# BEYOND DATA ANALYTICS

**P.O Box 109-60400, Nairobi**

**Email: beyonddataanalytics@gmail.com**

## SHORT COURSE IN MODELING AND DATA ANALYTIC TECHNIQUES USING

- R & RSTUDIO
- PYTHON
- STATA
- SPSS
- Excel
- MICROSOFT EXCEL

## TABLE OF CONTENT

**STATISTICAL ANALYSIS USING STATISTICAL SOFTWARES** .................................... 3

**1.1. Course Overview** .................................... 3

**1.2. Course Description** .................................... 3

**1.3. Rationale** .................................... 3

**1.4. Course Content** .................................... 3

**1.5. Purpose of short course** .................................... 3

**1.6. Target Group** .................................... 4

**1.7. Course Outcomes** .................................... 4

**1.8. Teaching Strategies** .................................... 4

**1.9. Course Materials and Equipment** .................................... 5

**1.10. Model of Delivery** .................................... 5

**1.11. Course Venue** .................................... 6

**1.12. Statistical Training Course outlines** .................................... 7

**EXCEL Training Outline** .................................... 7

**SPSS Training Outline** .................................... 10

**STATA Training Outline** .................................... 13

**R & Rstudio Training Outline** .................................... 18

**References** .................................... 23

# STATISTICAL ANALYSIS USING STATISTICAL SOFTWARES

## 1.1. Course Overview

This course aims to provide researcher and policy with essential statistical knowledge in order to be able to understanding the emerging statistical concepts and information. Besides, this course will provide a foundation of key techniques, concepts to understand and navigate through large data set.

## 1.2. Course Description

This is course is designed for both beginners and those at an advanced level. The course is designed to cover a wide range of concepts from the basic statistical concepts to a more advanced concepts and methods. The course will be taught using quite a number of statistical software programs including but not limited to SPSS, STATA, Rstudio and Python, however, the statistical programming software of interests include Rstudio and Python. This course aims to students, equip researcher and policy makers with key skills of data handling and management. It is important to note that data visualization which is one of the primary techniques covered in the course is a key statistical computation since such technique helps illustrate statistical findings in a more insightful manner. Furthermore, this course will introduce students the most emerging statistical concept of machine learning algorithms, which is helpful in big data analytics. Machine learning concepts covered in this course include both supervised and unsupervised. The final touches of the course cover the practical hand-on exercise to create a dynamic learning environment.

## 1.3. Rationale

A structured study for statistical analysis, data visualization and machine leaning algorithms will be developed. These are key statistical techniques and approaches to data analytics. The course is designed to provide students, policy makers and researchers a hands-on exercise and experience to maneuver through the industrial needs and requirement by filling the gap between the theoretical concepts and practical procedures.

## 1.4. Course Content

- ❖ Descriptive Statistics
- ❖ Data Visualization
- ❖ Parametric methods
- ❖ Linear Regression and correlation analysis
- ❖ Non-linear model
- ❖ Design and Analysis of Experiments
- ❖ Non-Parametric Tests
- ❖ Machine Learning Techniques
  - ✓ KNN
  - ✓ Random forest
  - ✓ Vector support machine

## 1.5. Purpose of short course

The course primary purpose is to enable students, policymakers, and researchers analyze, interpret and report statistical findings in a more meaning and insightful manner to be able to make informed decisions. It is worth noting that data has become a very important focal point of decision making and therefore any decision made has to be backed up by data and statistical findings.

### 1.6.Target Group

The target group include but not limited to;

- ❖ Business Professionals
- ❖ Students
- ❖ Entrepreneurs and Small Business Owners
- ❖ Government and Nonprofit Professionals
- ❖ Marketing and Sales Professionals
- ❖ Healthcare Professionals
- ❖ Educators and Researchers
- ❖ Finance and Banking Professionals
- ❖ Supply Chain and Logistics Professionals

### 1.7. Course Outcomes

The specific outcomes of the course include but not limited to;

- ❖ Understanding of Data Concepts
- ❖ Data Collection and Cleaning
- ❖ Statistical Analysis
- ❖ Data Visualization
- ❖ Programming Skills
- ❖ Machine Learning and Predictive Analytics
- ❖ Ethical and Legal Considerations
- ❖ Communication Skills
- ❖ Problem-Solving Skills
- ❖ Industry Applications

### 1.8. Teaching Strategies

1. **Hands-On Labs**
   - Provide opportunities for students to work with real datasets using popular analytics tools like Python, R & Studio, Gretl and guide them through practical exercises to reinforce theoretical concepts.
2. **Interactive Lectures**
   - Foster class discussions and encourage students to ask questions and use interactive tools and platforms to engage students actively during lectures.
3. **Data Visualization Techniques**
   - Emphasize the importance of effective data visualization in conveying insights and teach students how to use tools like ggplot2, tidyverse, dplyr, Matplotlib, Seaborn, or Tableau for creating compelling visualizations
4. **Continuous Assessment**
   - Implement regular quizzes, assignments, and assessments to gauge student understanding and provide constructive feedback to help students improve
5. **Collaborative Learning**
   - Promote teamwork and collaboration on data analytics projects and encourage students to share insights, code, and collaborate on problem-solving

## 1.9. Course Materials and Equipment

**Educational resources**

1. **Textbooks and Reference Materials**
   - "Python for Data Analysis" by Wes McKinney
   - "The Art of Data Science"  by Roger D. Peng and Elizabeth Matsui
2. **Interactive Platforms**
   - Rstudio and Jupyter Notebooks and Visual studio code: Ideal for experimenting with code, visualizing data, and creating data narratives
   - Kaggle: A platform for data science competitions, collaborative coding, and datasets.
3. **Data Analytics Chuka University website**

**Software Tools**

1. **Programming Languages**
   - Python: Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn.
   - R Programming: Especially useful for statistical analysis
2. **Database Management**
   - SQL (Structured Query Language): Learn to work with relational databases.