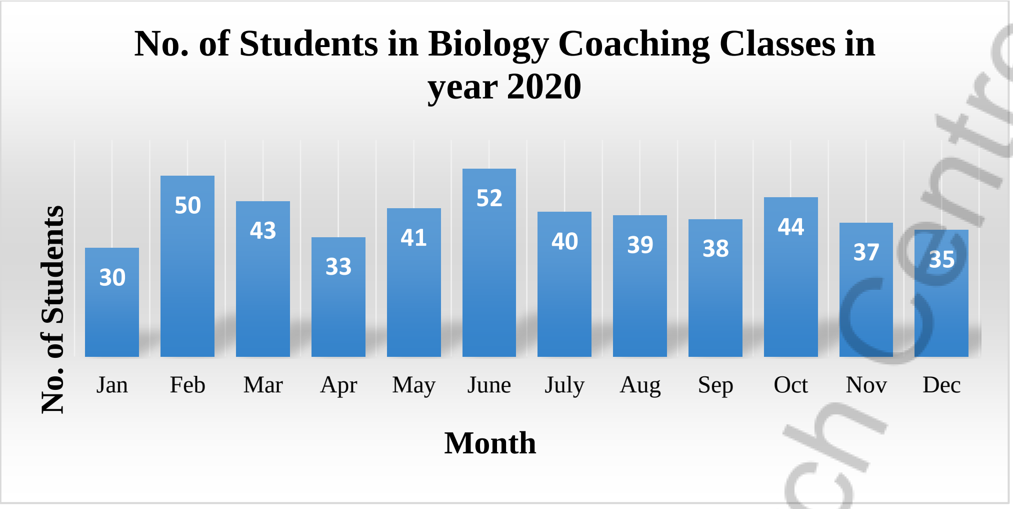
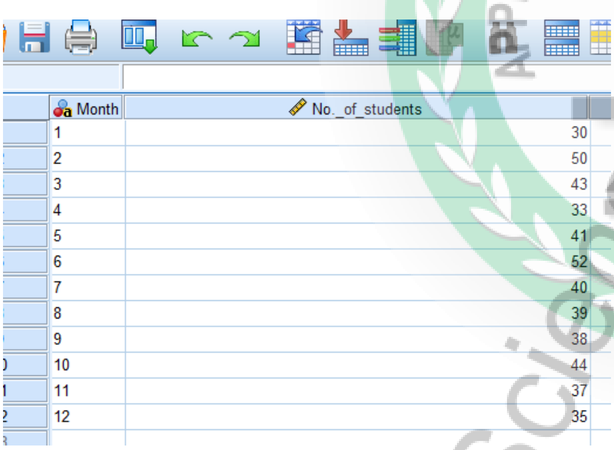


Figure 2.5. Simple bar graph showing No. of students who joined biology coaching class

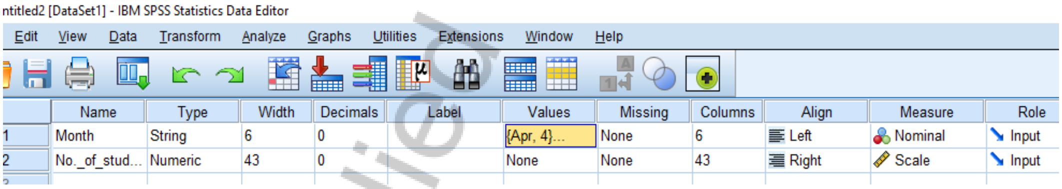
2.2.7.5. Key to make simple bar graph in IBM-SPSS

Open data sheet → enter data → open variable view → label months →go to graph builder → select simple bar graph → drag variables on axis → click OK



	Month	No._of_students
1	1	30
2	2	50
3	3	43
4	4	33
5	5	41
6	6	52
7	7	40
8	8	39
9	9	38
10	10	44
11	11	37
12	12	35

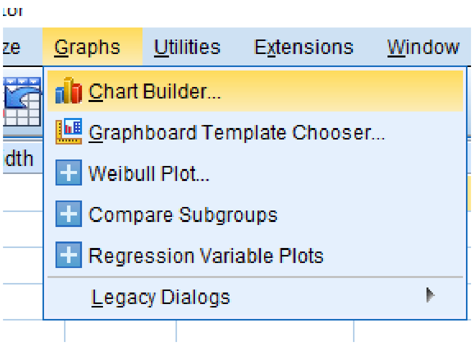
Step 1:
Data Entry


Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

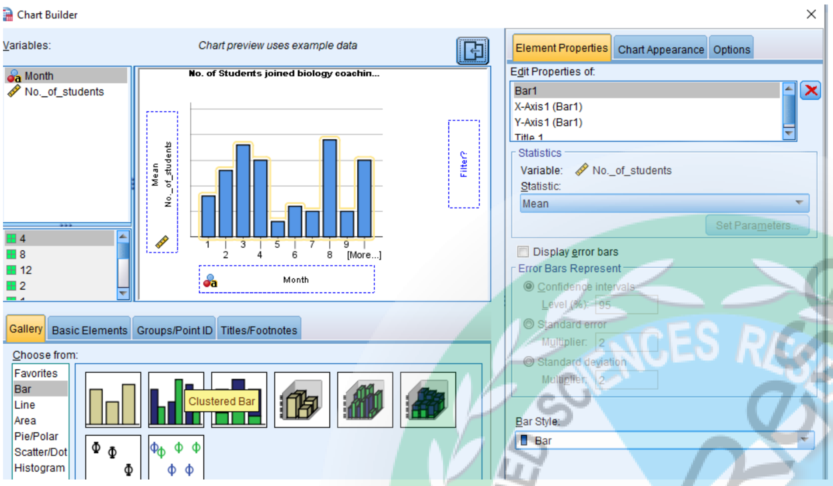
ntitled2 [DataSet1] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Month	String	6	0		{Apr, 4}...	None	6	Left	Nominal	Input
2	No._of_stud...	Numeric	43	0		None	None	43	Right	Scale	Input

Step 2:
Data Entry in variable view

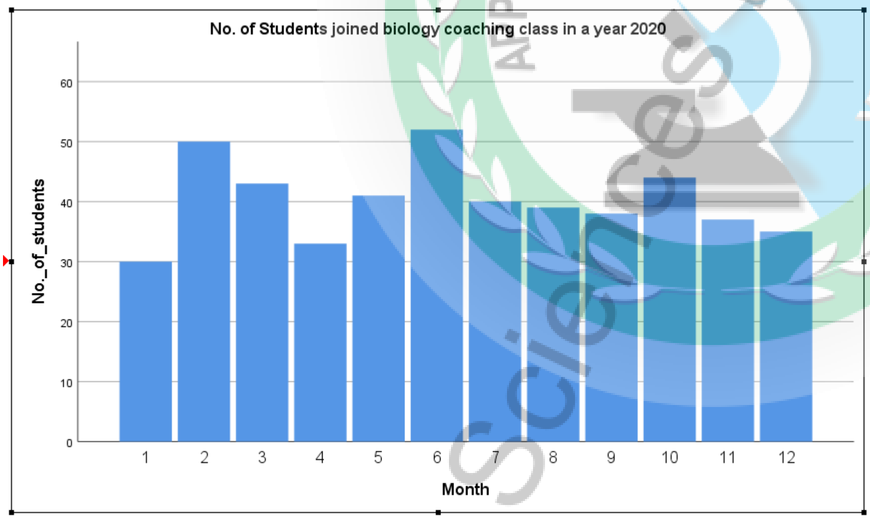


Step 3:
Click on graph and select Chart Builder



Variable: No._of_students
Statistic: Mean
Bar Style: Bar

Step 4:
Select Bar Graph



Step 5:
Graph appear on screen

2.2.7.6. Double Bar Graph

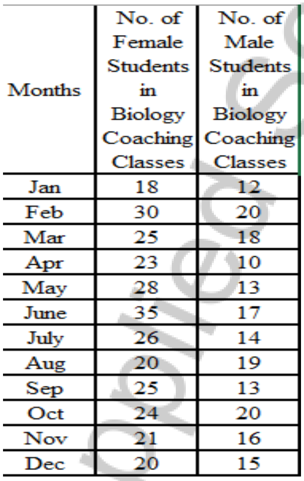
Double bar graph is formed when one set of variable has two components or various variables have one component. In this graph different components of a variables are represented in double bar. For example the no. of students can be further sorted by male and females and double bar graph is constructed through gender variable.

Months	No. of Female Students in Biology Coaching Classes	No. of Male Students in Biology Coaching Classes
January	18	12
February	30	20
March	25	18
April	23	10
May	28	13
June	35	17
July	26	14
August	20	19
September	25	13
October	24	20
November	21	16
December	20	15

Table 2.4. Male and female students who joined biology coaching class in year 2020

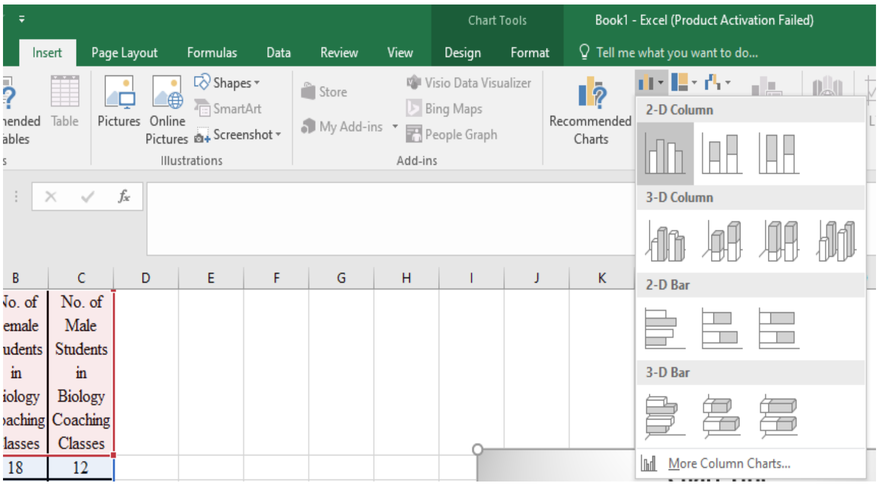
2.2.7.7. Key to make double bar graph in Excel

Enter data in excel sheet → on top of menu click on insert → select bar chart → select design→ graph appear →label axis

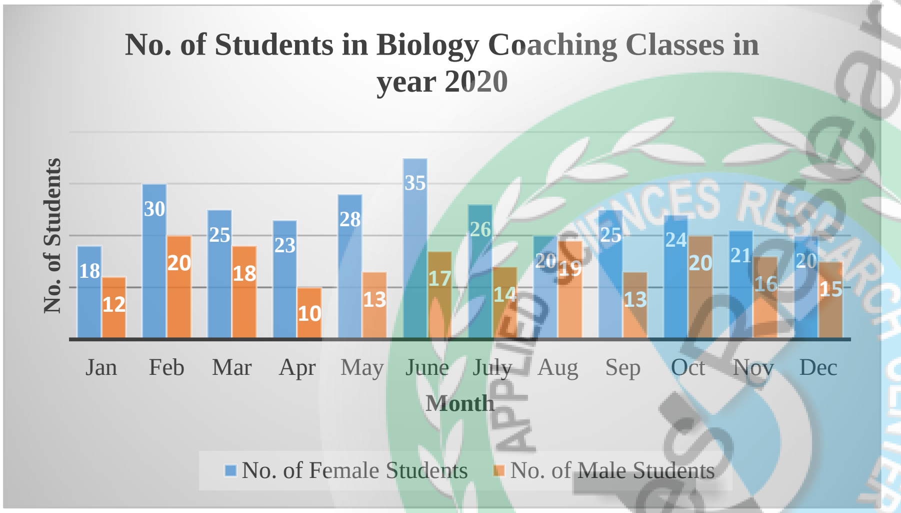


Months	No. of Female Students in Biology Coaching Classes	No. of Male Students in Biology Coaching Classes
Jan	18	12
Feb	30	20
Mar	25	18
Apr	23	10
May	28	13
June	35	17
July	26	14
Aug	20	19
Sep	25	13
Oct	24	20
Nov	21	16
Dec	20	15

Step 1:
Data Entry in excel sheet



Step 2:
Click on insert and select bar chart



Step 3:
Chart appear

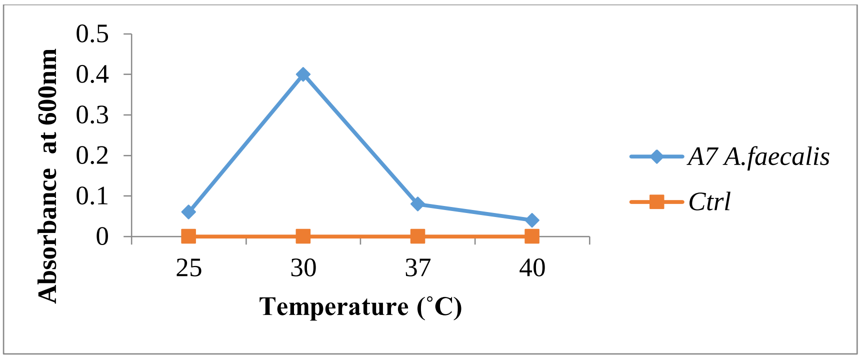
Figure 2.6. Double bar graph showing No. of male and female Students joined biology coaching class in year 2020

2.2.7.8. Line Graph

Line Graph is plotted to display continuous data and is used to predict future data. For example some bacteria show maximum growth at a specific temperature so by plotting line graph their optimum temperature can be predicted. The one variable is taken on x-axis while other variable on y-axis. Following graph is showing effect of temperature on growth of bacteria.

2.2.7.9. Key to make line graph in Excel

Enter data in excel sheet → on top of menu click on insert → select line chart → select design→ graph appear →label axis



Step 4: Line Graph in Excel

Figure 2.7. Line graph representing Optimum Temperature of A7 *Alcaligened faecalis*

2.2.7.10. Histogram

Histogram is the 2-D frequency density graph and it represents the class intervals and frequency in rectangle form. A histogram presents the categorical variables in bins which shows data point within range. First you form best fitted data intervals or range. The height of the bar represents the frequency. Histograms are of two types

i. With class intervals

ii. Without class intervals

Following steps are important in drawing histogram

- Class intervals should be exclusive and if intervals are inclusive then make them in exclusive form
- Class intervals make the base of rectangles whereas frequencies make the heights.
- Equal intervals represents the height of each rectangle is proportional to the class frequency.

Histogram can be formed when data showing class interval and their corresponding frequencies

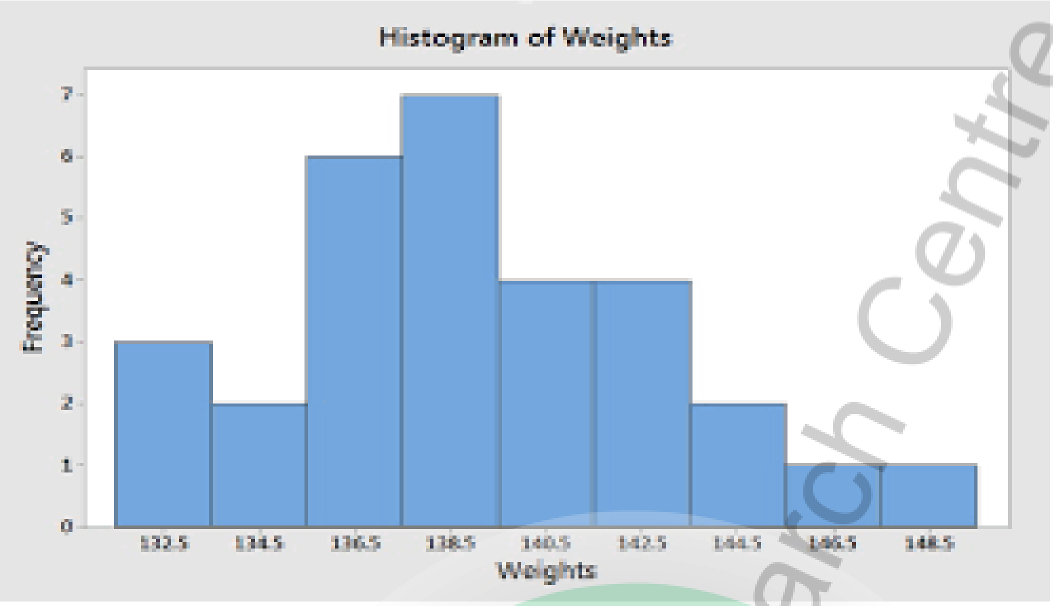


Figure 2.8. Histogram showing range of weights and its frequencies in a sample population

2.2.7.11. Frequency Table

Frequency table represents the number of pieces within the given interval in the given data. For example the height of individuals in a population form frequency table and it tells the ranger of heights within the population.

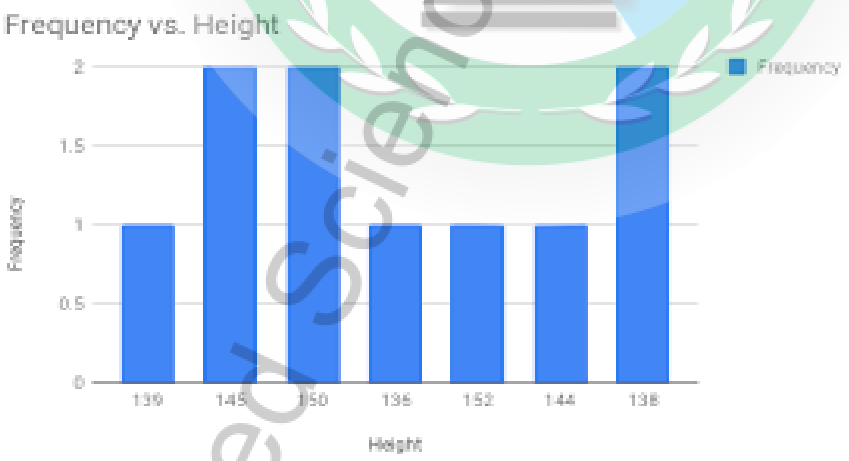
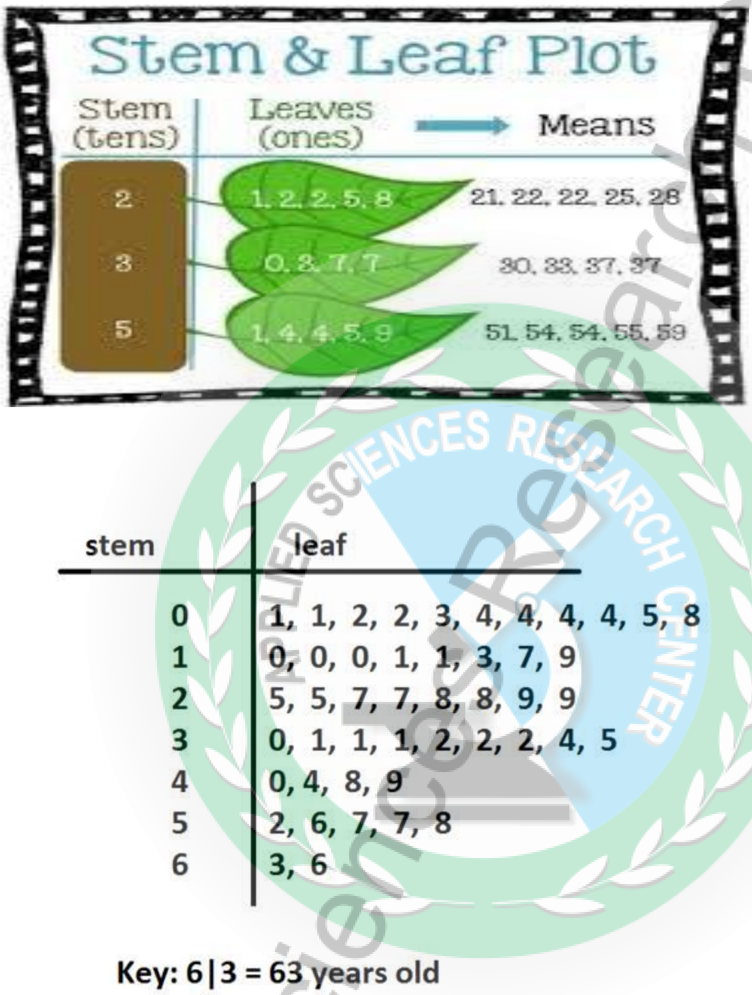


Figure 2.9. Frequency table of range of heights within a sample population

2.2.7.12. Stem and Leaf Plot

The data in stem and leaf plot, is organized in stem and leaf the tens in a figure represented in stem whereas the ones are represented in form of leaves



stem	leaf
0	1, 1, 2, 2, 3, 4, 4, 4, 4, 5, 8
1	0, 0, 0, 1, 1, 3, 7, 9
2	5, 5, 7, 7, 8, 8, 9, 9
3	0, 1, 1, 1, 2, 2, 2, 4, 5
4	0, 4, 8, 9
5	2, 6, 7, 7, 8
6	3, 6

Key: 6|3 = 63 years old

Figure 2.10. Stem and Leaf Plot

CHAPTER THREE

Descriptive Statistics

3.1. INTRODUCTION

The data presented and complied in form of tables and graphs then the next step is various measurements of the data for statistical analysis. The qualitative and quantitative data are differently analyzed through statistical means. Qualitative data is statistically analyzed through proportion, ratio, indices, percentage, test of independence, ranks and association whereas the quantitative data is analyzed through percentage, average, correlation, regression, variations, indices, analysis of variance, etc.

3.1.1. RATES

It is the particular (n) event that occurs in a specific population at a specific period of time. Following formula can be used to calculate the rate

$$\text{Rate} = \frac{n\ (A)}{n} \text{ x } K$$

n (A)= number of individuals involved in an even

n= number of individuals exposed to an event

K= base unit and its value can be taken as 1, 100, 1000, 10000 or 100000

For example if we want to check the prevalence rate of a disease in a population follow formula will be used

$$\text{Rate}= \frac{\text{number of individuals encounter the disease at a given time}}{\text{Number of individuals exposed to the disease}} \text{ x } K$$

3.1.2. RATIO

It is the comparison between two numbers that reflect the relation between these quantities. For example the gender ratio can be calculated as follow

$$\text{Gender Ratio} = \frac{\text{Number of females}}{\text{Number of males}}$$

3.2. MEASURES OF CENTRAL TENDENCY

It is the relation of the data set with its average value, as it is located in the middle so it is called as central value. Central tendency is the representative point of the data set and used in comparison between two or more data sets. It is also known as average and has two types. Central tendency can be defined as single value statistical representation of entire dataset.

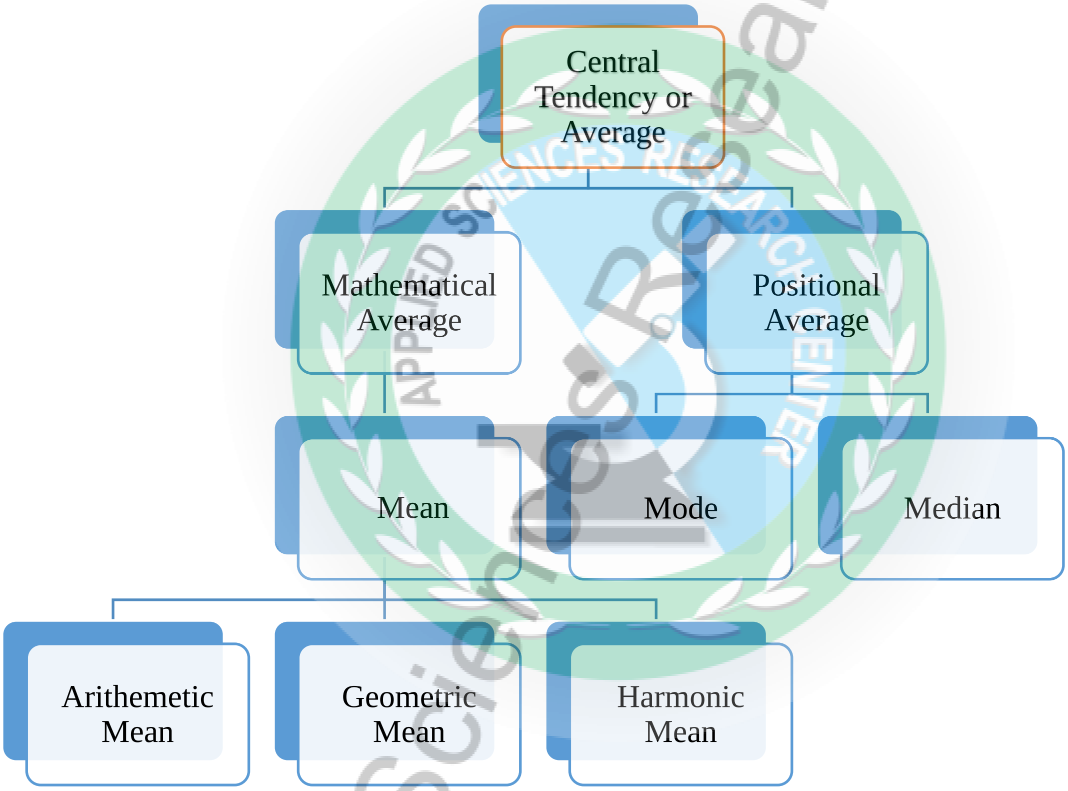


Figure 3.1. Types of Central tendency or Average

3.2.1. MEAN

It is the mathematical average value of the data and can be calculated by sum up all the dataset values and divided it with total number of values. For example biology class students' scores in a test were 75, 84,63,79,82, 91, 72. Following formula will be used to calculate the mean value:

$$\text{Mean} = \text{Sum of observations/ Total no. of observations}$$

$$= \frac{x_1+x_2+x_3+x_4+\ldots\ldots+x_n}{n} = \frac{\Sigma x_i}{n}$$

Whereas, n= Total number of values or observations

$$\text{Mean} = \frac{75+84+63+79+82+91+72}{7}$$

$$\text{Mean} = 73.7$$

Mean is further divided into three types

- Arithmetic mean
- Geometric mean
- Harmonic mean

It has been observed that if all values in dataset are same then all arithmetic, geometric and harmonic mean will be the same, different values in dataset will result in different mean values.

3.2.2. MODE

It is the most frequent number of the dataset. Mode is not an effective type of positional average because sometime no mode is present in a dataset or more than one mode are present in dataset. For example test scores of English class students are 7,8,9,7,8.5,6,10,7. As it can be seen that 7 is the most frequent number and it is the mode of above data. If there are two mode in data then average of the mode can be taken and it is called as average mode. For example the test score of biology class students are 6,7,7,7,8,8,9,9,9,10 so 7 and 9 are the mode so the average mode will be 7+9/2= 8

3.2.3. MEDIAN

It is the class midpoint or the middle value of the data set. If the dataset is given in the odd numbers then the middle of the dataset is the median. For example in the following given numbers **2,4,5,7,8,10,**11,**13,14,16,19,21,24** median is 11 because it is dataset midpoint. If the data set contain even numbers then the middle two values are taken and the average of these values can be median of dataset. For example 2,3,4,5,6,**7,8**,9,10,11,12,13 the median will be average of 7 and 8 i.e. 7.5. Following formula used to calculate median:

$$\text{Median} = 1/2\ [(n/2) + (n/2 + 1)]^{\text{th}}\ \text{observation}$$

3.3. MEASURES OF DISPERSION

It is used to measure the individual observations disperse around average. Central tendency or average value fails to describe the dispersion without variation of the observed data. Range and standard deviation are the measures of dispersion.

3.3.1. RANGE

It is the difference between the lowest and highest values of data like marks of students in particular subjects, levels of sugar in blood, blood pressure, etc. Following formula is used to calculate range:

$$\text{Range} = \text{Highest value} - \text{Lowest value}$$

Example:

Consider the following data to calculate the mean mode, median and range:

80, 74, 73, 88, 79, 94, 73, 65, 77, 90, 80, 69, 65, 89, 85, 53, 47, 61, 27, 80

Total No. of observations (N/n) = 20

Mean = Sum of observations/ Total no. of observations

= (80+74+73+88+79+94+73+65+77+90+80+69+65+89+85+53+47+61+27+80)/20

Mean = 72.5

Median = 1/2 [(n/2) + (n/2 + 1)]th observation

= 1/2 [10+11]th observation

= 1/2 [90+80]

= 85

Mode is the most frequent observation which is 73 in the given data.

Range = Highest value – Lowest value

= 94- 27

= 67

3.3.2. STANDARD DEVIATION

It is the most widely used type of dispersion and can be defined as square root of variance. Standard deviation is used to find the values of dispersed data, in other words it is the measurement of deviation of data from an average mean. It is denoted as SD and its lower value indicates that it is very close to their average. Higher values of SD reflects that values of the data are far from mean value. Its values are always positive and can never be in negative. Following formula is used to calculate the SD.

$$\text{Population SD} \qquad \sigma = \sqrt{\frac{\sum(X-\mu)^2}{n}}$$

$$\text{Sample SD} \qquad s = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$$

Whereas,

σ = Standard Deviation

x_i = Observations in the Data

x̄ = Mean

n = Total number of observations

3.3.3. STANDARD ERROR

In a sample the measure of uncertainty is called as standard error. It is one of the most used mathematical tool in biostatistics to estimate the variability. Standard error is denoted ad SE. Following formula is used to calculate the SE

$$SE_{\overline{x}} = \frac{S}{\sqrt{n}}$$

S is the standard deviation and n is the observation numbers.

Example:

In a survey 8 students were asked that how much time per day they spend on an average to study. The answers were 2, 4,3,2,4,5,6,3. Calculate the standard deviation and standard error.

Solution: Mean (x̄) = 2+4+3+2+4+5+6+3/8 = 3.1

Firstly make a table

x_1	$x_1 – \bar{x}$	$(x_1 – \bar{x})^2$
2	-1.1	1.2
4	0.9	0.81
3	-0.1	0.01
2	-1.1	1.2
4	0.9	0.81
5	1.9	3.61
6	2.9	8.41
3	-0.1	0.01

$$\Sigma(x_1 – \bar{x})^2 = 16.1$$

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

$$= \sqrt{16.1/8\text{-}1}$$

$$= \sqrt{2.1}$$

$$= 1.4$$

Following formula is used to calculate standard error of mean:

$$SE_{\bar{x}} = \frac{S}{\sqrt{n}}$$

$$= 1.4/\sqrt{8}$$

$$= 0.5$$

3.4. RELATIVE MEASURES

It is used for the comparison of dataset and it is free of unit. For example if we want to compare the two data sets of same time but he measuring units are different like weight can be measured in kilogram and pound then relative measures can be used. There are two types of relative measures.

a- Coefficient of variation

b- Z-score

3.4.1. COEFFICIENT OF VARIATION

When the two or more data sets have different units then coefficient of variance is used for comparing the datasets. It is denoted by C.V. Following formula is used to calculate C.V.

$$\text{C.V.} = \frac{\text{Sample standard deviation}}{\text{Sample mean}} \text{ x } 100$$

$$\text{Sample S.D} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

$$\text{C.V.} = \frac{\text{Population standard deviation}}{\text{Population mean}} \text{ x } 100\%$$

$$\text{Population S.D} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}}$$

Example:

Calculate the coefficient of variation of the following dataset:

1, 5, 6, 8, 10, 40, 65, 88

Sample mean = (1 + 5 + 6 + 8 + 10 + 40 + 65 + 88)/8 = 223/8 = 27.875

Variance:

$$\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{7578.875}{7} = 1082.696$$

Standard deviation:

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}} = \sqrt{1082.696} = 32.904$$

Coefficient of variation = 32.901/27.875=1.180

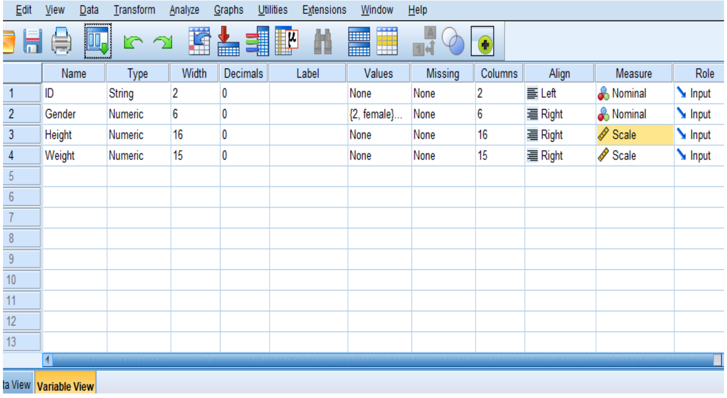
Example:

Calculate Mean, Mode Median, Standard deviation and variance from IBM-SPSS using the following dataset:

ID	Gender	Height in inches	Weight in pound
1	Male	64	128
2	Female	66	124
3	Male	70	153
4	Male	68	144
5	Female	62	136
6	Female	64	116
7	Female	64	122
8	Male	68	138
9	Female	66	151
10	Male	66	118

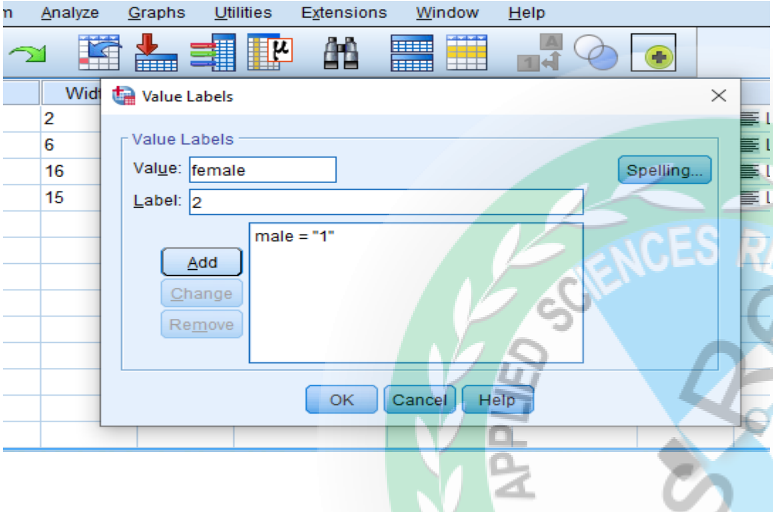
Solution key:

1. Enter the data in the data sheet of IBM-SPSS
2. Label the each variable
3. In the data sheet variable view assign no. 1 to males and 2 to females by clicking values and add label
4. Now click on Analyze on the top of the menu
5. Go to Descriptive Statistics and click on frequencies in sub menu
6. New window will open showing the variables now add the variable you want to analyze into the right side box by double clicking on the variable.
7. After adding the variables you want to analyze now click on statistics and add the central tendency and dispersion you want to calculate. Click ok and a window of desired statistics will be open.


Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
ta View
Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ID	String	2	0		None	None	2	Left	Nominal	Input
2	Gender	Numeric	6	0		{2, female}...	None	6	Right	Nominal	Input
3	Height	Numeric	16	0		None	None	16	Right	Scale	Input
4	Weight	Numeric	15	0		None	None	15	Right	Scale	Input
5											
6											
7											
8											
9											
10											
11											
12											
13											

Step 1:
Data Entry and labeling in variable view



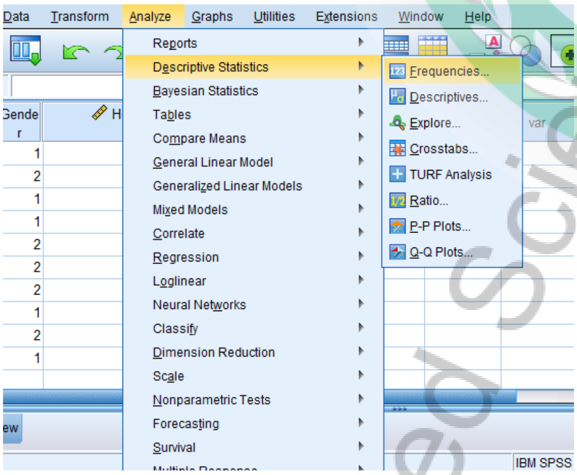
Value: female

Label: 2

male = "1"

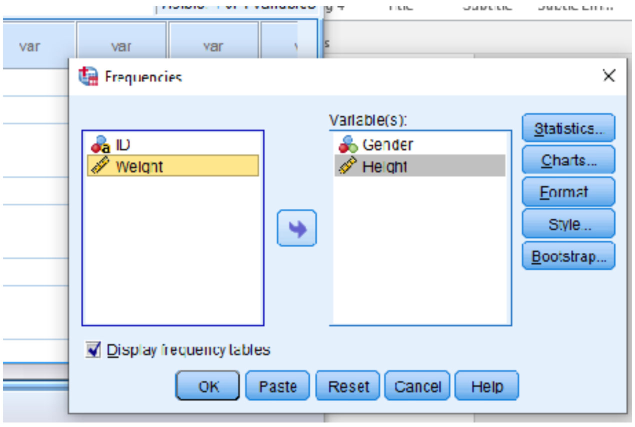
Step 2:
Value labels of Gender

] - IBM SPSS Statistics Data Editor



Gender
1
2
1
1
2
2
2
1
2
1

Step 3:
Click on analysis and select descriptive statistics and go to frequencies

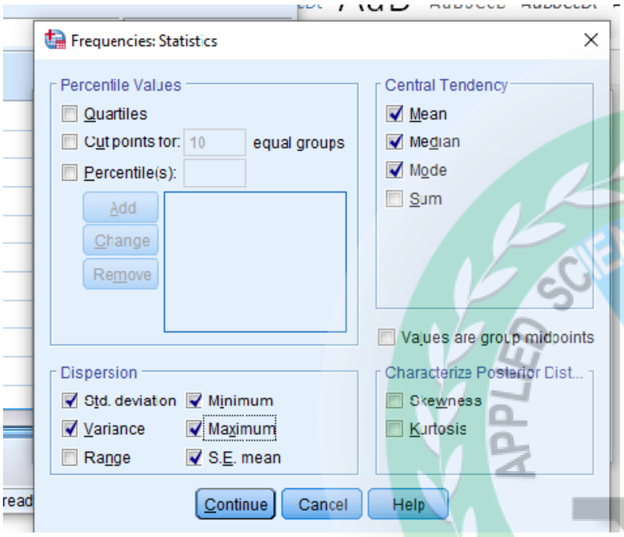


ID
Weight

Variable(s):
Gender
Height

☑ Display frequency tables

Step 4:
Add Variables



Percentile Values

☐ Quartiles
☐ Cut points for: 10 equal groups
☐ Percentile(s):

Central Tendency

☑ Mean
☑ Median
☑ Mode
☐ Sum

☐ Values are group midpoints

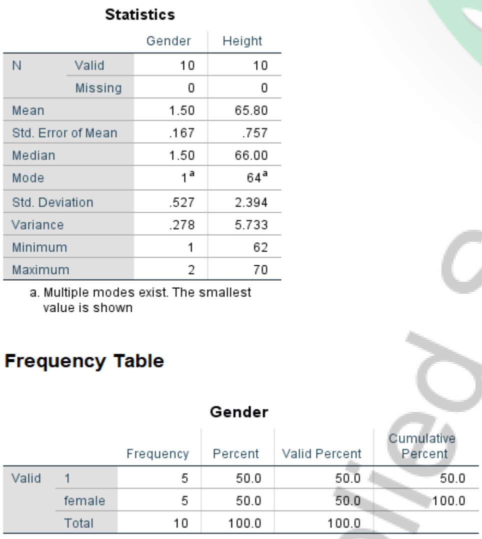
Dispersion

☑ Std. deviation ☑ Minimum
☑ Variance ☑ Maximum
☐ Range ☑ S.E. mean

Characterize Posterior Dist...

☐ Skewness
☐ Kurtosis

Step 5:
Select Central Tendency and dispersion



Statistics

		Gender	Height
N	Valid	10	10
	Missing	0	0
Mean		1.50	65.80
Std. Error of Mean		.167	.757
Median		1.50	66.00
Mode		1[a]	64[a]
Std. Deviation		.527	2.394
Variance		.278	5.733
Minimum		1	62
Maximum		2	70

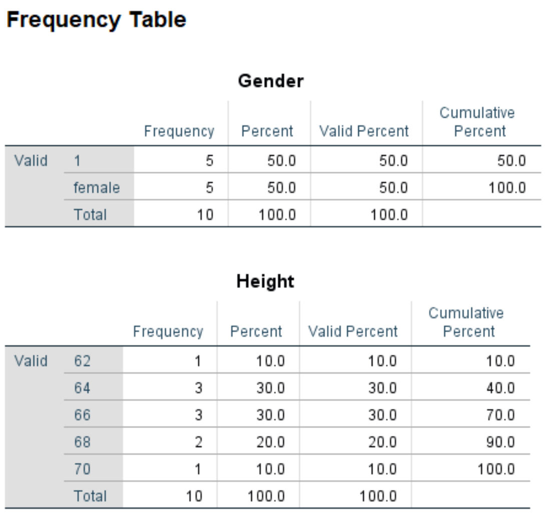
a. Multiple modes exist. The smallest value is shown

Frequency Table

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	50.0	50.0	50.0
	female	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

Step 6:
Values appear



Frequency Table

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	5	50.0	50.0	50.0
	female	5	50.0	50.0	100.0
	Total	10	100.0	100.0	

Height

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	62	1	10.0	10.0	10.0
	64	3	30.0	30.0	40.0
	66	3	30.0	30.0	70.0
	68	2	20.0	20.0	90.0
	70	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

Step 7:
Values appear

Example: Calculate central tendency and dispersion from the following grouped data.

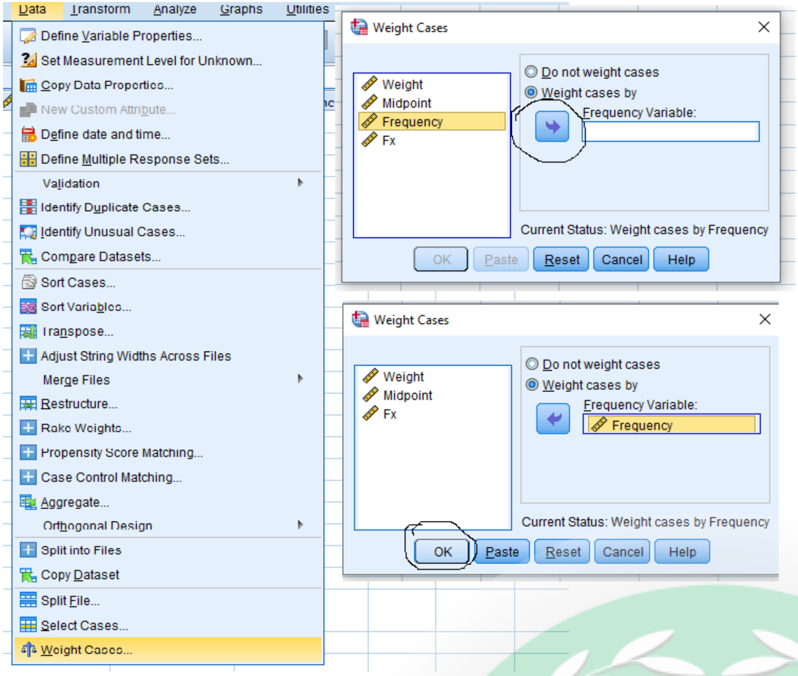
Solution: Calculate the mid-point and then enter the data in data sheet.

Weight (Pounds)	Mid-point(x)	Frequency (f)	fx
65-84	74.5	9	670.5
85-104	94.5	10	945
105-124	114.5	17	1946.5
125-144	134.5	10	1345
145-164	154.5	5	772.5
165-184	174.5	4	698
185-204	194.5	5	972.5

Solution Key:

Properly label the variables. Follow the given key by clicking on →

Data → Weight cases → Frequency → weight cases by → add frequency variable → click OK



Weight Cases

○ Do not weight cases
● Weight cases by
Frequency Variable: Frequency

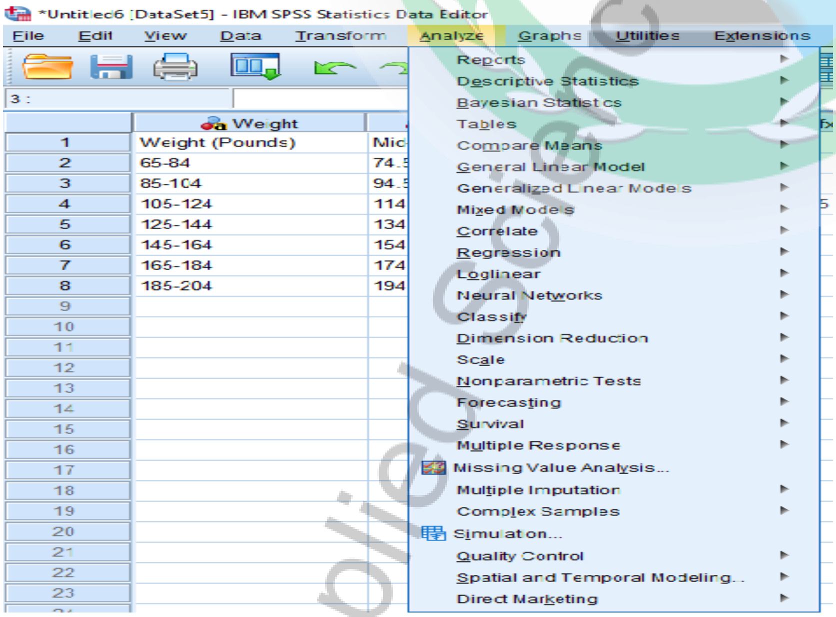
Current Status: Weight cases by Frequency

Step 1:
Go to data and select weight cases. Add weight cases

Now

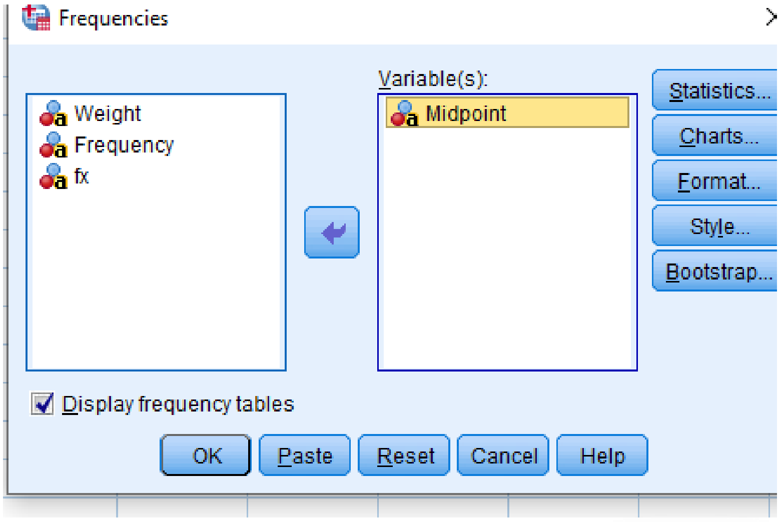
Click Analyze → Descriptive statistics → Frequencies → Shift/add mid-point to the variable → click statistics → Add Central tendency → Add dispersions

→ Click Ok → Values will appear.



	Weight	Mic
1	Weight (Pounds)	Mic
2	65-84	74.5
3	85-104	94.5
4	105-124	114
5	125-144	134
6	145-164	154
7	165-184	174
8	185-204	194

Step 2:
Click on Analyze and select descriptive statistics. Go to frequencies



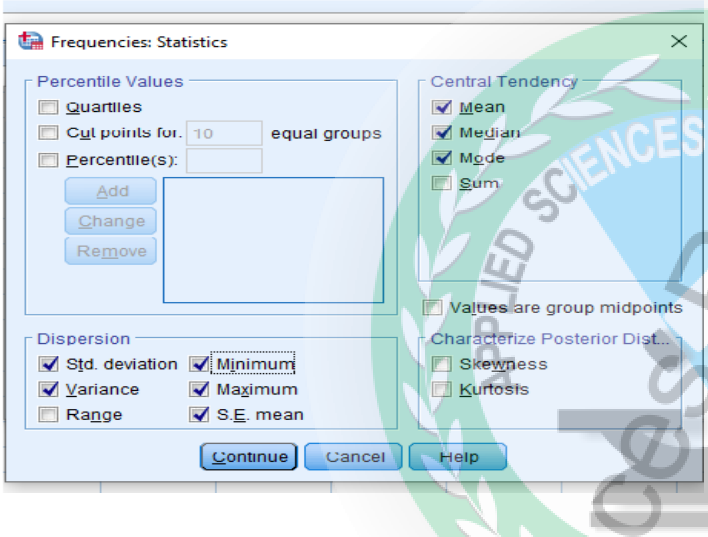
Frequencies

Variable(s):
- Midpoint

- Weight
- Frequency
- fx

☑ Display frequency tables

Step 3:
Add Variables



Frequencies: Statistics

Percentile Values
- □ Quartiles
- □ Cut points for: 10 equal groups
- □ Percentile(s):

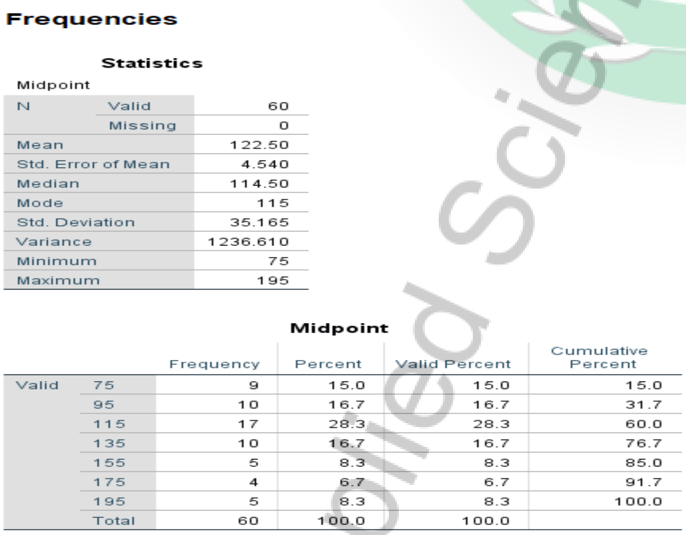
Central Tendency
- ☑ Mean
- ☑ Median
- ☑ Mode
- □ Sum

□ Values are group midpoints

Dispersion
- ☑ Std. deviation
- ☑ Minimum
- ☑ Variance
- ☑ Maximum
- □ Range
- ☑ S.E. mean

Characterize Posterior Dist...
- □ Skewness
- □ Kurtosis

Step 4:
Click on Statics and select central tendency/dispersion



Frequencies

Statistics

Midpoint

N	Valid	60
	Missing	0
Mean		122.50
Std. Error of Mean		4.540
Median		114.50
Mode		115
Std. Deviation		35.165
Variance		1236.610
Minimum		75
Maximum		195

Midpoint

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	75	9	15.0	15.0	15.0
	95	10	16.7	16.7	31.7
	115	17	28.3	28.3	60.0
	135	10	16.7	16.7	76.7
	155	5	8.3	8.3	85.0
	175	4	6.7	6.7	91.7
	195	5	8.3	8.3	100.0
	Total	60	100.0	100.0	

Step 5:
Values appear

CHAPTER FOUR

Population and Sampling

4.1. POPULATION

It is a group of individuals, creatures, things, objects or any cases. It is the group of objects or anything on which conclusion is drawn. The numbers of population may be finite or infinite. Statisticians also called population “universe”. Example of population include bacterial fauna of a region, number of patients in an institution, birds of a city, etc.

The finite population includes total number of students in an institution, total houses in a society, number of persons in a city, etc. whereas the infinite population includes number of stars in sky. In statistics the population is group of items which exhibit certain characteristics.

4.2. SAMPLE

Sample is the tiny or small part of population which is studied to make conclusions about population. Proper sample truly defines the characteristics of population and it is the representative of it. For example if someone wants to study the soil quality of a specific area, a small soil portion is collected from that area for study and this small portion is known as sample.

4.3. SAMPLING UNITS

Sampling Units are the individuals of the population which cannot be further subdivided and these are used to draw inferences about population. For example to find average marks in a particular subject of a class, each student is a unit.

4.4. SAMPLING FRAME

Sampling Frame is used to identify each sampling unit. It is defines as the list, scheme or map used for the identification of sampling unit. It is represented in form of numbers. For example, list of students attending a class, list of patients in a surgical ward or list of hospital staff.

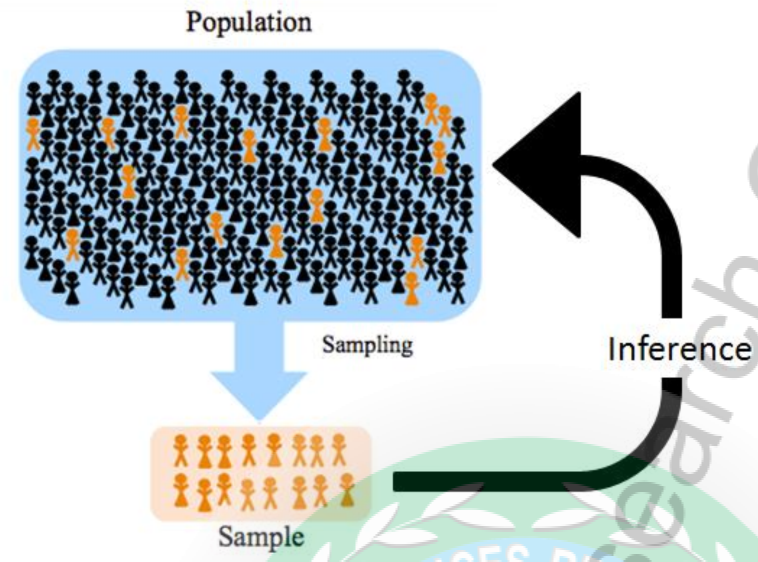


Figure 4.1. Difference between population and sample

4.5. SURVEY

It is an important step in sampling from population and a well-planned survey can be helpful in research. Various steps are part of conducting a successful survey.

4.5.1. OBJECTIVE OF SURVEY

Clearly describe the objective of survey this will clear the ongoing steps and it must be illustrated in a time frame and available resources.

4.5.2. DEFINE THE POPULATION TO BE STUDIED

The population which is going to be studied should be clearly defined. For example if you want to study the wheat field of an area then one must clearly define the size, shape, border, line, etc. of the field.

4.5.3. THE SAMPLING UNIT AND FRAME

Sampling units should be non-over lapping, clear and distinct. For example while conducting socio-economic study of a city, a person, a family, a block or a town can be sampling unit. Lists maps or schemes are the sampling frame of the study which determine the structure of the survey.

4.5.4. DATA COLLECTION

The data is collected according to the objectives of the survey. Too much irrelevant data can interrupt the survey. An outline is prepared in the form of table to collect the data for survey.

4.5.5. QUESTIONNAIRES OR SCHEDULE

After data collection the important step is the formation of questionnaires and schedule. The questionnaire is filled by the respondent while schedule is filled by the interviewer. In a survey the questions must be clear, to the point. Brief and in polite tone. The schedule preparation requires expertise, familiarity and special techniques on the subject so that more precise information can be collected.

4.5.6. METHODS OF DATA COLLECTION

There are various methods in collection information in a survey like interviews, mailed questionnaire, etc.

4.5.7. ANALYSIS OF DATA

The data collected from survey is summarized and analyzed. The data is first scrutinized and represented in form of tabulation for statistical analysis.

4.6. SAMPLE DESIGNING

The sample designing is comprises of work plan to collect sample from population.

4.7. SAMPLING ERRORS

There are two types of sampling errors in collection, process and analysis of sample.

1. Sampling error
2. Non-sampling error

4.7.1. SAMPLING ERROR

This type of error occurs when small part of population is utilized to study the population. The main reasons can be:

- Incomplete selection of population
- Wrong demarcation of sampling units
- Inappropriate statistical analysis of population parameters

4.7.2. NON-SAMPLING ERROR

This type of error occurs at the stage of processing of sampling data or material. It is present in complete enumeration and survey. Followings are the reasons of non-sampling errors:

- Inappropriate planning
- Faulty data compiling
- Responsive and non-responsive errors
- Errors in publication

4.8. TYPES OF SAMPLING

In research and study the sampling techniques are very important. The data can be collected through various sampling schemes which include.

- Subjective sampling
- Random sampling
- Mixed sampling

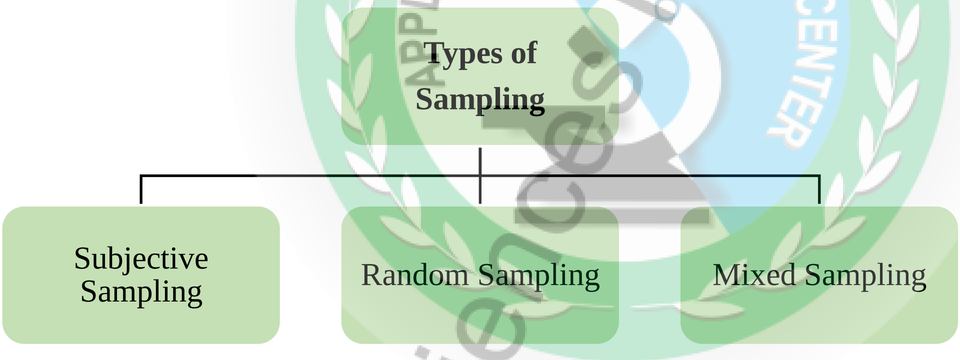


Figure 4.2. Various types of sampling

4.8.1. SUBJECTIVE SAMPLING

It is also known as purposive or judgement sampling. In this method sampling depends upon the decision of the sampler with a proper reason in mind. This sampling lacks true representation of the population because the selector select few units from population. So this type of sampling can be prejudice. For example if we want to study the TB patients we only sample from the TB wards in hospitals hence ignoring the TB individual sampling from population. This sampling method is used seldom and cannot be given preference because it targets certain units and ignore all other parameters.

4.8.2. RANDOM SAMPLING

It is also known as probability sampling and it is frequently used in scientific methods. Samples are taken in such a way that all units of population get equal chance of selection. There are different types of random sampling.

- Simple random sampling, when all units get equal chance of selection
- When units have different chances of selection
- When selection chances of units increases or decreases with sample size

4.8.2.1. Simple Random Sampling

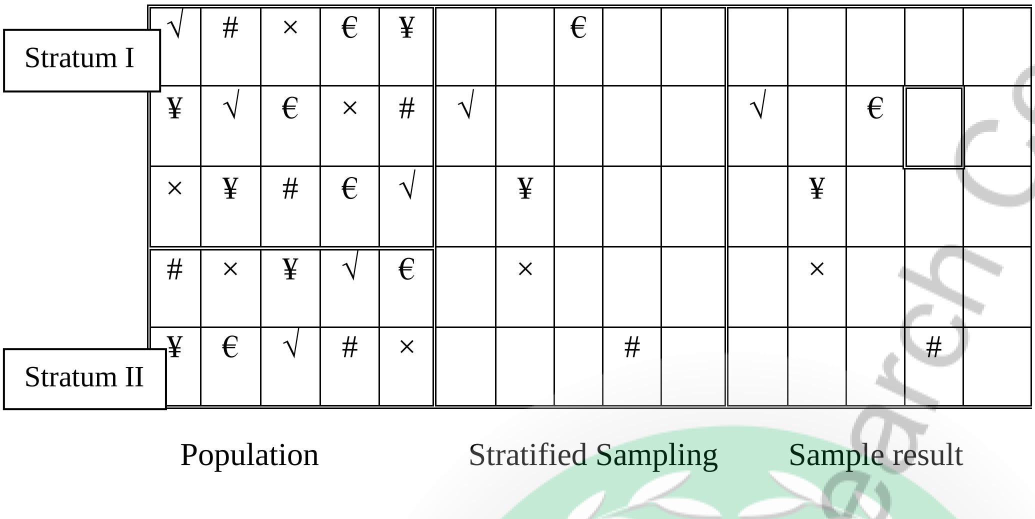
It is the most common mode of sampling in which all the units have equal chances of selection. For example if we want to study the bacterial fauna of soil then the samples of soil will be collected from surface layer, middle layer and a bit deeper layer of soil just to make sure that all the bacterial strains have equally collected in soil samples.

√	#	×	€	¥										
¥	√	€	×	#	¥					¥	×	#	€	√
×	¥	#	€	√			#	€						
#	×	¥	√	€		×		√						
¥	€	√	#	×										
Population					Random Sampling					Sampling result				

Figure 4.3. Simple random sampling

4.8.2.2. Stratified random sampling

In a heterogeneous population simple random sampling does not work. So firstly the population under certain criteria is divided into homogenous groups and these groups are known as strata. Random sampling from these strata is known as stratified random sampling. This type of sampling is very important in large heterogeneous population. For example if we want to study the average level of hemoglobin in a population it is better to divide the individuals according to age group and gender like infants, adult males, adult females and old ones. This strata formation will make it easy to estimate the average level of hemoglobin among individuals of different age groups in a population.



√	#	×	€	¥			€							
¥	√	€	×	#	√					√		€		
×	¥	#	€	√		¥					¥			
#	×	¥	√	€		×					×			
¥	€	√	#	×				#					#	

Figure 4.4. Stratified random sampling

4.8.3. MIXED SAMPLING

The samples are selected according to fixed sampling and probability sampling rules. So the sample is mixed and known as mixed sampling.

4.9. ADVANTAGES OF SAMPLING

Sampling is very important in research studies because it is not easy to study the whole population. Some significant advantages of sampling are:

1. Reduction of cost and time
2. Reduction in resource utilization
3. Intensive Data
4. Conclusion about population

4.9.1. REDUCTION OF COST AND TIME

Sampling not only reduces the time of research analysis but it also minimizes the cost of the research. If we want to study a specific disease in human population it is impossible to study all the individuals. So a small portion of the population is studies to draw inferences.

4.9.2. REDUCTION IN RESOURCE UTILIZATION

Sampling reduces the utilization of resources because while analyzing samples the numbers of objects have been reduced and fewer resources are utilized to study them.

4.9.3. INTENSIVE DATA

The sample data is intensive and thorough due to less number of respondents. Each respondent is studied thoroughly because of adequate time.

4.9.4. CONCLUSION ABOUT POPULATION

All the sample data is collected from a population so the results drawn from sample studies can be applied to the population.

4.10. ESTIMATION OF SAMPLE SIZE

In research studies the important point is estimation of population that is how large a population should be? It is crucial for researchers to determine the sample size in order to minimize the sampling error. Sample size is the choice of planner but it should always be determined with care because it is the portion of population so it should neither be too large to difficult to handle and nor too small to make it less reliable. The size of the sample depends upon the cost and time of the study. So the optimum sample size can minimize the sampling error.

4.10.1. ESTIMATION OF SAMPLE SIZE FOR ABSOLUTE PRECISION

The following formula is used to calculate sample size for absolute precision:

$$\mathbf{n = \frac{Z^2_{(1-\alpha/2)}\, p(1-p)}{d^2}}$$

d = difference between estimated and actual value
(It is usually taken as 5% or 0.05 in decimal terms)

p = proportion

For 95% probability or confidence level $Z^2_{(1-\alpha/2)} = 1.96$

For 90% probability or confidence level $Z^2_{(1-\alpha/2)} = 1.645$

For 98% probability or confidence level $Z^2_{(1-\alpha/2)} = 2.58$

Example:

The health department wanted to estimate the prevalence of Pneumonia among children below 5years age. How much the sample size should be so the prevalence must be estimated within 5% with 95% confidence level? The proportion or true rate must not exceed 15%.

Solution:

In example the p=0.15 and 1-p= 0.85

For 95% probability or confidence level $Z^2_{(1-\alpha/2)} = 1.96$

$d^2=0.05$

Therefore, according to formula

$$\mathbf{n = \frac{(1.96)^2(0.15)(0.85)}{(0.05)^2}}$$

n = 196 for 95% confidence level

Thus, the sample size will be 196 for 95% probability or confidence level

4.10.2. ESTIMATION OF SAMPLE SIZE FOR RELATIVE PRECISION

The following formula is applied to calculate the sample size for relative precision:

$$\mathbf{n = \frac{Z^2_{(1-\alpha/2)}\, p(1-p)}{d^2 p}}$$

Example: Health department wishes to estimate the prevalence of diabetes among persons of age above 60 years. From the past data the prevalence was estimated around 25%. How large the population sample should be with 95% confidence level?

p = 0.25 $\qquad$ 1-p = 0.75

d = 0.05 $\qquad$ at 95% confidence level Z2(1-α/2) = 1.96

$$\mathbf{n = \frac{(1.96)^2(0.75)}{(0.05)^2(0.25)} = 4610}$$

CHAPTER FIVE

Probability

5.1. INTRODUCTION

In an event “probability” can be defined as relative frequencies of particular occurrence in long run. Like, if an event happens X times out of N times, then probability of event can be expressed as X÷N as N gets infinitely large. More simply, if you are interested to study whether use of probiotics cure the microbial infection in fish or cause more deleterious effect over the fish health, it is not possible to study the microbial infected fish population on large scale. You will investigate a small group of microbial infected fish group selected from the larger group of fish. The summary of this experiment may show that the use of probiotics cures the microbial infection in fish or not. You can say that fish that take probiotics may have more chances of getting rid of microbial infection than the infected fish that didn’t take probiotics. Here, the term “chances” in statistic language is defined as “Probability”. The idea of probability is critically important in expressing the experimental data statistically. It aids in inferring the P value and the terms significance or non-significance of the experimental data.

Every time, one study the probability, one must face the word “experiment”. The term experiment can be defined as process of assembling the observations or making the measurements one or more than one experimental units. During a complex study an experiment can be replicated serval time. Each replication of an experiment is referred as experimental trial. From each experimental trial more than one outcomes can be results. Consider, if particular event E is an experimental trial occurs many times. The probability of any specific outcome is the number of times that specific outcomes appears divided by the total number of trials. Here the probability of event E can be defined as P(E) = Number of times Event (E) occurs in an experimental trial / total number of trial in an experiment.

Probability of an event can be determined theoretically. For example, if you flip a coin, the chance of getting the tail of coin or head is ½. Similarly, if the coin is flipped 10 times, there is no surety that heads would be appeared with the ratio ½. The coin is then filliped again 10,000 times the chance of number of heads appears to the total number of trails will be near to ½. Here frequencies of head appear in each case may differ from 0 to 10,000 as in each case probability of head appears is ½.

5.1.1. PROBABILITY CONCEPTS

Example:

Following data relate to serum total protein level (gm) for 25 *Aeromonas hydrophila* challenged fish.

Serum total protein level (gm)	Number of *Aeromonas hydrophila* challenged fish	Relative Frequency
75.8 - 82.0	4	4/25
82.0 - 99.3	7	7/25
99.3 - 116.7	3	3/25
116.7 - 135.9	2	2/25
135.9 - 140.0	5	5/25
140.0 - 158.3	2	2/25
158.3 - 167.2	2	2/25
Total	**25**	**1.00**

Table 5.1. Distribution of *Aeromonas hydrophila* challenged fish by serum total protein level in ascending order

Suppose a fish is selected randomly, the probability that the *Aeromonas hydrophila* challenged fish have serum total protein level of 135.9-140.0 is 5/25, expressing the relative frequency of a group. Simply its means that out of 25 fish, 5 belongs to the group 135.9-140.0.

Example:

In a recent study relation between particular fish type and disease type in fish, a sample of diseased fish from cartilaginous fish (CF), a sample of diseased fish from bony fish (BF), a sample of diseased fish from A*ctinopterygii* fish (AF) and a sample of control that are free of the disease are classified into the disease types.

Particular fish type	Disease type			Control	Total	Probability
	Hemorrhages	Ulcer	Skin lesions			
Cartilaginous fish (CF)	536	462	985	1231	3214	**0.335**
Bony fish (BF)	982	735	428	2165	4310	**0.449**
***Actinopterygii* fish (AF)**	123	205	749	986	2063	**0.216**
Total	**1641**	**1402**	**2162**	**4382**	**9587**	**1.00**
Probability	**0.17**	**0.146**	**0.225**	**0.46**	**1.00**	

Table 5.2. Distribution of fish by particular fish type and disease type

In explaining this example one can quickly calculate the probability of a fish selected at random will belongs to particular fish type, cartilaginous fish (CF), bony fish (BF), *Actinopterygii* fish (AF) or it may suffer from disease type hemorrhages, ulcer, and skin lesions. The chances of a fish selected at random from 9587 cases (cartilaginous fish)will be

P (CF) = 3214/9587 = 0.335

Similarly, the chances that a fish selected at random from 9587 cases (ulcer group) will be

P (Ulcer group) = 1402/9587 = 0.146

If one would add the probabilities of particular fish type, cartilaginous fish (CF), bony fish (BF), *Actinopterygii* fish (AF), it will be 1.00 (Table 4.2). The value could be zero if no fish belongs to particular group and could be 1.00 if all the fish belongs to that particular group. Therefore, two important results can be drawn here:

1- The total of all the probabilities of all possible results of a study is 1.00
2- The chances of each result (microbial infection type or disease type) is more than or equal to zero but can't be more than 1 or less than zero.

Therefore, it is concluded generally that probability of any outcome lies between zero and **0≤P(A)≤1**.

5.2. RULES OF PROBABILITY

5.2.1. ADDITIVE RULE OF PROBABILITY

The addition rule states the probability of two events is the sum of the probabilities of two events that will happen minus the probability of both the events that will happen. Mathematically, the addition rule of probability is expressed as:

$$\mathbf{P\ (A \cup B) = P\ (A) + P\ (B) - P\ (A \cap B)}$$

5.2.1.1. Mutually Exclusive Events

It is important to perceive the idea of what an event and a collaborated/mutually exclusive event is before to get the idea about the rules of probability. An **event** might be describe as a one set of outcomes of study/experiment. Events can be **mutually exclusive**, if things can't happen at the same time. Like one can't run backward and forward at the same time. So running forward and running backward are mutually exclusive events. Similarly flipping of a coin can't give both head and tail at the same time. Thus, flipping a coin head and flipping a coin tail are mutually exclusive events.

To understand the **additive rule of probability**, consider the above quoted example 2. A fish can't be a cartilaginous fish and bony fish at the same time, therefore for a fish to belong to either the cartilaginous fish group or bony fish group are mutually exclusive events.

Suppose the probability of cartilaginous fish (CF) is **0.335** whereas the probability of bony fish (BF) is **0.449**. The probability of cartilaginous fish (CF) or) bony fish (BF) can be drawn as

$$\mathbf{P\ (CF\ or\ BF) = P\ (CF) + P\ (BF) = 0.335 + 0.449 = 0.784}$$

This is known as additive rule of probability for mutually exclusive events.

5.2.1.2. Non-Mutually Exclusive Events

In an experiment, two events can be **non-mutually exclusive,** if the same things happen at the same time or if there intersect is not **zero.** For example, we roll a fair dice with the event A for the multiple of 2 number and event B for the multiple of 3 number. Determine whether event A and event B are non-mutually exclusive events.

Event A and B are: **A = {2, 4, 6}; B = {3, 6, 9}.** Then we have:

$$\mathbf{A \cap B = \{6\}}$$

Since event A and event B intersect, they are non-mutually exclusive events.

It is crucial to examine the situation to find out the probability that either of two events occurs, when they are not mutually exclusive. For example, (Refer to Example 2) the type of fish disease ulcer and fish belongs to bony fish (BF) type is non-mutually exclusive events. Here the additive law of probability can be modified, otherwise the probability that both events occur would be added twice into the calculated probability. The probability that a randomly selected fish has skin lesions is = 2162/9587**= 0.225** and the fish belongs to bony fish group (BF) is = 4310/9587 = **0.449**. Here the combined probability of having skin lesions and belongs to bony fish group type has been added twice. This joint probability of having skin lesions and fish belongs to bony fish group (BF) is 428/9587= **0.044** must be subtracted from the calculated probability,

P (Skin lesions and Bony fish (BF) type) = P (Skin lesions) + P (Bony fish (BF) type)

– P (Skin lesions and Bony fish (BF) type)

= 0.225 + 0.449 – 0.044 = 0.63

Thus, the additive rule of probability for no- mutually exclusive events can be defied as

The probability that either the event A or an event B or both occurs is

P (A or B) = P (A) + (B) – P (A and B)

5.2.2. MULTIPLICATIVE RULE OF PROBABILITY

According to the multiplication rule of probability, the probability of occurrence of both the events A and B is equal to the product of the probability of B occurring and the conditional probability that event A occurring given that event B occurs.

5.2.2.1. Independent Events

The two events A and B are referring as independent events, if the fact that one event has occurred and does not affect the probability that the other event will occur. Thus if two events A and B are independent the probability that both A and B occur is equal to the product of their respective probabilities. Hence, the probability of two independent events can be summarized as:

P (A and B) = P (A) P (B)

Suppose two coins are flipped. The probability that heads occur on both coins would be

P (2 heads) = P (H1 and H2) = P (H1) P (H2)

Where the H1 denotes the head on first coin and H2 denotes the head on second coin.

Since P (H1) = P (H2) = ½ therefore P (H1 and H2) = 1/4.

This is called a multiplicative rule of probability.

5.2.2.2. Conditional Probability

Conditional probability is the probability of one event that occurs with the same relationship to one or more other events. For example, **Event A** is that it is raining outside, and it has 20% chance of raining today. **Event B** is that you need to go outside and that has the probability of 50%. The conditional probability would be at these two events A and B that occurs in a relationship with one another, i.e. the probability that it is raining and you will need to go outside. Hence,

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

And,

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

In the example 2, the probability that a fish selected at random has a skin lesions given that it is belonging to the cartilaginous fish (CF) type. The conditional probability for these two events could be drawn as

P (Skin lesions | Cartilaginous fish (CF) type) = 985/ 9587= 0.102

This may also be calculated using:

P (Skin lesions and Cartilaginous fish (CF) type) = P {Skin lesions | Cartilaginous fish (CF) type} X P {Cartilaginous fish (CF) type}

$$= \frac{985}{9587} \text{ X } \frac{9587}{3214} = \frac{985}{3214} = 0.30$$

5.2.2.3. Non-Independent Events

Refer to the above cited example 2, occurrence of hemorrhages is event A and the occurrence of fish belonging to the *Actinopterygii* fish (AF) is event B. The probability of A and B would be

P (A and B) = P (A) P (B|A) = P(B) P (A|B)

P (Hemorrhages and *Actinopterygii* fish (AF) type) = P (Hemorrhages) P (*Actinopterygii* fish (AF) type **|** Hemorrhages)

$$\left(\frac{1641}{9587}\right)\left(\frac{123}{1641}\right) = 0.0128$$

5.3. PROPERTIES OF PROBABILITY

1. Any event occurs with the probability that lies either zero or 1 and between 0 and 1.
2. If one lists all the possible events, the sum of their probabilities is always 1.
3. If the two events A and B are mutually exclusive, then the probability that either A or B occurs is equal to P (A) + P(B).
4. If two events A and B are independent, then the probability of both A and B occurring together is equal to the product of their probabilities P (A and B) = P (A) P (B).
5. If two events A and B are not mutually exclusive, the probability that either A and B or both occur is equal to P (A) + P (B) – P (AB). If A and B are mutually exclusive, then P (A and B) = 0
6. The probability of an event A, given that B has already occurred, is called the conditional probability of A given B is P (A|B).
7. The probability that both event A and B occur is P (A and B) P (A) P(A|B) =P (B) P (B|A)

5.4. PROBABILITY DISTRIBUTION

Probability distribution tells us what the probability of an event is. Probability distributions can show the simplest events e.g. flipping of a coin, or they may show much more complexed events e.g. probability of certain antibiotics treating bacterial infections in fish successfully. Before getting understanding of probability distributions, it is important to get the idea of particular terms.

5.4.1. RANDOM VARIABLES

In a statistical experiment, numerical description of the outcome is referred as random variable. There are two type of random variables

1. **Discrete random variables**

A random variable that considers only the finite and infinite number/sequence of values is known as discrete random variables. It is also referred as variable with countable set of values.

2. **Continuous random variables**

A random variable that consider any values with interval on real number line is known as continuous random variables

Probability distribution can be a table or a formula enlisting all the possible values that a random variable can take along with associated probabilities. If the random variable is discrete then this distribution is known as **discrete probability distribution**. Similarly, if the random variable is continuous than this distribution is known as **continuous probability distribution**.

5.4.2. TYPES OF PROBABILITIES DISTRIBUTIONS

However, the basic rule of probability distribution is that the sum of all the probabilities is always 100% or 1 in decimal. In statistics, there are different types of probabilities distributions

1. Binomial probability distribution
2. Poisson probability distribution
3. Normal probability distribution

5.4.2.1. Binomial Probability Distribution

Binomial distribution is named after a Swiss mathematician James Bernoulli (1654-1705). It is a type of discrete probability distribution. In biosciences experiments usually conducted with the two possible outcomes to test either the positivity or negativity of the outcomes of an event. Let’s suppose patients suffering with diabetes will referred as positive group and without diabetes will referred as negative group. The outcomes are called as success or the failure. The probability of positive results is represented by “p” while the probability of negative is represented as “q”.

Experiment with binomial probability possesses the following characteristics

1. In an experiment each parameter results in an outcome with “success” or “failure”
2. Random variable “x” counts the number of “success” or “failure” in “n” number of identical trials.
3. In an experiment, the probability of single parameter of success represented by “p” remains constant from one trial to another trial.
4. In an experiment the outcome is independent of the outcome of another experiment.

The binomial distribution proposed that the specific outcome occur in a known number of independent trials. The rule of binomial distribution can be applied to model the happening of biological events in body, e.g. occurrence of specific reaction patterns such as particular amount of interleukin-15 to activate JAK-STAT pathway which is pivotal in development and proper functioning of immune system.

To understand the binomial distribution concept; let’s suppose, 10 coins were flipped. Consider there are 8 heads and tails. While flipping the coin, the chance of head comes taken as success whereas the chance of tail come taken as failure. Thus, the probability of success is denoted by p and the probability of failure is denoted by q (q= 1-p). sSince the trials are independent, according to the multiplicative rule of probability, the probability of sequence is - S, S, S, F, F, S, S, S, S, S:

$$\mathbf{P\ (S, S, S, F, F, S, S, S, S, S) = p\ p\ p\ q\ q\ p\ p\ p\ p\ p = p^8 q^2}$$

As the chances of head or tail occurrence is equal and is 0.5, hence

$$\mathbf{P\ (S, S, S, F, F, S, S, S, S, S) = 0.5^8 \ x\ 0.5^2 = 0.00390\ x\ 0.25 = 0.000975}$$

If we make all the possible arrangements of 8 heads and 2 tails, it will appear in 20 possible ways. Thus, the probability of 8 heads when the coins are flipped will be

$$\mathbf{P\ (8\ heads\ and\ 2\ tails) = 20\ (0.5)^8 \ x\ (0.5)^2 = 0.0195}$$

5.4.2.2. Poisson Probability Distribution

Poisson probability distribution used to determine the probability of infrequent trials, e.g. probability of outcomes occurs with specified number of times with the large number of trials. This distribution was named after French mathematician S.D. Poisson. It is the basic type of discrete probability distribution. Poisson distribution is extensively used to formulate the biological and medicinal models.

Poisson distribution is useful to count the number of cells in a known volume of fluid, e.g. counting of fungal spores growing in certain volume of fluid/medium.

Poisson experiment carries following characteristics:

1. In one interval of time the occurrence of number of outcomes are free of the number of outcomes that occurs in any disjoint time intervals
2. The probability of single outcome of an event would occur in a very short time interval is proportional to the length of interval.
3. The probability of outcome of an event is independent of number of outcomes that occur before this particular time interval.
4. Thus the probabilities of other outcomes of an event that occurs in this short time interval are counted negligible.

Consider all these important points; we can draw the formula to calculate the Poisson probability. Let's suppose, X is a random variable taking on one of the values (0, 1, 2, 3, ………). This random variable is known as a Poisson random variable with the parameter μ. If μ is taken as μ >0, then its probability can be:

$$P(X) = \frac{e^{-\mu}\mu^{x}}{X!}$$

X = 0, 1, 2, 3, 4, and so on

In the above cited formula e stands for constant (e = 2.7183), μ is distribution parameter and describe the maximum number of outcomes that occurs in a short time interval as an event happened. Following are some examples of random variables for understanding that follows Poisson distribution rule.

1. The number of fish in 10 acre feet pond living for 20 years.

2. The number of deaths of healthy fish suddenly in a pond during short time interval.
3. The number of bacterial cell / colonies in known amount of gowning medium.

Example:

The probability that fish mortality occurs with bacterial infection is 0.002. Find the probability that

1. Less than 5 of the next 2000 bacterial infected fish will die.
2. Exactly 5 fish will die.

So here;

P= 0.002, n=2000, X= 5, mean = μ = np = 2000 x 0.002 = 4

(a) P (X< 5) = P (X ≤ 4) = 0.629 (Table 2.4)

(b) P (X= 5) = $\frac{e^{-4}\ 4^{5}}{5!} = \frac{0.0183 \text{ X } 124}{120} = 0.156$

Poisson probability distribution can be calculated by using the cumulative probability tables.

Probability of μ can also be calculated by using the cumulative probability table

P (X< 5) = P (X ≤ 5) – P (X ≤ 4) = 0.785 – 0.629 = 0.156

5.4.2.3. Normal Probability Distribution

It is most useful model frequently used in mathematics, medical and social sciences. It is also called as Gaussian distribution. Normal probability distribution is continuous unlike the Poisson and binomial probability distributions.

Following is the normal distribution formula:

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

x = Variable

μ = Mean

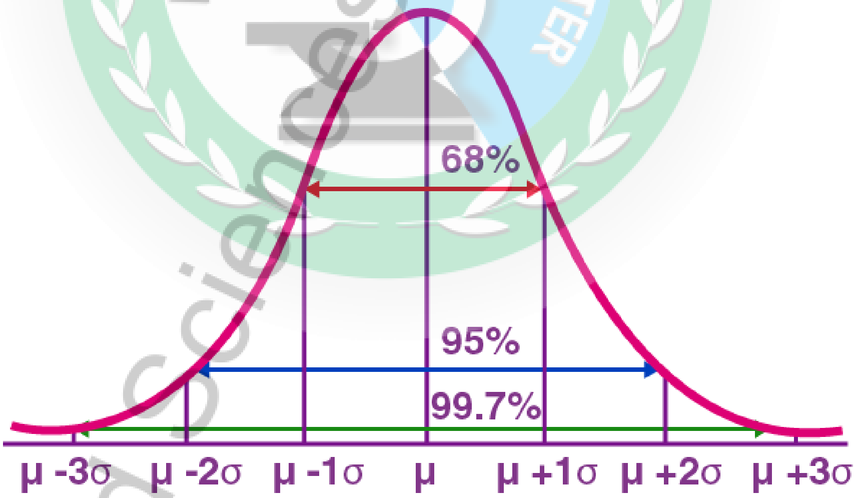
σ = Standard Deviation

Normal Distribution Curve

The graphical representation of normal distribution is referred as normal curve. The curve is bell shaped and area below normal curve represents the reference value to draw conclusion about any observation or experiment.

The variables values that are randomly distributed can estimate any unknown value prior to ranger value. The range can be from –∞ to + ∞ and a smooth curve is formed. These randomly distributed variables are called continuous variables. It has been seen that normal distribution has positive SD and this standard deviation help in calculating the spread of data. Low SD reflects that data is close to each other thus resulting in narrow graph. Whereas, larger SD reflects more dispersed data thus resulting in wide graph. In the normal distribution curve the sub-divided sections reflect the percentage of data. According to Empirical Rule using one SD:

- Approx. 68% of data falls between mean- one SD and mean + one SD
- Approx. 95% of data falls between mean- two SD and mean + two SD
- Approx. 99.7% of data falls between mean- three SD and mean + three SD



Normal Distribution Curve

Normal Distribution Properties:

Following are the normal distribution properties:

- Mean, mode and median are equal in normal distribution.

- Total area under the curve is equal to one.
- The center of the curve should be symmetric.
- Half of the values should be on right and half should be on left.
- Mean and SD defines the normal distribution curve.
- The curve should be unimodal i.e. it have one curve only.

Normal Distribution vApplication:

Normal distributions have following application:

- Blood pressure
- Heights of persons
- Objects size produced through machine
- Scores in tests

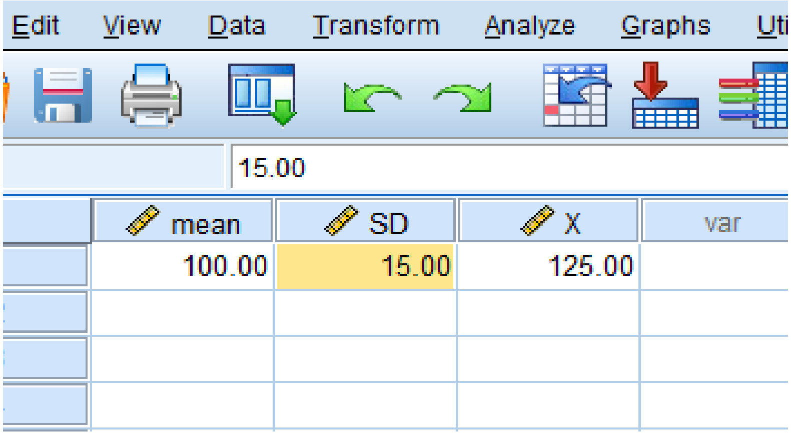
Example 1:

Assume adults have IQ score that are normally distributed with mean value of 100 and standard deviation of 15. Find probability using IBM-SPSS that a randomly selected individual has and IQ less than 125. (x<125).

Solution Key:

Open data sheet of SPSS → in 1st variable write mean and add 100 →
In 2nd variable column write SD and add 15 → in 3rd variable write x and add 125.

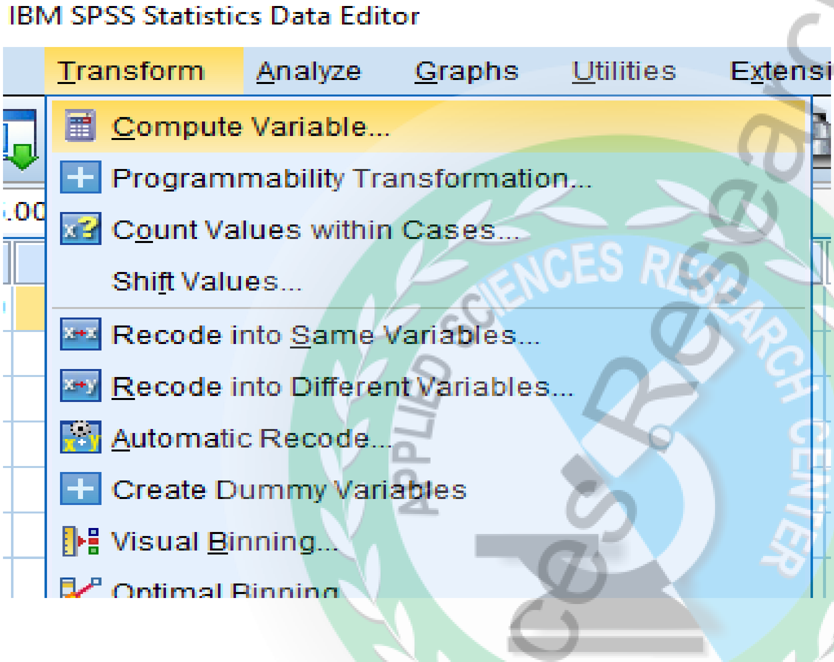
Now, go to top of page and click on transform → select compute variable →
In function group select CDF & non-central PDF → in functions and special variables select → Select CDF normal → on left side in the box CDF normal appear → click on the above arrow ⬆ → in numeric expression box add (x, mean and SD) → in target variable write p → and click OK → probability appear in data sheet.



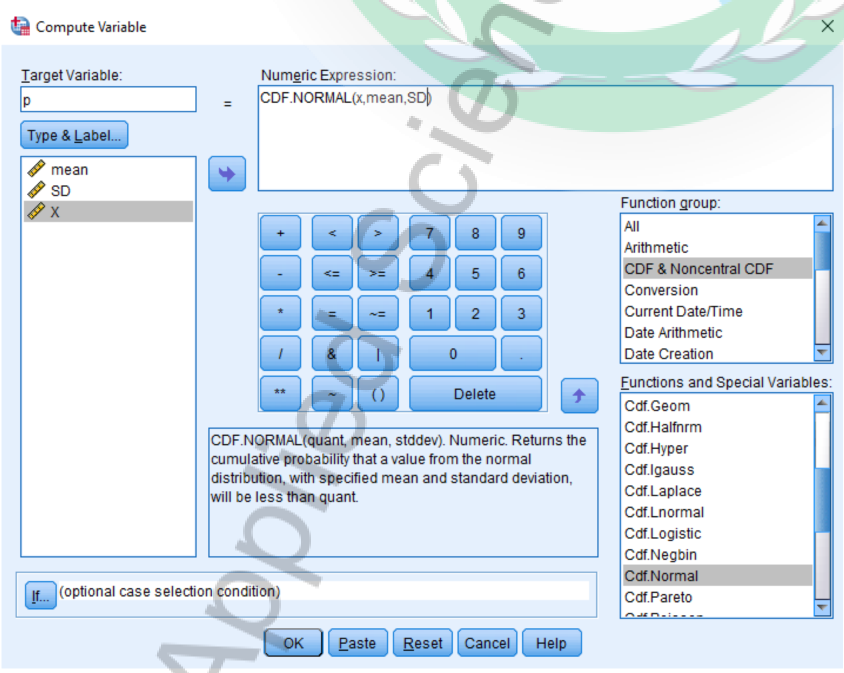
15.00

mean	SD	X	var
100.00	15.00	125.00	

Step 1:
Add Data


IBM SPSS Statistics Data Editor

Step 2:
Click on Transform and go to compute variable

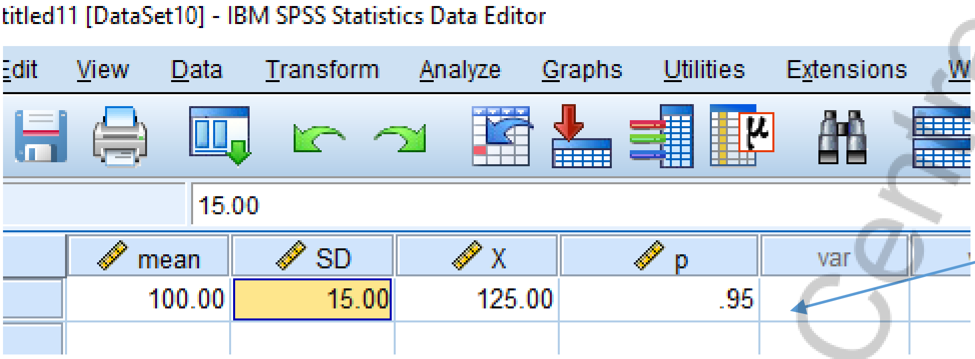

Compute Variable

Target Variable: p

Numeric Expression: CDF.NORMAL(x,mean,SD)

CDF.NORMAL(quant, mean, stddev). Numeric. Returns the cumulative probability that a value from the normal distribution, with specified mean and standard deviation, will be less than quant.

Step 3:
Select CDF & non-central CDF



titled11 [DataSet10] - IBM SPSS Statistics Data Editor

mean	SD	X	p	var
100.00	15.00	125.00	.95	

Step 4:
Value of probability appear

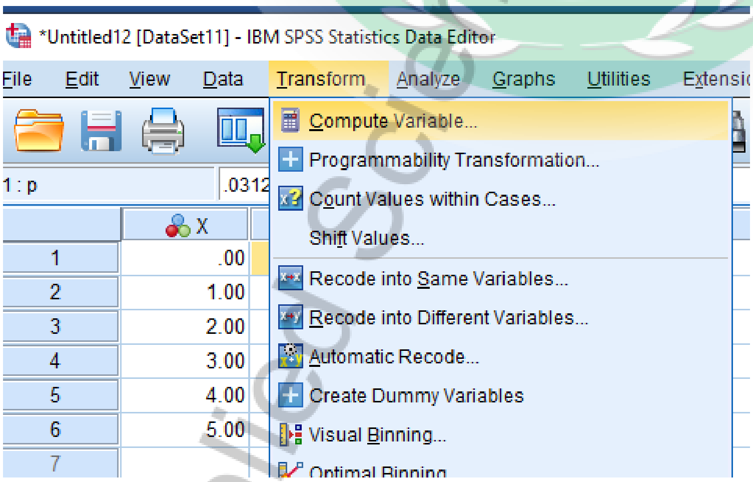
Example 2:

Estimate binomial probability distribution when x= 0, 1, 2, 3, 4, 5.

N = 5 and P = ½ = 0.2

Solution Key:

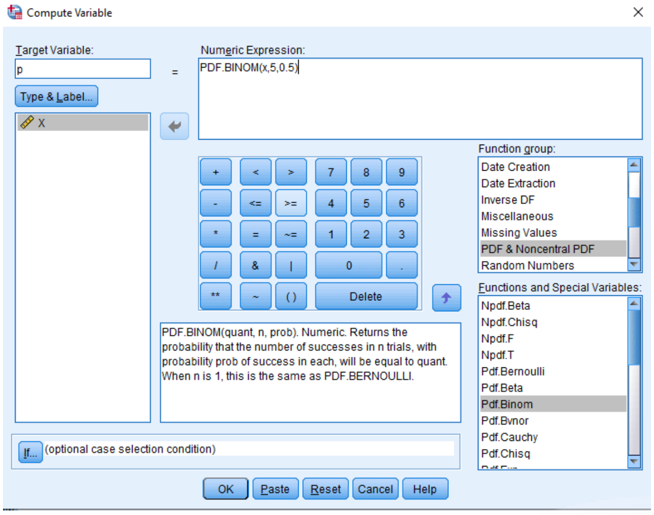
Open data sheet of SPSS → in 1st variable write x and add 0, 1, 2, 3, 4, 5. Now, go to top of page and click on transform → select compute variable → In function group select PDF & non-central PDF → in functions and special variables select PDF Binom → on left side in the box PDF appear → click on the above arrow ⬆ in numeric expression box add (x,5,0.5) → in target variable write p and click OK probability appear in data sheet.



*Untitled12 [DataSet11] - IBM SPSS Statistics Data Editor

	x
1	.00
2	1.00
3	2.00
4	3.00
5	4.00
6	5.00
7	

Step 1:
Click on Transform and go to compute variable

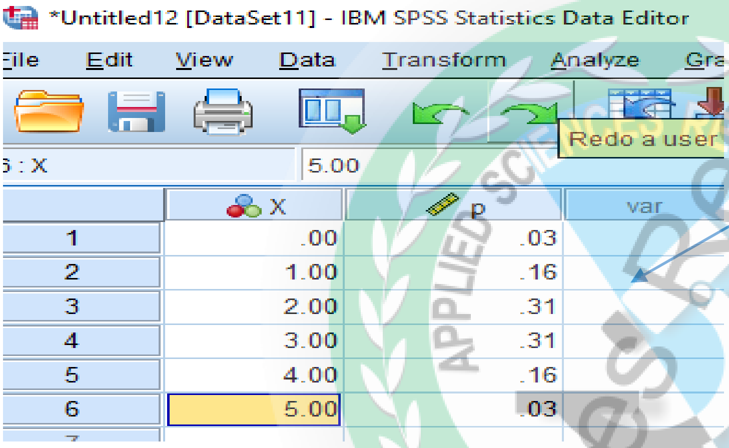


Target Variable: p

Numeric Expression: PDF.BINOM(x,5,0.5)

PDF.BINOM(quant, n, prob). Numeric. Returns the probability that the number of successes in n trials, with probability prob of success in each, will be equal to quant. When n is 1, this is the same as PDF.BERNOULLI.

Step 2:
Select PDF & non-central PDF



6 : X 5.00

	X	p
1	.00	.03
2	1.00	.16
3	2.00	.31
4	3.00	.31
5	4.00	.16
6	5.00	.03

Step 3:
Select CDF & non-central CDF

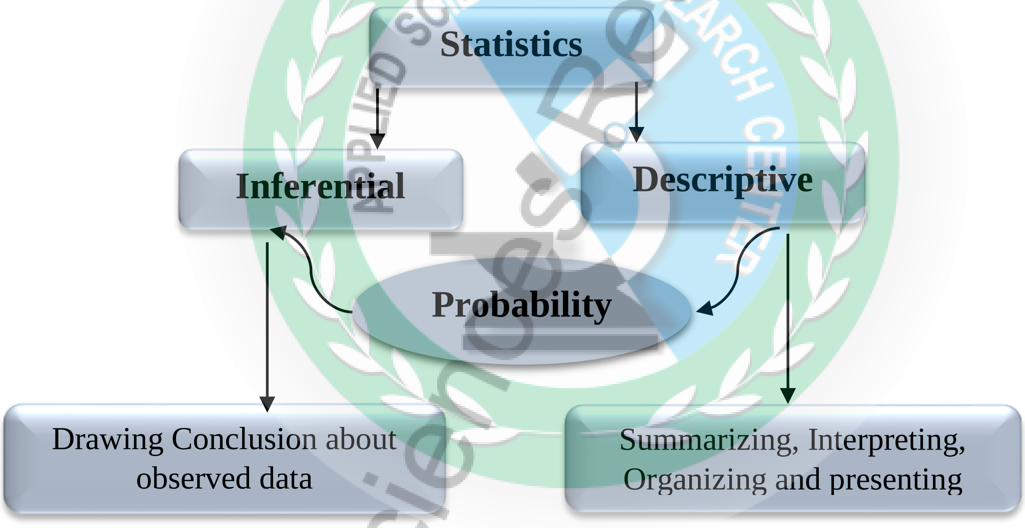
CHAPTER SIX

Procedures for Hypothesis Testing

6.1. INTRODUCTION

Generally, population parameters' inferences can be designed by two widely used and available methods. i.e.

1. Inference may be made through confidence limits.
2. Inference may be made about specific value of the population through hypotheses testing.


Statistics
Inferential
Descriptive
Probability
Drawing Conclusion about observed data
Summarizing, Interpreting, Organizing and presenting

Confidence intervals provide us an estimate of the precision for our parameter value and a range of possible values. Hypothesis tests articulate how assured we are in drawing conclusions by using our sample about the population parameter. Following example will explain it in best possible way.

Example:

Suppose a research worker, working for the Environmental Protection Agency [EPA] wants to determine whether the mean level of a certain type of pollutant released into the atmosphere by a certain chemical company meets the guidelines set by the EPA. If 4parts per million is the upper limit allowed by the EPA then the research worker will usea sample data (i.e. daily pollution measurements) to decide whether the mean is greater than 4. If, for example, 95% confidence interval for mean contains numbers greater than 4, then the research worker would suspect that the mean exceeds the established limits.

6.2. STATISTICAL HYPOTHESIS

In Statistics, a hypothesis is defined as a formal statement, which gives the explanation about the relationship between the two or more variables of the specified population.

There are two types of statistical hypothesis:

1. Null hypothesis H_0
2. Alternative hypothesis H_1

6.2.1. NULL HYPOTHESIS

A null hypothesis is a type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations. In statistics, the null hypothesis is usually denoted by letter H with subscript '0' (zero), such that H_0. It is pronounced as H-null or H-zero or H-nought.

The formula for the null hypothesis is:

$$\mathbf{H_0: p = p_0}$$

The Null Hypothesis states that there is no difference between the specified or stated value (μ_0 = mean or P_0= proportion) and actual unknown values of μ, or P of the parameters.

Example:

In every study or experiment, researchers assess an effect or relationship. This effect can be the effectiveness of a new drug, building material, or other intervention that has benefits. There is a benefit or connection that the researchers hope to identify. Unfortunately, no effect may exist. In statistics, we call this lack of an effect the null hypothesis.

6.2.2. ALTERNATIVE HYPOTHESIS

The alternative hypothesis is contradictory to the null hypothesis and states that a population parameter does not equal a specified value.

The alternative hypothesis, typically denoted with H_a or H_1 and using less than, greater than, or not equals symbols, i.e., (≠, >, or <) according to the problem. On the base of these problems, it can be specified as one tailed test and two tailed tests.

6.3. TAILS IN STATISTICS

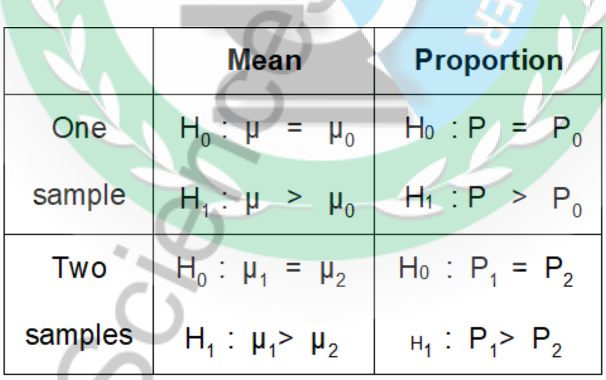
The tail refers to the end of the distribution of the test statistic for the particular analysis that you are conducting. For example, an analysis of variance (ANOVA) uses the F distribution and a t-test uses the t distribution.

6.3.1. ONE-TAILED TEST

In statistical hypothesis, one-tailed test is used to identify either sample mean would be higher or lower than the population mean, but not both. One tailed test is directional so it can be either less (<) or greater (>).

According to one tailed test, the analyst measures the possibility of relationship in one direction of interest and disregard in another direction. The requirements to run the one-tailed test are to establish a probability value (p-value), null hypothesis and an alternative hypothesis.

One tailed test for one and two sample population:

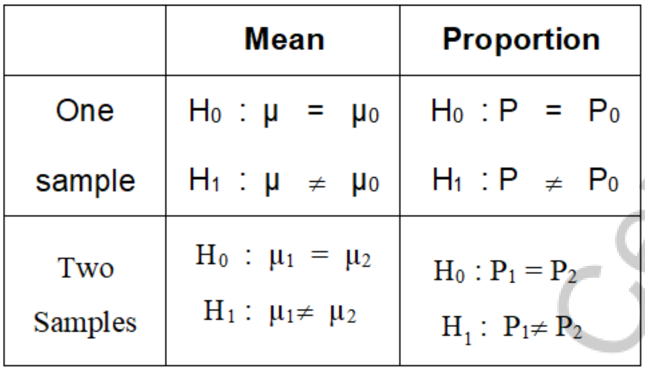


	Mean	Proportion
One sample	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	$H_0 : P = P_0$ $H_1 : P > P_0$
Two samples	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 > \mu_2$	$H_0 : P_1 = P_2$ $H_1 : P_1 > P_2$

6.3.2. TWO TAILED TEST

Two-tailed tests identify the difference in sample and population means and compare them to check either these means are statistically significant. The sample and population means are different, although both are assumed to be normally distributed and to compare these distributions, z-scores will be used.

Two tailed tests for one and two sample population.



	Mean	Proportion
One sample	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : P = P_0$ $H_1 : P \neq P_0$
Two Samples	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$	$H_0 : P_1 = P_2$ $H_1 : P_1 \neq P_2$

6.4. SIGNIFICANCE LEVEL (α)

To determine the statistical significance level, the null hypothesis is probably rejected when its true or by citing an alpha level. Significance level is selected either 1% or 5% and signified by the Greek letter α (Alpha) . 5% significance level mean that there are 5 in 100 probabilities that the null hypothesis is rejected when it is actually true and 95% null hypothesis is accepted and we are confident regarding our decision.

Tests of significance:

1. t-test or student's t-test
2. f-test or variance ratio test
3. z-test or fisher's z-test
4. Chi-Square Test (X^2-Test)

6.5. T-TEST OR STUDENT'S T-TEST

A statistical t-test is used to compare the means of two groups. It is frequently used in hypothesis testing to measure either treatment actually has an effect on the population of interest or these two groups are unlike.

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

t = Student’s t test

m = mean

μ = theoretical value

s = standard deviation

n = variable set size

A t-test can only be used to compare the means of two groups (a.k.a. pairwise comparison) but to compare the more than two groups or to do multiple pairwise comparisons, an **ANOVA test** or a post-hoc test will be used.

The t-test works as a **parametric test** of difference, as it makes the same assumptions like other parametric tests. The t-test undertakes data as:

1. Independent
2. Normally distributed.
3. Variance's homogeneity

Non-parametric alternative to the t-test such as the Wilcoxon Signed-Rank test can be used for data with unequal variances. Types of t-test are as follows:

1. One-sample, two-sample, or paired t-test
2. One-tailed or two-tailed t-test

6.5.1.1. One-sample, Two-sample, or Paired t-test

- Paired t-test is performed when the groups come from a single population (before and after results from an experimental treatment).
- Two-sample t-test or independent t-test is performed when the groups come from two different populations (two different species).
- One-sample t-test is performed to compare a one group being against a standard value (comparison of liquid's acidity to a neutral pH of 7).

Example (one-sample t-test):

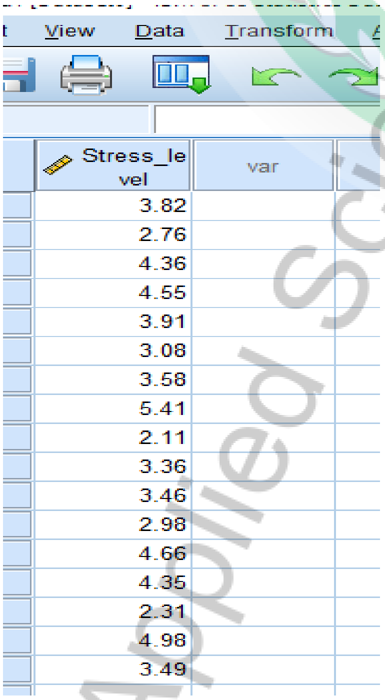
Stress levels in a selected group of 17 individuals were recorded after administration of an experimental drug. The observed stress levels were arranged in the form of tabulated dataset.

From this dataset calculate one sample t-test.

No. of Observations	Stress Level
1	3.82
2	2.76
3	4.36
4	4.55
5	3.91
6	3.08
7	3.58
8	5.41
9	2.11
10	3.36
11	3.46
12	2.98
13	4.66
14	4.35
15	2.31
16	4.98
17	3.49

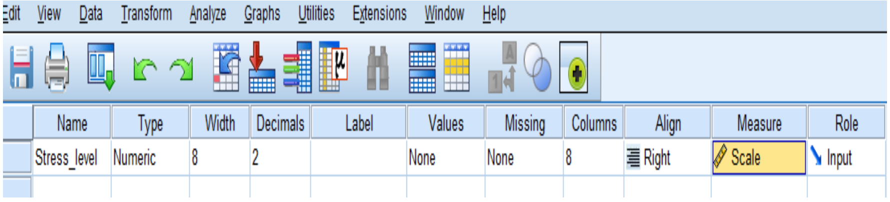
Solution Key:

Open data sheet → enter the data → click variable view → change measurement of stress level to scale → click analyze on top of menu → go to compare means → click one sample t-test → one sample t-test dialog box appear → select dependent value you want to compare → in text box write 1 because we are comparing one variable → click ok → results appear on screen



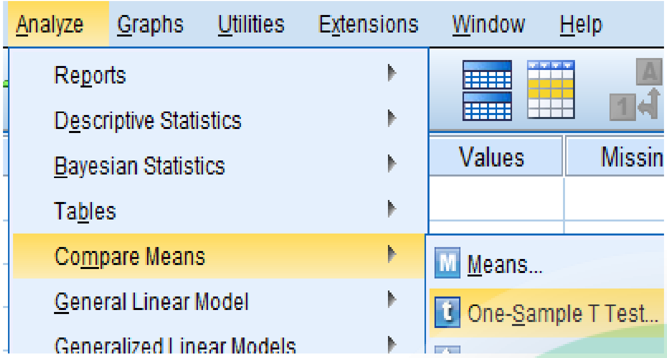
Stress_level	var
3.82	
2.76	
4.36	
4.55	
3.91	
3.08	
3.58	
5.41	
2.11	
3.36	
3.46	
2.98	
4.66	
4.35	
2.31	
4.98	
3.49	

Step 1: Enter data in data sheet

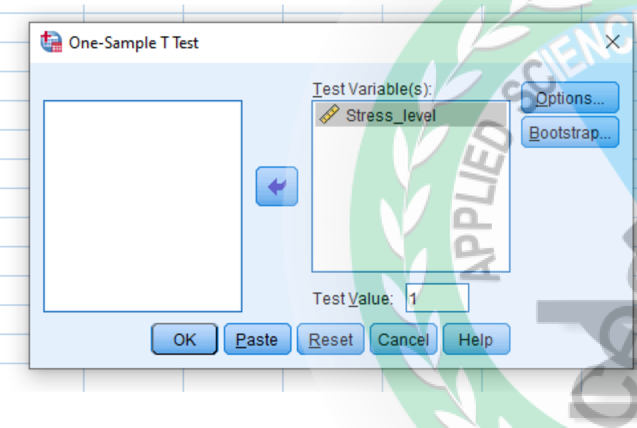

Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Stress_level	Numeric	8	2		None	None	8	Right	Scale	Input

Step 2: Adjust scale in variable sheet



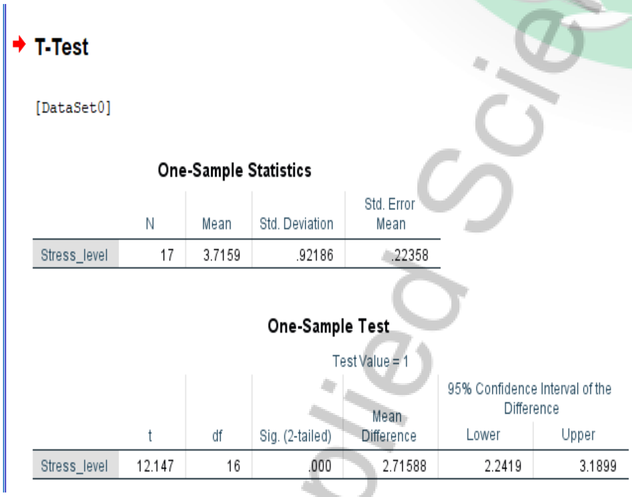
Step 3: Click on compare means and select one sample t-test



Test Variable(s): Stress_level

Test Value: 1

Step 4: Add variable



T-Test

[DataSet0]

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Stress_level	17	3.7159	.92186	.22358

One-Sample Test

	Test Value = 1					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Stress_level	12.147	16	.000	2.71588	2.2419	3.1899

Step 5: Sig. (2-tailed) values appear

Interpretation of result:

The results are significantly different as you can see the sig. (2-tailed) values are less than 0.05.

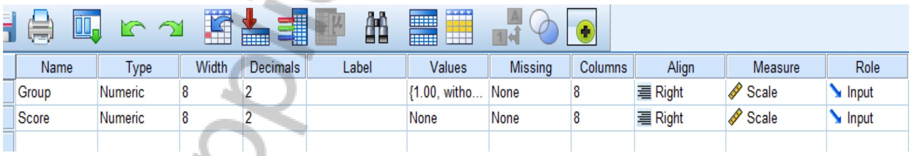
Example (two sample t-test):

From the following dataset calculate two sample t-test.

Group	Score
1	20
1	18
1	24
1	22
1	19
1	23
1	20
1	25
1	18
1	17
2	27
2	28
2	21
2	23
2	17
2	28
2	25
2	24
2	21
2	22

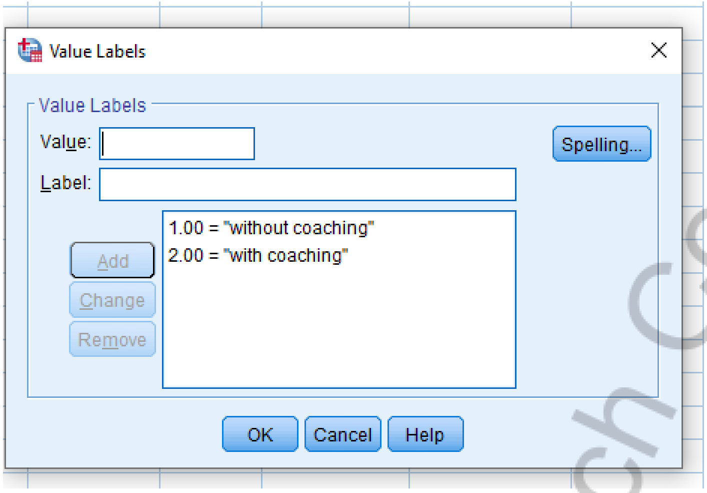
Solution Key:

Open data sheet → enter the data → click variable view → change measurement of Group and Score to scale → add values of group 1 by clicking on value and write without coaching→ write value of group 2 = with coaching click ok →click analyze on top of menu → go to compare means → click paired t-test → dialog box appear → drag the variables in box → click ok → results appear on screen



Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Group	Numeric	8	2		{1.00, witho...	None	8	Right	Scale	Input
Score	Numeric	8	2		None	None	8	Right	Scale	Input

Step 1:
Add data in variable sheet

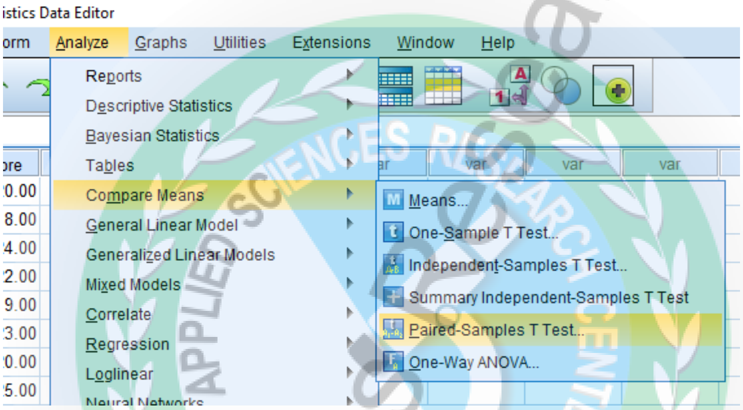


Value:

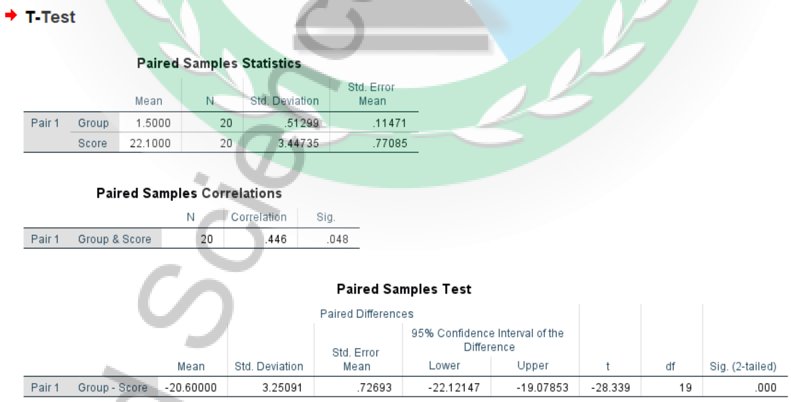
Label:

1.00 = "without coaching"
2.00 = "with coaching"

Step 2:
Label values



Step 3:
Click on compare means and select paired sample t-test



T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Group	1.5000	20	.51299	.11471
	Score	22.1000	20	3.44735	.77085

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Group & Score	20	.446	.048

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Group - Score	-20.60000	3.25091	.72693	-22.12147	-19.07853	-28.339	19	.000

Step 4:
Sig. (2-tailed) values appear

Interpretation of result:

The results are significantly different as you can see the sig. (2-tailed) values are less than 0.05.

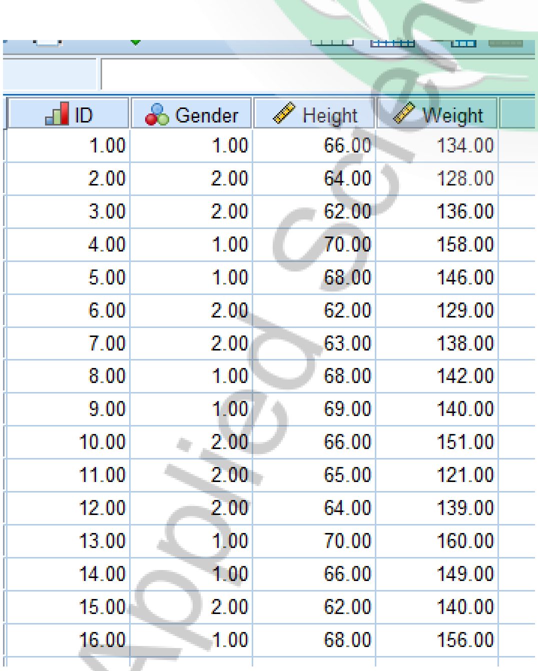
Example (Independent sample t-test):

From the following dataset calculate Independent sample t-test.

ID	Gender	Height	Weight
1	1	66	134
2	2	64	128
3	2	62	136
4	1	70	158
5	1	68	146
6	2	62	129
7	2	63	138
8	1	68	142
9	1	69	140
10	2	66	151
11	2	65	121
12	2	64	139
13	1	70	160
14	1	66	149
15	2	62	140
16	1	68	156

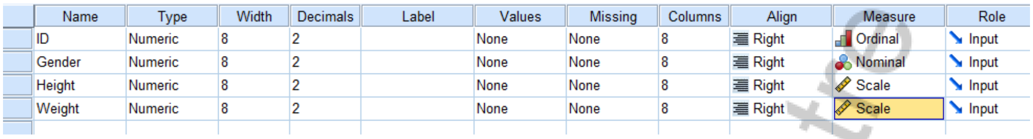
Solution Key:

Open data sheet → enter the data → click variable view → change measurement of Gender to nominal Open data sheet → enter the data → click variable view → change measurement of Group and Score to scale →change measurement of height and weight to scale → label gender by adding value 1= male, 2= female → click ok → click on analyze on top of menu → go to compare means → select independent sample t-test → dialog box appear → add height and weight variables in test variables → gender in grouping variables → define groups by writing 1 and 2→ Click ok → results appear on screen



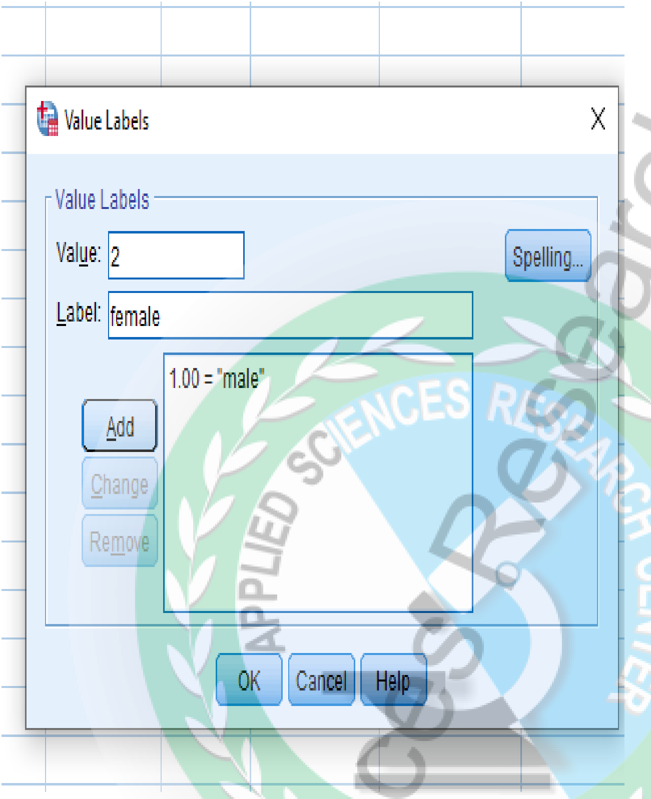
ID	Gender	Height	Weight
1.00	1.00	66.00	134.00
2.00	2.00	64.00	128.00
3.00	2.00	62.00	136.00
4.00	1.00	70.00	158.00
5.00	1.00	68.00	146.00
6.00	2.00	62.00	129.00
7.00	2.00	63.00	138.00
8.00	1.00	68.00	142.00
9.00	1.00	69.00	140.00
10.00	2.00	66.00	151.00
11.00	2.00	65.00	121.00
12.00	2.00	64.00	139.00
13.00	1.00	70.00	160.00
14.00	1.00	66.00	149.00
15.00	2.00	62.00	140.00
16.00	1.00	68.00	156.00

Step 1:
Enter data in software



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
	ID	Numeric	8	2		None	None	8	Right	Ordinal	Input
	Gender	Numeric	8	2		None	None	8	Right	Nominal	Input
	Height	Numeric	8	2		None	None	8	Right	Scale	Input
	Weight	Numeric	8	2		None	None	8	Right	Scale	Input

Step 2:
Label data in variable sheet

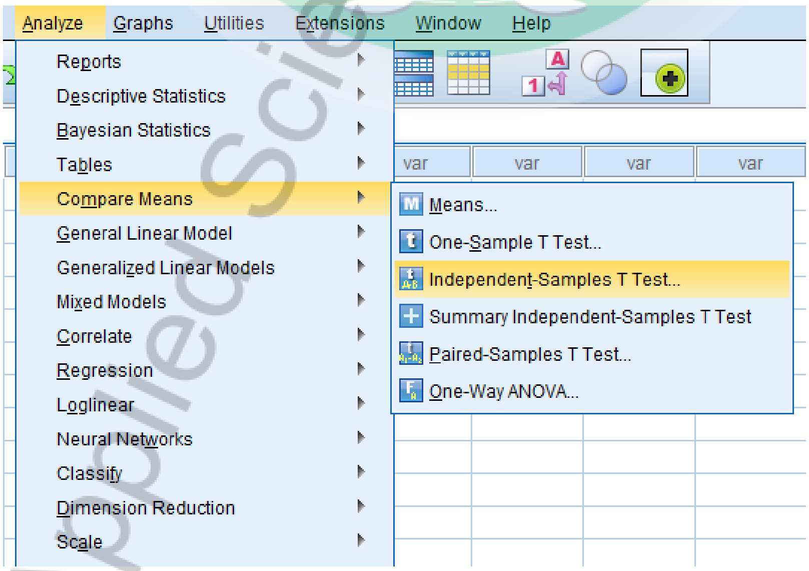


Value: 2

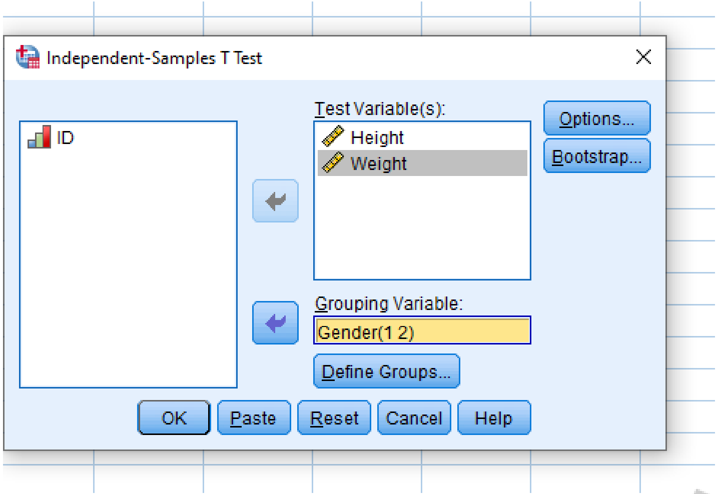
Label: female

1.00 = "male"

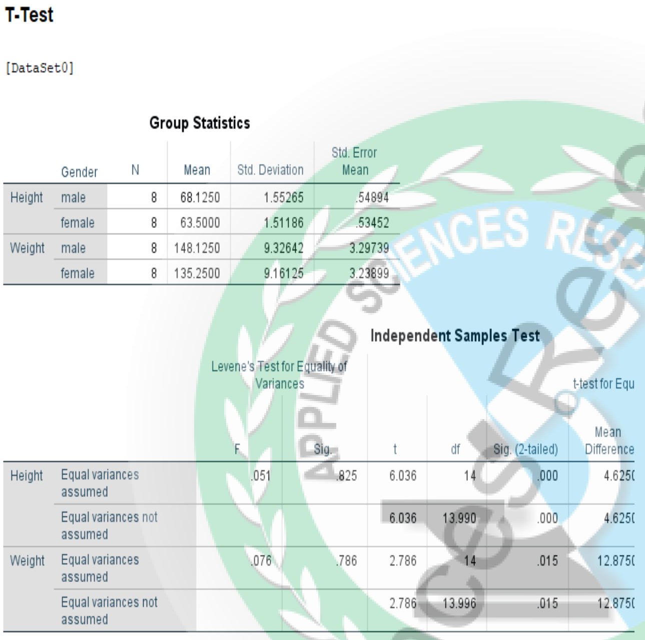
Step 3:
Value labels



Step 4:
Click on compare means and select independent sample t-test



Step 5:
Click on compare means and select one sample t-test



T-Test

[DataSet0]

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Height	male	8	68.1250	1.55265	.54894
	female	8	63.5000	1.51186	.53452
Weight	male	8	148.1250	9.32642	3.29739
	female	8	135.2500	9.16125	3.23899

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equu			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
Height	Equal variances assumed	.051	.825	6.036	14	.000	4.625(
	Equal variances not assumed			6.036	13.990	.000	4.625(
Weight	Equal variances assumed	.076	.786	2.786	14	.015	12.875(
	Equal variances not assumed			2.786	13.996	.015	12.875(

Step 6:
Sig. (2-tailed) values appear

Interpretation of result:

The results are significantly different as you can see the sig. (2-tailed) values are less than 0.05.

6.5.1.2. One-tailed or Two-tailed t-test

- Two-tailed t-test is performed when the two populations are different from one another.
- One-tailed t-test is performed to check either one population mean is greater than or less than the other.

6.6. VARIANCE RATIO TEST (F TEST)

The F test is used to compare two variances, σ1 and σ2. This statistical test was named after a British statistician called Ronald Aylmer Fisher (17 February 1890 - 29 July 1962), in which the test statistic has an F-distribution under the null hypothesis.

In order to compare and identify which model best fits the population from which the data were sampled, the F test is used. The value of F is equal to the ratio of variance.

Formula:

$$F = \frac{\textbf{Estimate variance between groups (including treatment effects)}}{\textbf{Estimate variance within groups (excluding treatment effects)}}$$

The F Value is calculated using the formula:

$$F = \frac{(SSE_1 - SSE_2 / m)}{SSE_2 / n\text{-}k}$$

SSE = residual sum of squares

m = number of restrictions

k = number of independent variables

In ANOVA, the F test is performed to identify the presence or absence of a significant difference among multiple samples

6.6.1. ONE-WAY CLASSIFICATION OF VARIANCE

One-way ANOVA is used to test the difference between two sets of groups without any further categorization with the concern of other factors.

Steps of finding F value in ANOVA:

Step 1: Transcribe null hypothesis.

$$\frac{\text{there is no significant difference}}{\text{variation among the treatments (different concentration of hormones)}}$$

Step 2: Sum of square (Total SS) calculation by the means of this formula:

square (Total SS) calculation by the means of this formula:

$$\frac{\sum x^2 - (\sum x)^{2}}{n}$$

$(\sum x)^{2}/n$ = correction factor

n = number of observed samples

$(\sum x)^{2}/n$ = correction factor

n = number of observed samples

By putting values:

$$\begin{aligned} SS &= (3^2+4^2+8^2+9^2) - (7+17)^2/4 \\ &= 170 - 144 \\ &= 26. \end{aligned}$$

Step 3: Compute treatment Sum of square by the means of this formula:

$$\frac{\text{(total of 1st treatment)}^2/n + \ldots + \text{(total of last treatment)}^2/n - (\sum x)^{2}/}{n}$$

n = number of observations under one treatment

By putting values

$$\begin{aligned} \text{Treatment SS} &= (7^2/2 + 17^2/2) - (7+17)^{2}/4 \\ &= 169 - 144 \\ &= 25. \end{aligned}$$

Step 4: Analyze Residual Sum of square by the means of this formula:

Residual SS= Total SS− Treatment SS

Residual SS= 26 − 25 = 1

Step5: Estimate degrees of freedom (df) for total, treatment and residual.

df for total and treatment = n−1

df for residual = total degrees of freedom− treatment degrees of freedom.

By putting values:

treatment df = 2−1 = 1

residual df = (4−1) − (2−1) = 2

Step 6: Mean square for treatment and residual using this formula:

$$\text{Mean square} = \frac{\text{Sum of square}}{\text{degrees of freedom}}$$

By putting values:

Treatment mean square = 25/1 = 25

Residual mean square = 1/2 = 0.5

Step7: Estimate variance ratio or F value using this formula

$$\text{F Value} = \frac{\text{Mean square of treatment}}{\text{Mean square of Residual}}$$

Step 8: Draw ANOVA table including all the headlines below-

Source of Variation	Sum of Square	Degrees of freedom	Mean Square	Variance ratio	Tabulated F value
Treatment					
Residual					

Step 9: At the end, compare the calculated F value with F distribution table and note either calculated value is greater than the tabulated value, then the null hypothesis is rejected and vice versa.

6.6.2. TWO-WAY CLASSIFICATION OF VARIANCE

Two-way ANOVA is used to test the difference between two sets of groups with further categorization with the concern of two or more factors.

There are quite different steps to find the F value in Two-way ANOVA.

Step1 and 2: First two steps are same as one-way ANOVA.

Step3: Find out the Treatment Sum of square for both factors and treatments applied by using this formula:

$$\textbf{SS for Treatments} = \frac{(\text{total of 1st treatment})^2/n + \ldots + (\text{total of last treatment})^2}{n - (\sum x)^2/n}$$

$$= \frac{(C1)^2/n + (C2)^2/n + (C3)^2}{n - (\sum x)^2/n}$$

$$\textbf{SS for factors} = \frac{(F1)^2/n + (F2)^2/n + (F3)^2/n + (F4)^2}{n - (\sum x)^2/n}$$

Step4: Find out Residual Sum of square by using this formula:

Residual SS = Total SS − (Treatment SS + Factor SS)

Step5: As in one way ANOVA, find out Compute degrees of freedom for Total, Treatment, Factor and Residual.

Step6: Find out the Mean square for two Treatment, sum of squares for different treatments (concentration) and factor, residual Sum of squares by using this formula:

Mean square = Sum of square/degrees of freedom.

Step 7: Variance ratio or F value for both Mean square values will be calculated by:

F value for treatments (concentration) = mean square of treatment (concentration)/ mean square of residual.

F value for factors = mean square of factors/ mean square of residual

Step8: Draw ANOVA table including all the headlines below-

Source of Variation	**Sum of Square**	**Degrees of freedom**	**Mean Square**	**Variance ratio**	**Tabulated F value**
Treatment (Concentration)					
Treatment (Factors)					
Residual					

Step9: At the end, compare the calculated F value with F distribution table and note either calculated value is greater than the tabulated value, then the null hypothesis is rejected and vice versa.

6.7. Z-TEST OR FISHER'S Z-TEST

The Fisher Z-test or transformation is used to change Pearson's correlation coefficient (r) into z_r value to calculate a confidence interval for Pearson's correlation coefficient.

The **Z-test** formula is as follows:

$$\mathbf{z_r} = \quad \ln((1+r) / (1-r)) / 2$$

For example,

if r = 0.55, then z_r would be:

$$z_r = \ln((1+r) / (1-r)) / 2$$

$$z_r = \ln((1+.55) / (1-.55)) / 2$$

$$z_r = 0.618$$

The important step is that the sampling distribution of this transformed variable follows a normal distribution and it permits us to analyze a confidence interval for a Pearson correlation coefficient. We will not be able to calculate a unfailing confidence interval for the

Pearson correlation coefficient without performing this Fisher Z test. This can be best explained by the following example.

Example: Calculate confidence interval for correlation coefficient.

To analyze the correlation coefficient between weight and height of residents, select a random sample of 60 residents and find the following information:

Sample size **n** = 60

Correlation coefficient between height and weight **r** = 0.56

Following steps are to find a 95% confidence interval for the population correlation coefficient:

Step 1: Use Z test to Perform Fisher transformation.

$$z_r = \ln((1+r) / (1-r)) / 2$$

$$= \ln((1+.56) / (1-.56)) / 2$$

$$= 0.6328$$

Step 2: Find log lower and upper bounds by using the following formula:

Lower $= z_r - (z_{1-\alpha/2} / \sqrt{n-3})$

$= .6328 - (1.96 / \sqrt{60-3})$

$= \mathbf{.373}$

Upper $= z_r + (z_{1-\alpha/2} / \sqrt{n-3})$

$= .6328 + (1.96 / \sqrt{60-3})$

$= \mathbf{.892}$

Step 3: calculate confidence interval by using the following formula:

Confidence interval $= [(e^{2L}-1)/(e^{2L}+1),\ (e^{2U}-1)/(e^{2U}+1)]$

Confidence interval $= [(e^{2(.373)}-1)/(e^{2(.373)}+1),\ (e^{2(.892)}-1)/(e^{2(.892)}+1)]$

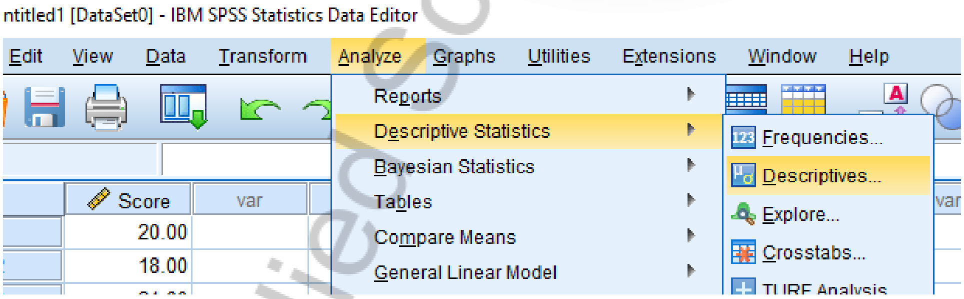
$= \mathbf{[.3568, .7126]}$

Example: From the following dataset perform Z test.

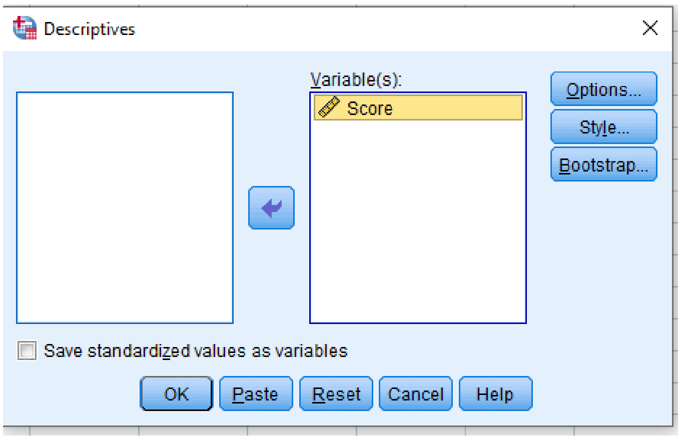
Group	Score
1	20
1	18
1	24
1	22
1	19
1	23
1	20
1	25
1	18
1	17
2	27
2	28
2	21
2	23
2	17
2	28
2	25
2	24
2	21
2	22

Solution Key:

Open data sheet → enter the data → click variable view → change measurement of Group and Score to scale → add values of group 1 by clicking on value and write without coaching→ write value of group 2 = with coaching click ok →click analyze on top of menu → go to descriptive statistics → dialog box appear → drag score in variables box →check save the standardized values as variables → click ok → results appear on screen.



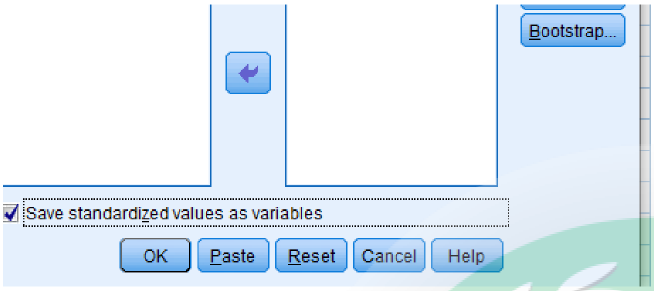
Step 1:
Go to descriptive statistics



Variable(s): Score

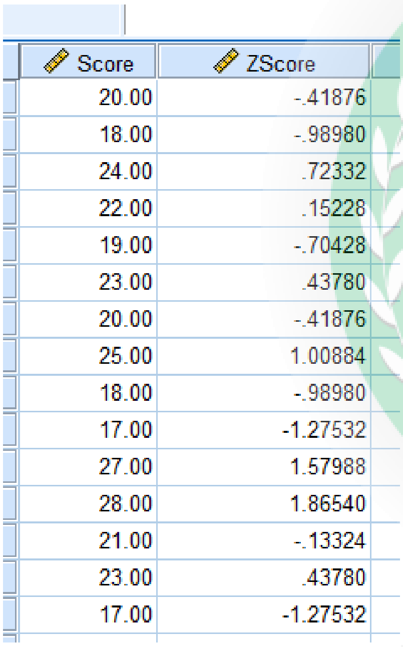
□ Save standardized values as variables

Step 2:
Add variables



☑ Save standardized values as variables

Step 3:
Check save standardized values



Score	ZScore
20.00	-.41876
18.00	-.98980
24.00	.72332
22.00	.15228
19.00	-.70428
23.00	.43780
20.00	-.41876
25.00	1.00884
18.00	-.98980
17.00	-1.27532
27.00	1.57988
28.00	1.86540
21.00	-.13324
23.00	.43780
17.00	-1.27532

Step 4:
Z-scores appear in column

Interpretation of result:

The value of Z score reflects how much standard deviation you are away from mean. 1 score indicates 0 standard deviation. Whereas, -1 score indicate that it is one SD below the mean. Positive 1 score indicate it is 1 standard deviation above the mean.

6.8. CHI-SQUARE TEST (X^2-TEST)

A statistical test, chi-square is used to do a comparison among observed results with expected results.

6.8.1. TYPES OF CHI-SQUARE

- Chi-square goodness of fit test (determines either **sample** data matches a **population**).
- Chi-square test for independence (compares two variables in a contingency table to check their relatability).

Both of these use the chi-square statistic and distribution for different purposes.

6.8.2. CHI-SQUARE FORMULA

The formula for the chi-square statistic is:

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

c = degrees of freedom
O = observed value
E = expected value

The summation symbol presents the requirement of calculation for every single data item in your data set. Calculations can be done in:

- Chi Square Test in SPSS
- Chi Square P-Value in Excel

There are two types of variables in statistics:

1. Numerical (countable) variables
2. Non-numerical (categorical) variables

The relationship between these two variables can be measured by chi-square. The chi-squared statistic estimates how much difference exists between your observed and expected counts when there was no relationship at all in the population.

6.8.3. CHI SQUARE P-VALUES

P value will be found out by using chi square test that show the significance of results. Before performing chi square test, P value should be calculated.

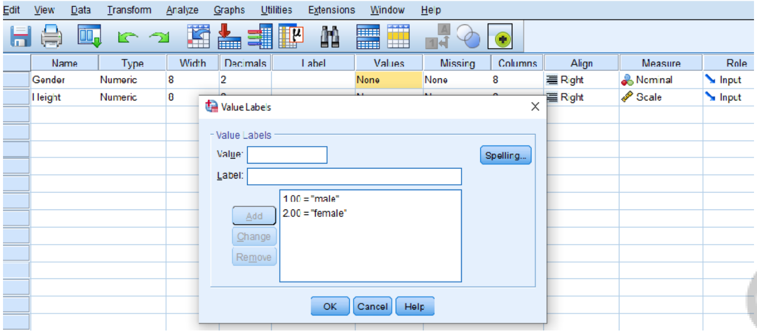
Example:

From the following dataset calculate the Chi-square test.

Gender	Height
1	66
2	64
2	62
1	70
1	68
2	62
2	63
1	68
1	69
2	66
2	65
2	64
1	70
1	66
2	62
1	68

Solution key:

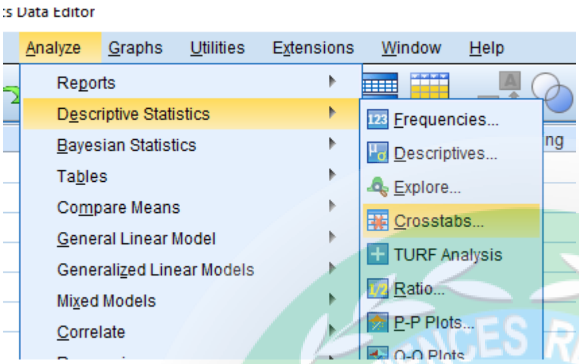
Open data sheet → enter the data → click variable view → change measurement of Gender to nominal Open data sheet → enter the data → click variable view → change measurement of Group and Score to scale →change measurement of height to scale → label gender by adding value 1= male, 2= female → click ok → click on analyze on top of menu → go to Descriptive statistics → select crosstabs → dialog box appear → place gender in rows and place height in columns → click statistics → Chi-square, Phi and Cramer's box → click continue and Ok → Results appear on screen.


Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Value Labels
Spelling...
Add
Change
Remove
OK
Cancel
Help

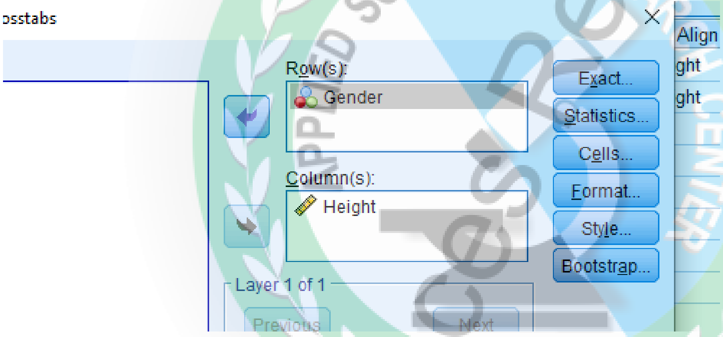
Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Gender	Numeric	8	2		None	None	8	Right	Nominal	Input
Height	Numeric	8	0		None			Right	Scale	Input

Value Labels
Value:
Label:
1.00 = "male"
2.00 = "female"

Step 1:
Add value labels



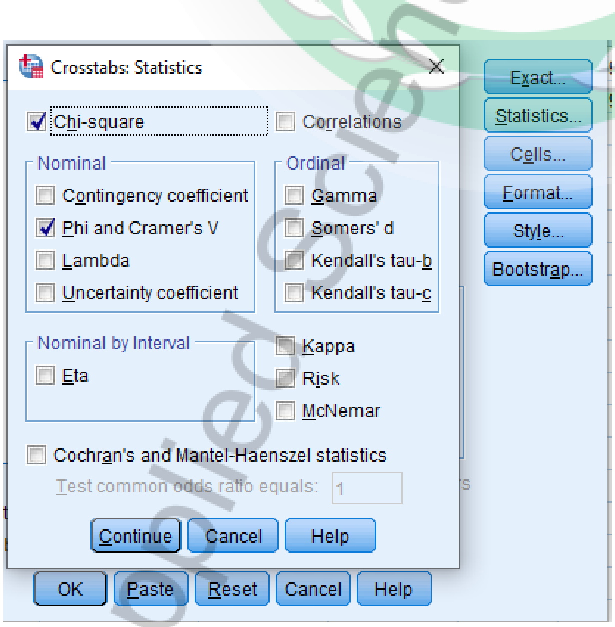
Step 2:
Go to descriptive statistics and select crosstabs



Row(s): Gender

Column(s): Height

Step 3:
Add variables



☑ Chi-square □ Correlations

Nominal
□ Contingency coefficient
☑ Phi and Cramer's V
□ Lambda
□ Uncertainty coefficient

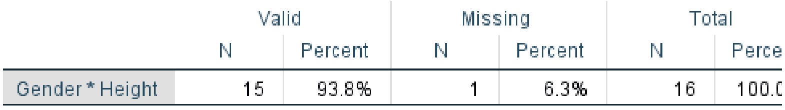
Ordinal
□ Gamma
□ Somers' d
□ Kendall's tau-b
□ Kendall's tau-c

Nominal by Interval
□ Eta

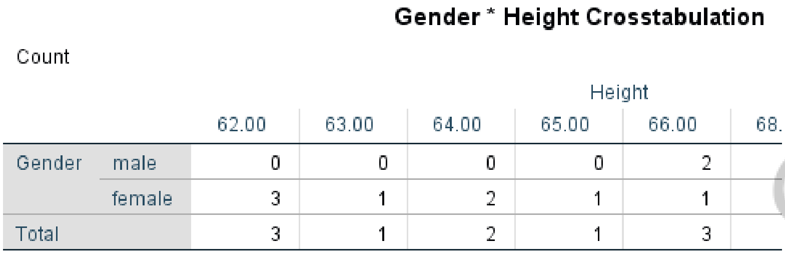
□ Kappa
□ Risk
□ McNemar

□ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1

Step 4:
Check Chi-square



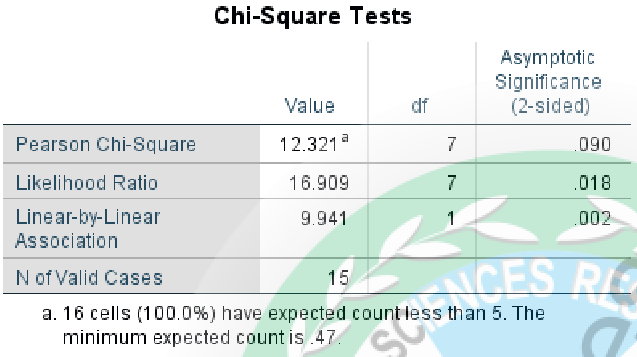
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Perce
Gender * Height	15	93.8%	1	6.3%	16	100.(



Gender * Height Crosstabulation

Count

		Height					
		62.00	63.00	64.00	65.00	66.00	68.
Gender	male	0	0	0	0	2	
	female	3	1	2	1	1	
Total		3	1	2	1	3	



Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	12.321[a]	7	.090
Likelihood Ratio	16.909	7	.018
Linear-by-Linear Association	9.941	1	.002
N of Valid Cases	15		

a. 16 cells (100.0%) have expected count less than 5. The minimum expected count is .47.

Step 5:
Pearson's chi-square values appear

Interpretation of result:

The results of Chi square appear in Pearson Chi square column and value is 12.321. Higher Chi square score indicate strong association between variables and more likely the rejection of null hypothesis.

CHAPTER SEVEN

Analysis of Variance (ANOVA)

7.1. INTRODUCTION

When we test the hypothesis with two independent samples usually t-test is applied but to study more than two independent samples or population t-test is not applicable. In experimental work there are more than two populations which cannot be studied through t-test. Sir R.A. Fisher and his mates constructed data analysis technique for more than two population which is known as Analysis of variance or ANOVA. In decision making situation when there are many samples, then these samples are divided into sub-groups.

> "The systematic procedure in which the total variation of the data is classified into non-overlapping groups is called Analysis of variance or ANOVA".

Initially this statistical method was designed to study the experimental data but later it is used to analyze survey and data from descriptive study. In ANOVA the dependent variable is measured on ratio scale or interval and it is known as **metric.** Whereas the independent variables are categorical and have more than one category.

For example a shoe company sell its product all over the country. The company has sub-divided into four sub-groups (East, West, North, and South) for administrative purposes. The sale data is collected randomly from all four sub-groups. The variation in sales can be seen among all four sub-groups and even within the sub-group. The sources of variations can be analyzed in scales and there may be two types of variation sources:

- Variation within Sub-group: Sales within the sub-groups may vary.
- Variation among Sub-groups: Sale of all four sub-groups wouldn't be same and there may be variation among sub-groups.

So the total variation in sales can be partitioned into two ways: variation between sub-groups and within sub-groups. The comparison can be made among regions to check whether the sale difference is substantial or not. If the variations are not in close agreement then it can be concluded that in four sub-groups the sales are not same and a substantial difference of sales occur among all four sub-groups.

7.2. TYPES OF ANALYSIS OF VARIANCE (ANOVA)

If the independent variable consists of one factor only which affects the values of response variable or dependent variable then the data is analyzed through one-way analysis of variance or one-way ANOVA. The above shoe company example comes under one-way ANOVA. The different aptitude of medical graduate students among various subjects can be analyzed through one-way ANOVA. On the other hand, if there are two independent variables which effect the measurements of dependent variable then two-way ANOVA is applied to analyze the data. For example if the sale of shoes are analyzed among four sub-regions then another factor like type of area can be added in the sub region like urban or rural outlet. This type of sales study will be analyzed through two-way ANOVA. Likewise if we are studying bacterial soil fauna from four different areas of Lahore then one-way ANOVA will be applied and on the other hand we study bacterial fauna from soil and sewage samples of different areas of Lahore then analysis will be studied through two-way ANOVA. Another term multivariate analysis of variance MANOVA is applied to analyze the study when multiple independent variables are present. For example if we want to study the bacterial fauna from soil, sewage and air of different regions of Lahore. This type of study is analyzed through MANOVA.

There are two categories of two-way ANOVA.

1. Two-way ANOVA with one observation per combination
2. Two-way ANOVA with multiple observations per combination

7.2.1. TWO-WAY ANOVA WITH ONE OBSERVATION

In this type of ANOVA there will be one category per factor. So the two factors will have one category each. Suppose there are two independent variables A and B and each of the variables will have category m and n simultaneously A (m) and B (n). So, N= m*n

7.2.2. TWO-WAY ANOVA WITH MULTIPLE OBSERVATIONS

In this type of ANOVA, each of the independent variable consist of multiple observations or categories per factor.

7.3. BACKGROUND OF ANOVA

Linear Model is the concept behind the analysis of variance.

X_1, X_2, X_3, X_4,..................... X_n are the observable quantities and expressed as:

$X_i = \mu_i + e_i$

μ_i is the true value due to assignable causes.

e_i is the all error term due to random causes.

These are independent and are distributed normal variable with common variance and mean zero.

μ_i can be considered to have linear function of t_1, t_2, t_3, t_4,...,t_k and this is known as effect. There are two types of effects:

1. **Fixed-effect model:** When all effects are constant in a linear model.
2. **Random-effect model:** When all effects are random in a linear model.

7.4. ASSUMPTIONS OF ANOVA

1. The samples which have been collected from a population must have normal distribution.
2. The sample collection should be random and independent.
3. There should be a common variance per each group which means that the different groups have equal variability in the dependent variable.
4. If the sample is large then the minor deviations will not affect the linear model of ANOVA.

7.5. ONE-WAY ANALYSIS OF VARIANCE

It is a parametric test to check that the statistical evidence of samples are significant or not. This is also known as one-way ANOVA or one-Factor ANOVA. There are two variables in this test:

- Dependent variable
- Independent variable or factor

Variable divides into groups or level. One-way ANOVA is used to test:

- Experimental studies
- Field studies

And it analyze mean of two or more groups and calculate statistical differences.

7.5.1. DATA REQUIREMENTS FOR ONE-WAY ANOVA

Following data is required for one-way ANOVA:

1. Continuous dependent variable (Interval/ratio level).
2. Categorical independent variable (2 or more groups).
3. Experiments have values of both independent and dependent variables.
4. Random sampling.
5. Dependent variable should have normal distribution.
6. Homogenized variances.

7.5.2. NULL HYPOTHESES OF ONE-WAY ANOVA

Null or alternative hypotheses of one-way ANOVA can be expressed as:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

$$H_1: \text{At least one } \mu_i \text{ different}$$

μ_i is population mean of i group

This test indicates whether a study is significant or not. If p value is less than 0.05 then the null hypotheses is rejected and the test or study is significant. Value of p more than 0.05 indicate non-significance of the results. Test statistics of one-way analysis of variance is named as F for independent variable and k is denoted for its groups. The F values reflect whether the results or the tests are significant or not. Following table shows the F statistic components:

	Sum of Squares	df	Mean Square	F
Treatment	SSR	df_r	MSR	MSR/MSE
Error	SSE	df_e	MSE	
Total	SST	df_T		

where

SSR = the regression sum of squares

SSE = the error sum of squares

SST = the total sum of squares (SST = SSR + SSE)

dfr = the model degrees of freedom (equal to dfr = k - 1)

dfe = the error degrees of freedom (equal to dfe = n - k)

k = the total number of groups (levels of the independent variable)

n = the total number of valid observations

dfT = the total degrees of freedom (equal to dfT = dfr + dfe = n - 1)

MSR = SSR/dfr = the regression mean square

MSE = SSE/dfe = the mean square error

Then the F statistic itself is computed as

7.5.3. DATA SET-UP FOR ONE-WAY ANOVA

The data must include minimum two variables used for analysis. The independent variable is categorical (Either ordinal or nominal) and consist of minimum two groups. Whereas, dependent variable in continuous (either interval or ratio). Each row of dataset reflect a unique title.

One-way analysis of variance is helpful in determining the significance differences on the dependent variable exist among two or more groups. Through Post-Hoc test indicate where the differences occur.

Example:

For the following dataset, test the level of significance using one-way ANOVA Post Hoc test.

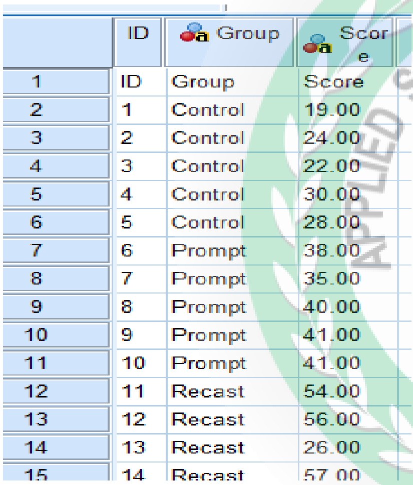
ID	Group	Score
1	Control	19.00
2	Control	24.00
3	Control	22.00
4	Control	30.00
5	Control	28.00
6	Prompt	38.00

7	Prompt	35.00
8	Prompt	40.00
9	Prompt	41.00
10	Prompt	41.00
11	Recast	54.00
12	Recast	56.00
13	Recast	26.00
14	Recast	57.00
15	Recast	58.00

Solution Key:

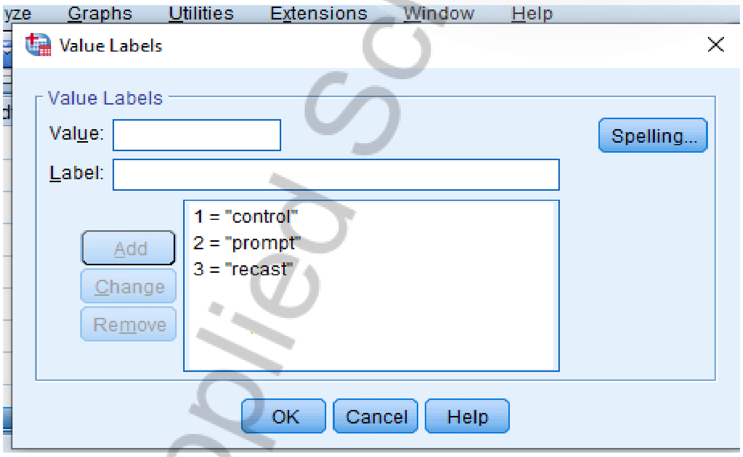
Enter data in datasheet → Click on variable view → add name in 1st row → click on label → write 1 in value and write Control in label box→ click add and then write 2 in value box and Prompt in label box → click add and write 3 in value box and recast in label box→ click add and ok→ In measure column select nominal.

In 2nd row write score and label it with score → in measure column select scale.



	ID	Group	Score
1	ID	Group	Score
2	1	Control	19.00
3	2	Control	24.00
4	3	Control	22.00
5	4	Control	30.00
6	5	Control	28.00
7	6	Prompt	38.00
8	7	Prompt	35.00
9	8	Prompt	40.00
10	9	Prompt	41.00
11	10	Prompt	41.00
12	11	Recast	54.00
13	12	Recast	56.00
14	13	Recast	26.00
15	14	Recast	57.00

Step 1:
Enter data



Value Labels

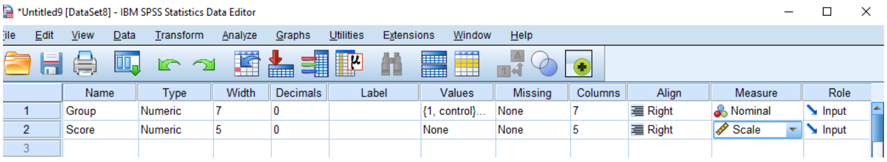
Value Labels

Value:

Label:

1 = "control"
2 = "prompt"
3 = "recast"

Step 2:
Value labels

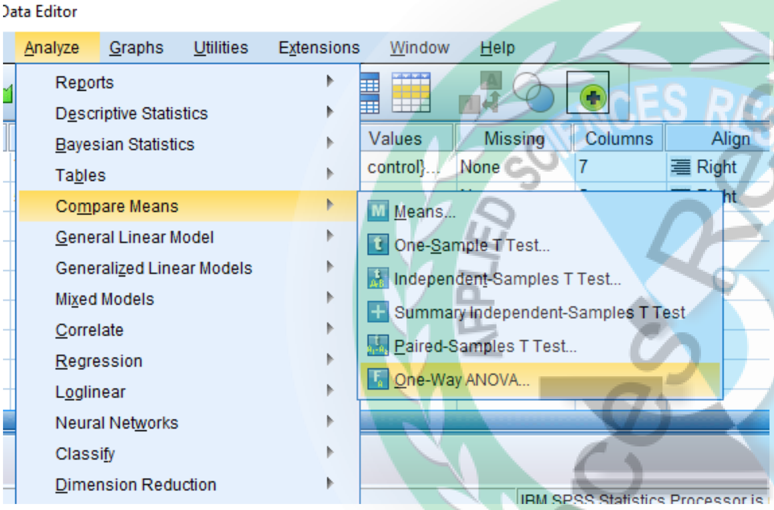

*Untitled9 [DataSet8] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Group	Numeric	7	0		{1, control}...	None	7	Right	Nominal	Input
2	Score	Numeric	5	0		None	None	5	Right	Scale	Input
3											

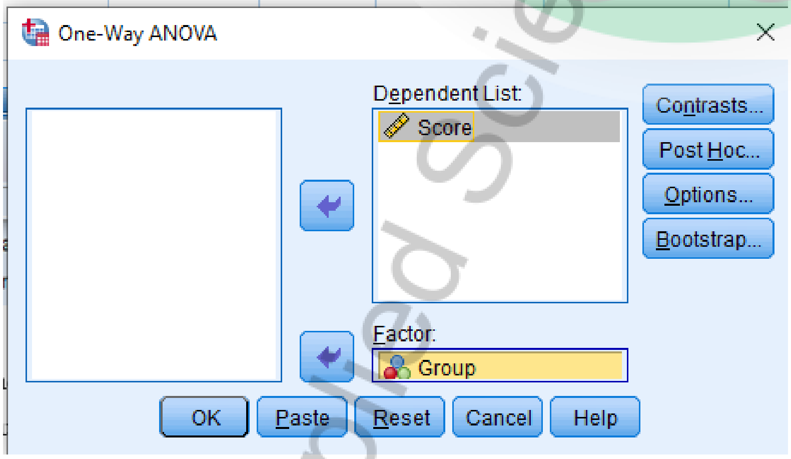
Step 3:
Add data in variable view

Note: For independent variable select Nominal and for dependent variable select scale option in measure column.

Now on top click on Analyze→ Click on compare mean → select one-way ANOVA → In independent variable add Group and in Dependent variable add Score. → Click on Post Hoc → check Tukey box → ok → now click on options → Check descriptive → Check exclude cases analysis by analysis → click Ok → Values will appear.



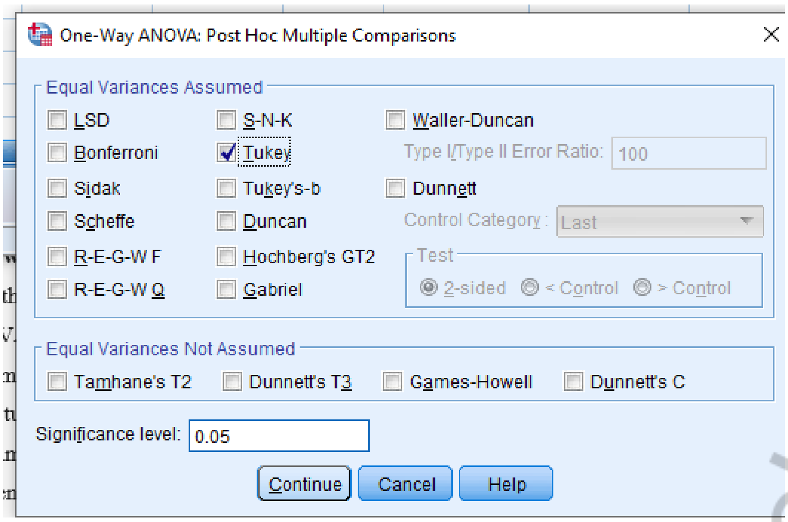
Step 4:
Go to Analyze and Click on One-way ANOVA



Dependent List: Score

Factor: Group

Step 5:
Add variables



One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

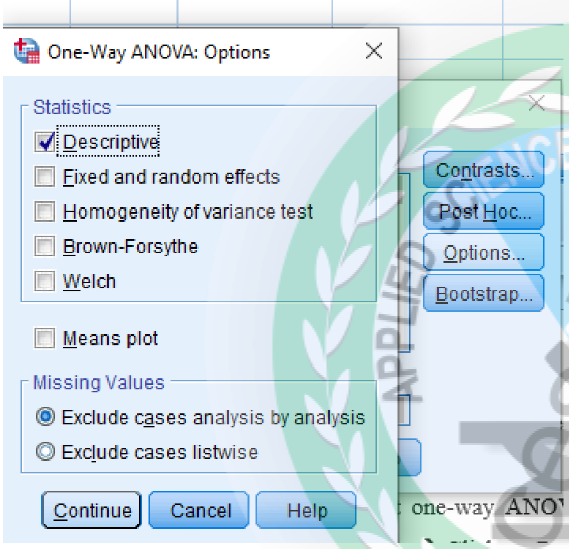
□ LSD	□ S-N-K	□ Waller-Duncan
□ Bonferroni	☑ Tukey	Type I/Type II Error Ratio: 100
□ Sidak	□ Tukey's-b	□ Dunnett
□ Scheffe	□ Duncan	Control Category: Last
□ R-E-G-W F	□ Hochberg's GT2	Test
□ R-E-G-W Q	□ Gabriel	● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C

Significance level: 0.05

Step 6:
Click Post Hoc and Select Tukey



One-Way ANOVA: Options

Statistics

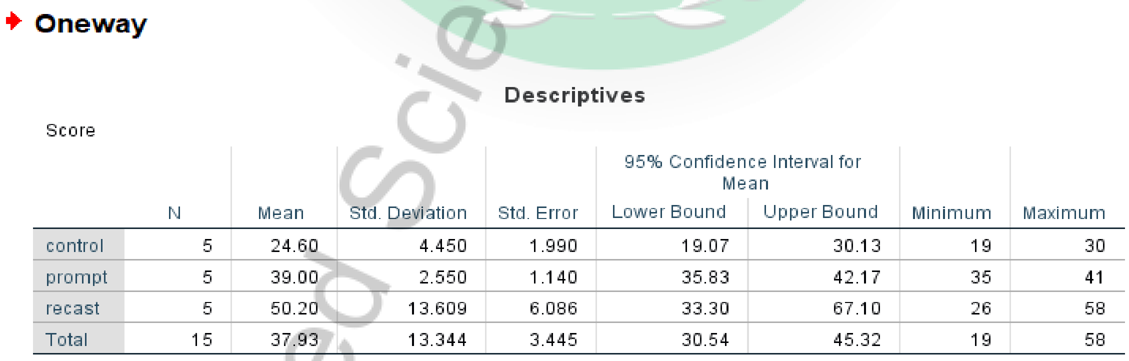
☑ Descriptive
□ Fixed and random effects
□ Homogeneity of variance test
□ Brown-Forsythe
□ Welch

□ Means plot

Missing Values

● Exclude cases analysis by analysis
○ Exclude cases listwise

Step 7:
Click options and select Descriptive

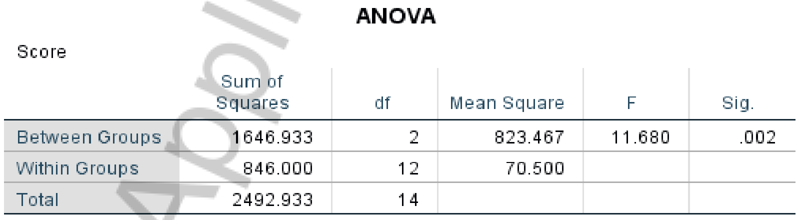


Oneway

Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
control	5	24.60	4.450	1.990	19.07	30.13	19	30
prompt	5	39.00	2.550	1.140	35.83	42.17	35	41
recast	5	50.20	13.609	6.086	33.30	67.10	26	58
Total	15	37.93	13.344	3.445	30.54	45.32	19	58

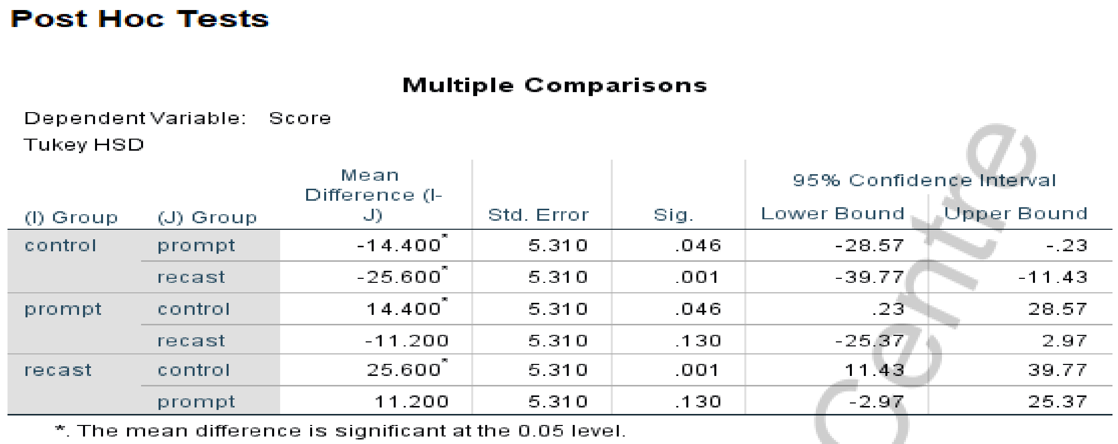


ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1646.933	2	823.467	11.680	.002
Within Groups	846.000	12	70.500		
Total	2492.933	14			

Step 8:
Values appear



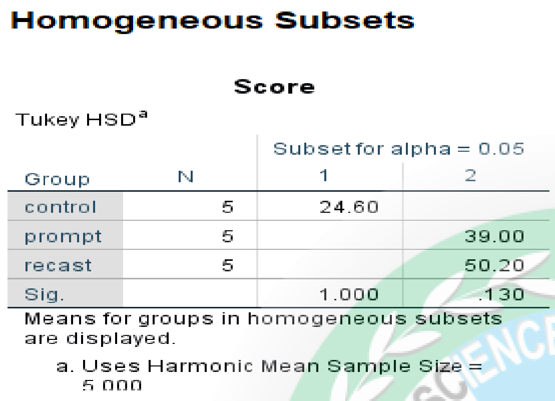
Post Hoc Tests

Multiple Comparisons

Dependent Variable: Score
Tukey HSD

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
control	prompt	-14.400*	5.310	.046	-28.57	-.23
	recast	-25.600*	5.310	.001	-39.77	-11.43
prompt	control	14.400*	5.310	.046	.23	28.57
	recast	-11.200	5.310	.130	-25.37	2.97
recast	control	25.600*	5.310	.001	11.43	39.77
	prompt	11.200	5.310	.130	-2.97	25.37

*. The mean difference is significant at the 0.05 level.



Homogeneous Subsets

Score

Tukey HSD[a]

Group	N	Subset for alpha = 0.05	
		1	2
control	5	24.60	
prompt	5		39.00
recast	5		50.20
Sig.		1.000	.130

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5.000

Step 9:
P value appear

Interpretation of results:

The value of p determines the level of significance of two or more groups, if the value of p is less than 0.005 then the statistical analysis is significant and if value is 0.001 then it is highly significant. P value above 0.005 indicates non-significant.

7.6. TWO-WAY ANALYSIS OF VARIANCE

It is applied to calculate the significant mean difference of more than three independent groups that are split on two variables (factors). Generally two-way ANOVA is applied when two factors affect the response variable and on the response variable there may or may not be an interaction effect between two factors.

In a study a scientists wants to check how sunlight and watering affect the plant growth. An experiment was set in which 20 plant seeds were grown in different sunlight conditions and water frequency. In this study:

Plant growth = Response variable

Sunlight and water = Factors

In this study there are two factors so for statistical analysis two-way ANOVA will be applied to check the significant mean difference of the two factors on response. If only one factor is applied then one-way ANOVA will be applied.

7.6.1. ASSUMPTIONS OF TWO-WAY ANOVA

Following are the assumptions of two-way analysis of variance:

- There is normal distribution of the response variable.
- The group variances should be approximately equal.
- Each group observations should be independent and obtained through random sampling.

Problem:

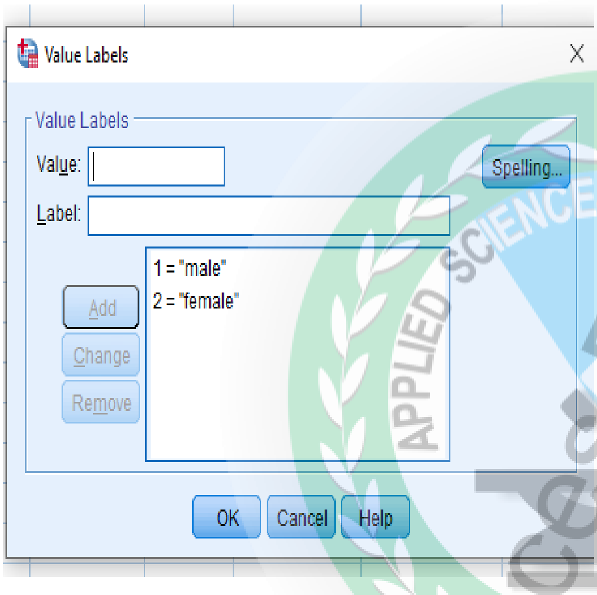
For the following dataset analyze the significance difference of mean between two factors.

ID	Gender	Group	Political-Interest-Score
1	Male	School	38
2	Male	School	40
3	Male	School	33
4	Male	School	36
5	Male	School	33
6	Female	School	37
7	Female	School	32
8	Female	School	30
9	Female	School	35
10	Female	School	36
11	Male	College	42
12	Male	College	48
13	Male	College	43
14	Male	College	46
15	Male	College	49
16	Female	College	44
17	Female	College	47
18	Female	College	41
19	Female	College	42
20	Female	College	45
21	Male	University	60
22	Male	University	58
23	Male	University	57
24	Male	University	52
25	Male	University	55
26	Female	University	53

27	Female	University	57
28	Female	University	51
29	Female	University	52
30	Female	University	54

Solution Key:

Enter data in data sheet → Click on variable view → Label two independent variable gender and education level → In gender column label 1 for male and 2 for female → In education level column label 1 for school, 2 for college and three for university level → the third column is for dependent variable that is Political-interest-score.

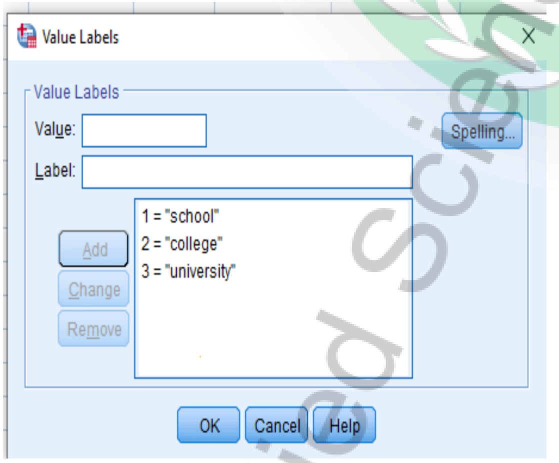


Value:

Label:

1 = "male"
2 = "female"

Step 1:
Data entry and value labels

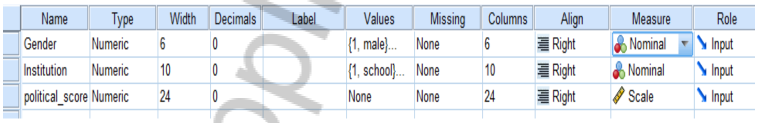


Value:

Label:

1 = "school"
2 = "college"
3 = "university"

Step 2:
Assign numeric to qualitative data

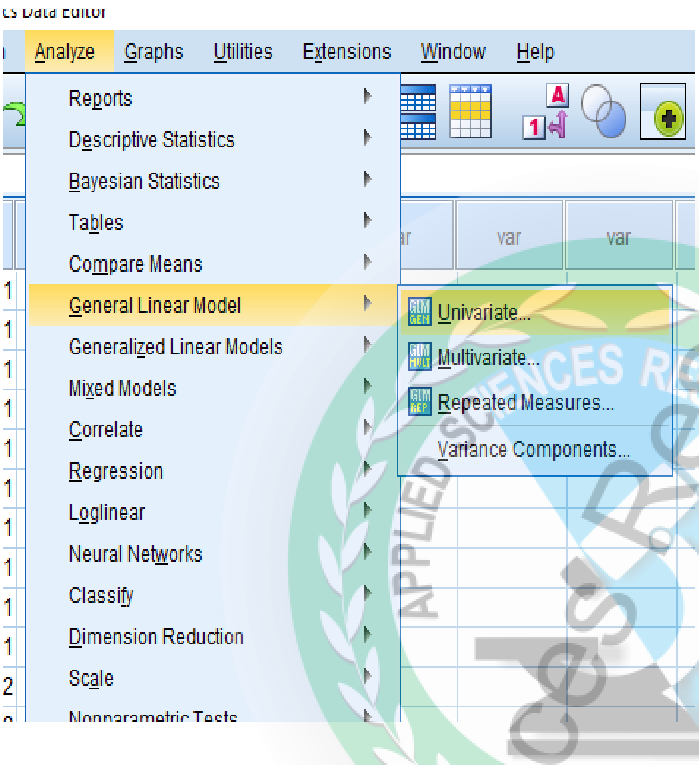


Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
Gender	Numeric	6	0		{1, male}...	None	6	Right	Nominal	Input
Institution	Numeric	10	0		{1, school}...	None	10	Right	Nominal	Input
political_score	Numeric	24	0		None	None	24	Right	Scale	Input

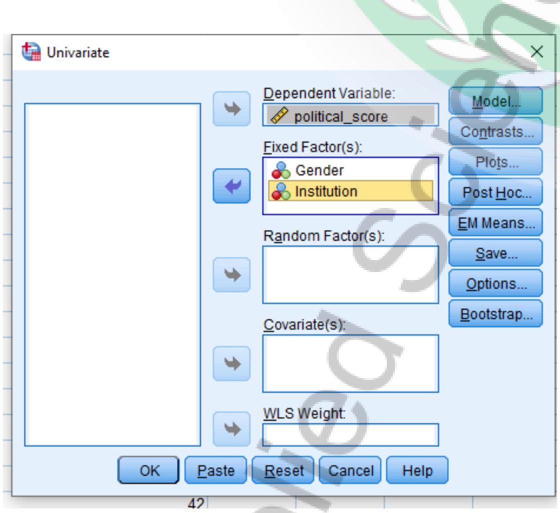
Step 3:
Modify data in variable view

Now click on Analyze → go to General Linear Model → select Univariate → place political-interest-score in dependent variable → gender and education-level in fixed factors - → Click OK →

Now click EM Means → add the three factors in display means for → check compare mean effect box → click OK → now click on options → select Descriptive statistics and homogeneity tests → significance level 0.05 → click continue → Analysis box appear.



Step 3:
Go to general linear model and select univariate



Dependent Variable: political_score

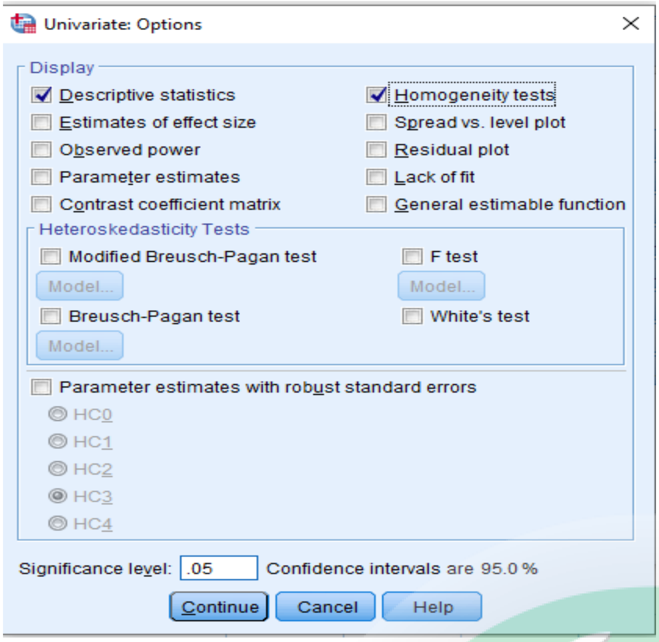
Fixed Factor(s): Gender, Institution

Random Factor(s):

Covariate(s):

WLS Weight:

Step 5:
Add Variables



Univariate: Options

Display

☑ Descriptive statistics
☐ Estimates of effect size
☐ Observed power
☐ Parameter estimates
☐ Contrast coefficient matrix
☑ Homogeneity tests
☐ Spread vs. level plot
☐ Residual plot
☐ Lack of fit
☐ General estimable function

Heteroskedasticity Tests

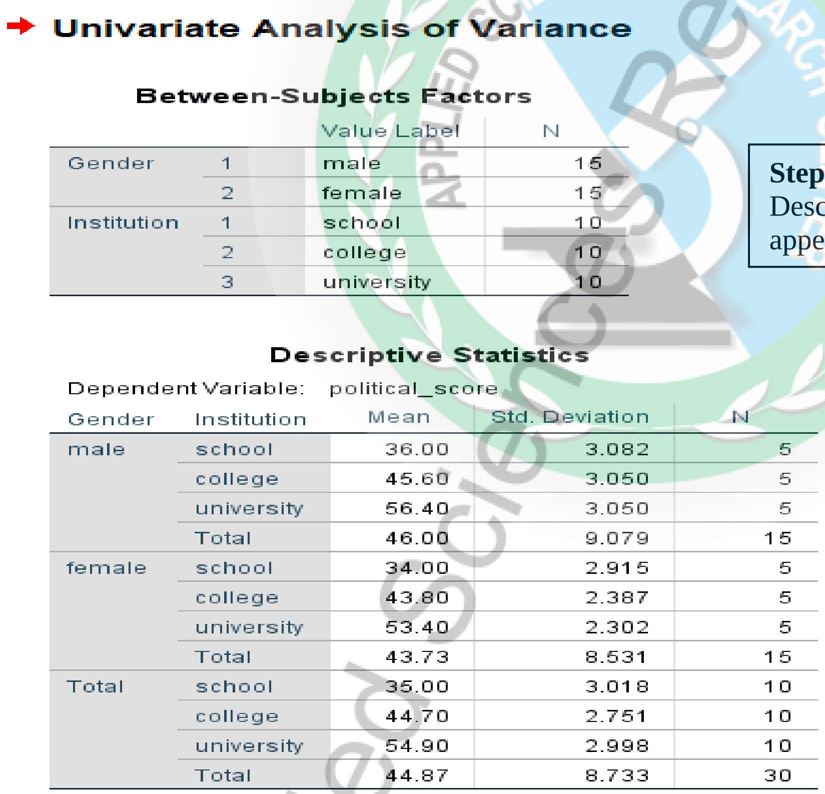
☐ Modified Breusch-Pagan test
☐ F test
☐ Breusch-Pagan test
☐ White's test

☐ Parameter estimates with robust standard errors

○ HC0
○ HC1
○ HC2
● HC3
○ HC4

Significance level: .05 Confidence intervals are 95.0 %

Step 6:
Click options and check descriptive statistics and homogeneity tests



Univariate Analysis of Variance

Between-Subjects Factors

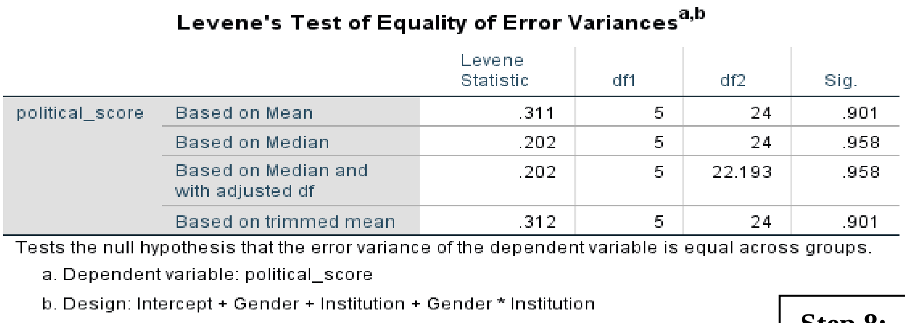
		Value Label	N
Gender	1	male	15
	2	female	15
Institution	1	school	10
	2	college	10
	3	university	10

Step 7:
Descriptive statistics appear

Descriptive Statistics

Dependent Variable: political_score

Gender	Institution	Mean	Std. Deviation	N
male	school	36.00	3.082	5
	college	45.60	3.050	5
	university	56.40	3.050	5
	Total	46.00	9.079	15
female	school	34.00	2.915	5
	college	43.80	2.387	5
	university	53.40	2.302	5
	Total	43.73	8.531	15
Total	school	35.00	3.018	10
	college	44.70	2.751	10
	university	54.90	2.998	10
	Total	44.87	8.733	30



Levene's Test of Equality of Error Variances[a,b]

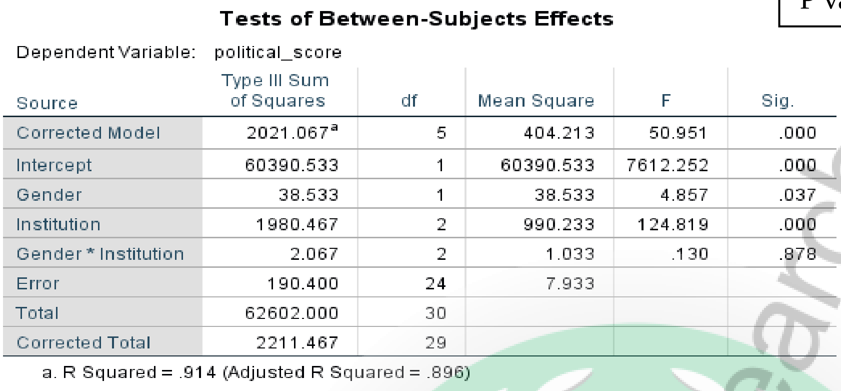
		Levene Statistic	df1	df2	Sig.
political_score	Based on Mean	.311	5	24	.901
	Based on Median	.202	5	24	.958
	Based on Median and with adjusted df	.202	5	22.193	.958
	Based on trimmed mean	.312	5	24	.901

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: political_score

b. Design: Intercept + Gender + Institution + Gender * Institution

Step 8:
P value appear

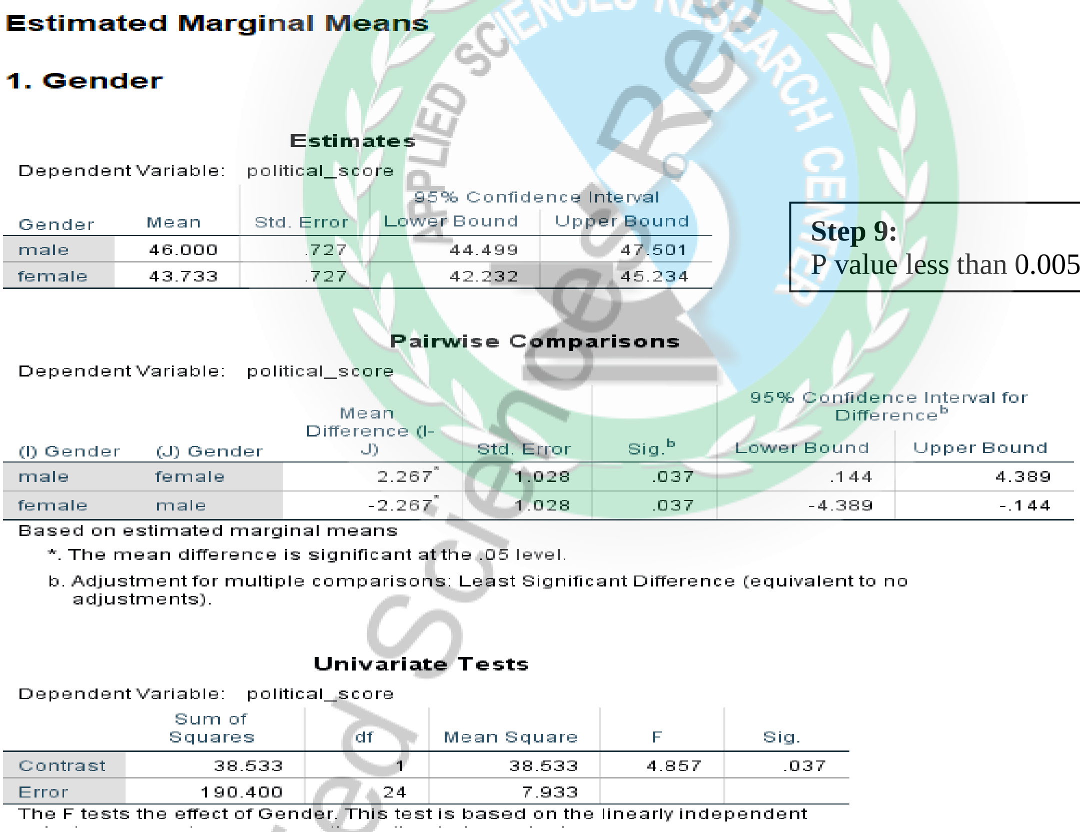


Tests of Between-Subjects Effects

Dependent Variable: political_score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2021.067[a]	5	404.213	50.951	.000
Intercept	60390.533	1	60390.533	7612.252	.000
Gender	38.533	1	38.533	4.857	.037
Institution	1980.467	2	990.233	124.819	.000
Gender * Institution	2.067	2	1.033	.130	.878
Error	190.400	24	7.933		
Total	62602.000	30			
Corrected Total	2211.467	29			

a. R Squared = .914 (Adjusted R Squared = .896)



Estimated Marginal Means

1. Gender

Estimates

Dependent Variable: political_score

Gender	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
male	46.000	.727	44.499	47.501
female	43.733	.727	42.232	45.234

Step 9:
P value less than 0.005

Pairwise Comparisons

Dependent Variable: political_score

(I) Gender	(J) Gender	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
male	female	2.267*	1.028	.037	.144	4.389
female	male	-2.267*	1.028	.037	-4.389	-.144

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

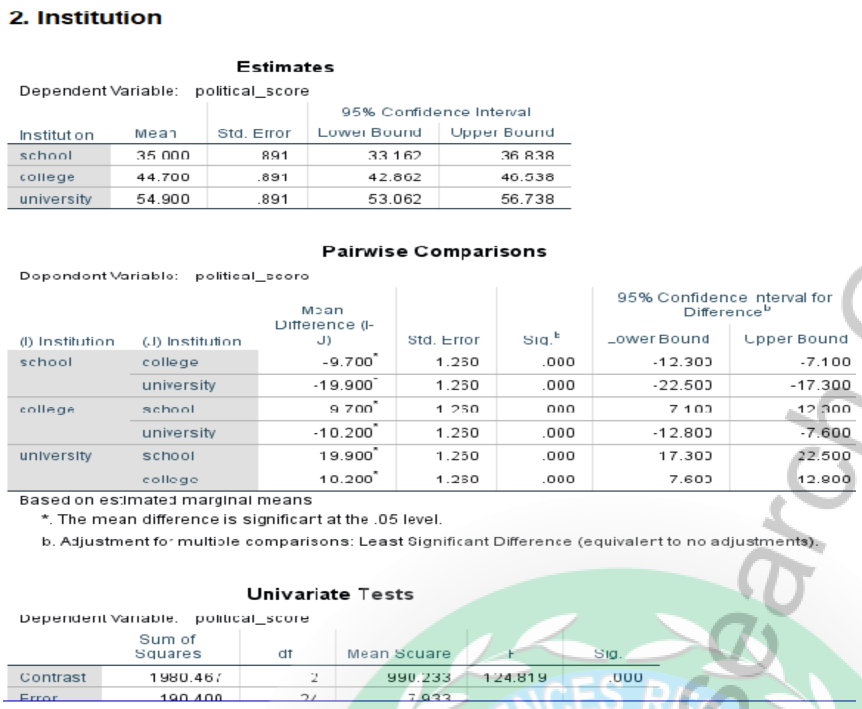
b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Univariate Tests

Dependent Variable: political_score

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	38.533	1	38.533	4.857	.037
Error	190.400	24	7.933		

The F tests the effect of Gender. This test is based on the linearly independent



2. Institution

Estimates

Dependent Variable: political_score

Institution	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
school	35.000	.891	33.162	36.838
college	44.700	.891	42.862	46.538
university	54.900	.891	53.062	56.738

Pairwise Comparisons

Dependent Variable: political_score

(I) Institution	(J) Institution	Mean Difference (I-J)	Std. Error	Sig.b	95% Confidence Interval for Differenceb	
					Lower Bound	Upper Bound
school	college	-9.700*	1.250	.000	-12.300	-7.100
	university	-19.900*	1.250	.000	-22.500	-17.300
college	school	9.700*	1.250	.000	7.100	12.300
	university	-10.200*	1.250	.000	-12.800	-7.600
university	school	19.900*	1.250	.000	17.300	22.500
	college	10.200*	1.250	.000	7.600	12.800

Based on estimated marginal means

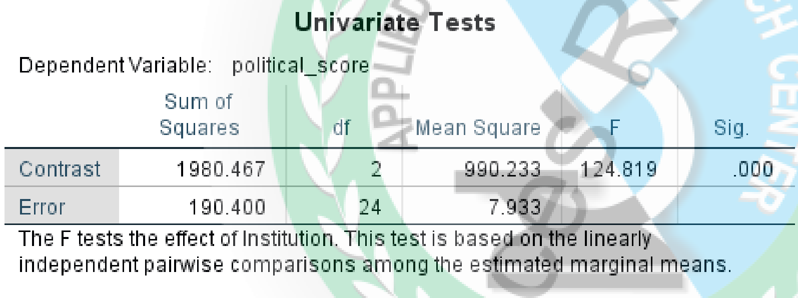
*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Univariate Tests

Dependent Variable: political_score

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	1980.467	2	990.233	124.819	.000
Error	190.400	24	7.933		

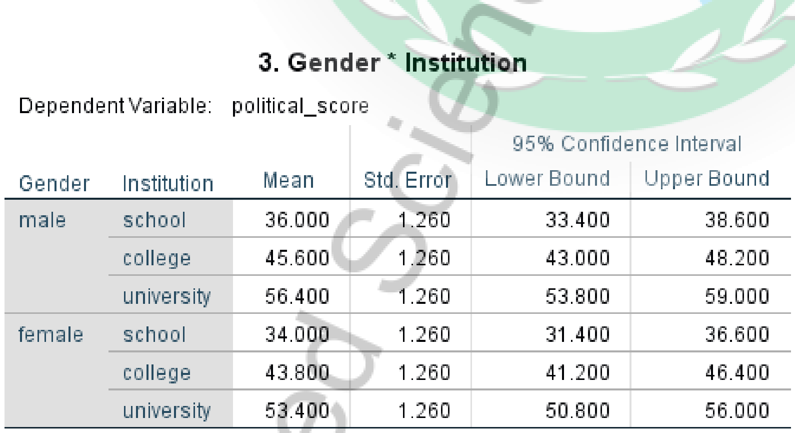


Univariate Tests

Dependent Variable: political_score

	Sum of Squares	df	Mean Square	F	Sig.
Contrast	1980.467	2	990.233	124.819	.000
Error	190.400	24	7.933		

The F tests the effect of Institution. This test is based on the linearly independent pairwise comparisons among the estimated marginal means.



3. Gender * Institution

Dependent Variable: political_score

Gender	Institution	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
male	school	36.000	1.260	33.400	38.600
	college	45.600	1.260	43.000	48.200
	university	56.400	1.260	53.800	59.000
female	school	34.000	1.260	31.400	36.600
	college	43.800	1.260	41.200	46.400
	university	53.400	1.260	50.800	56.000

Interpretation of result:

The value of p determines the level of significance of two or more groups, if the value of p is less than 0.005 then the statistical analysis is significant and if value is 0.001 then it is highly significant. P value above 0.005 indicates non-significant.

7.7. CORRELATION

It measures the relation between two variables. The correlation coefficient is used to measure the extent of relation between two variables. The most popular correlation coefficient is Pearson’s correlation and it is denoted as R. The correlation coefficient is used in linear regression.

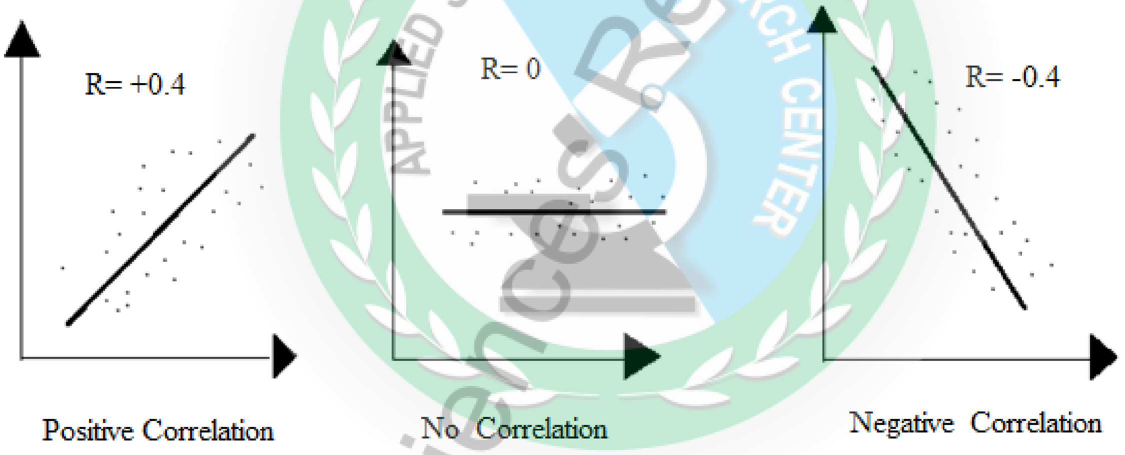
7.7.1. VALUE OF COEFFICIENT OF CORRELATION

The value of correlation coefficient ranges between -1 to +1

+1 value indicate strong positive relationship

-1 value indicate strong negative relationship

0 value indicate no relation between two variables



7.7.2. FORMULA OF COEFFICIENT OF CORRELATION

Following is the formula used to calculate coefficient of correlation:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

x = Value of x within sample

y = Value of y within sample

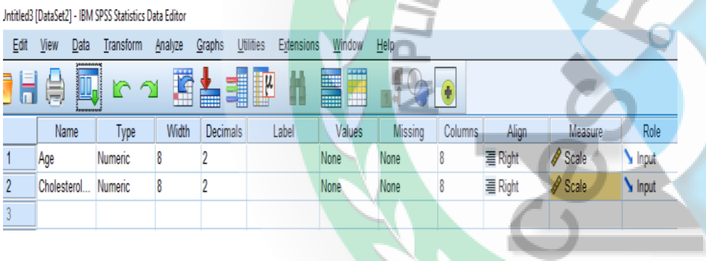
Example:

Calculate correlation coefficient of the following data using IBM-SPSS.

Observation	Age X	Glucose level Y	XY	X^2	Y^2
1	56	144	8064	3136	20736
2	37	120	4440	1369	14400
3	25	99	2475	625	9801
4	45	140	6300	2025	19600
5	53	135	7155	2809	18225
6	31	127	3937	961	16129
n=6	ΣX=247	ΣY=765	ΣXY=32,371	$\Sigma X^2=9964$	$\Sigma Y^2=98891$

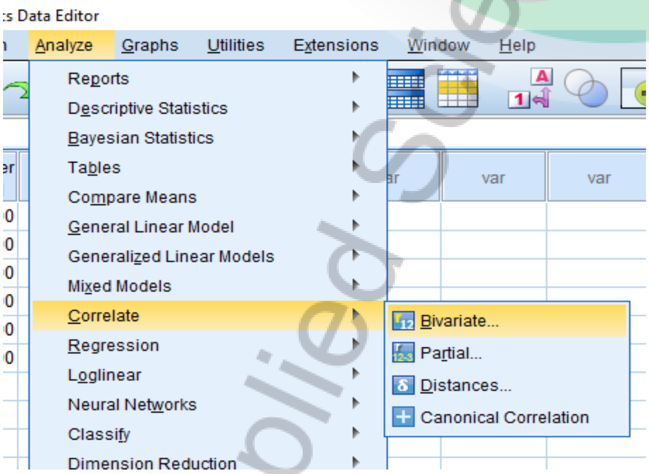
Solution Key:

Open data sheet → Enter values or X and Y → click variable view → change measurement to scale → click Analyze on top of Menu → go to correlation → bivariate → dialog box appear → click on variables on left side of box and drag the variables on right side → check the Pearson's box → check two-tailed box → click Ok → results appear on screen

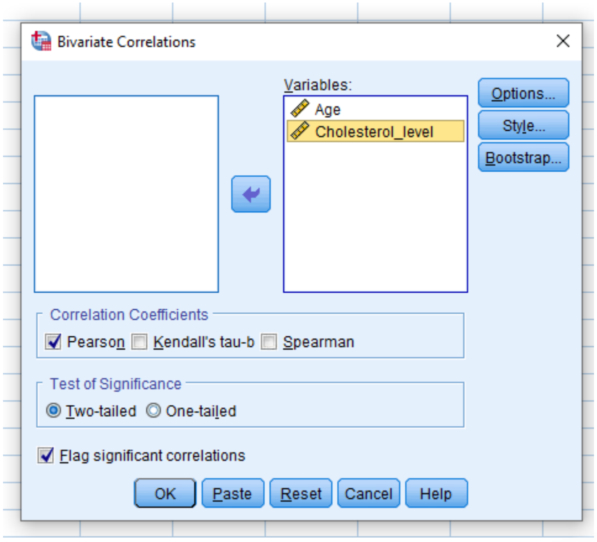

Jntitled3 [DataSet2] - IBM SPSS Statistics Data Editor
Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Age	Numeric	8	2		None	None	8	Right	Scale	Input
2	Cholesterol...	Numeric	8	2		None	None	8	Right	Scale	Input
3											

Step 1:
Enter data



Step 2:
Go to Analyze and select Correlate Bivariate



Bivariate Correlations

Variables:
- Age
- Cholesterol_level

Correlation Coefficients

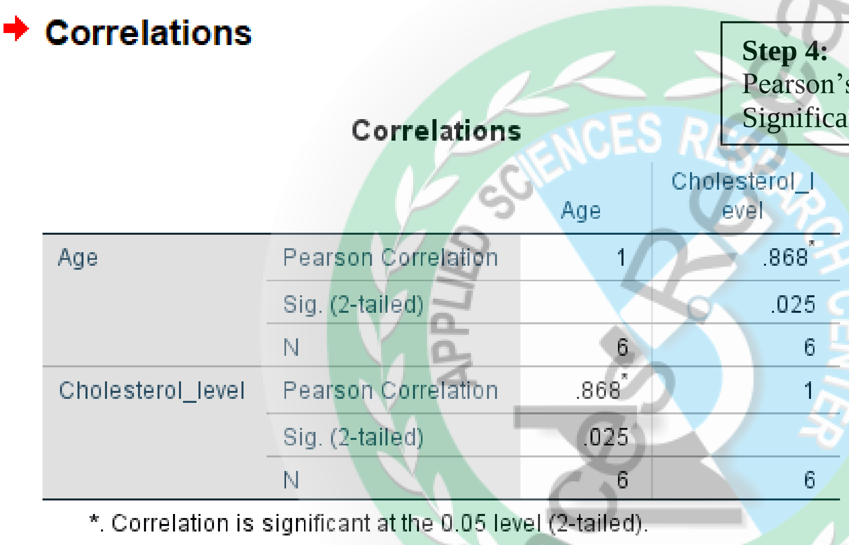
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

Step 3:
Add variables



Step 4:
Pearson’s Correlation and 2-tailed Significance appear

Correlations

Correlations

		Age	Cholesterol_level
Age	Pearson Correlation	1	.868*
	Sig. (2-tailed)		.025
	N	6	6
Cholesterol_level	Pearson Correlation	.868*	1
	Sig. (2-tailed)	.025	
	N	6	6

*. Correlation is significant at the 0.05 level (2-tailed).

Interpretation of result:

A Pearson correlation of -1 indicates a strong negative correlation, and a correlation value of 1.0 indicates strong positive correlation. If the correlation coefficient is greater than zero, it is a positive relationship. Conversely, if the value is less than zero, it is a negative relationship. A zero value reflects no relation between two variables. Two-tailed significance is less than 0.05 which indicate significance level.

Applied Sciences Research Centre
Copyrights 2024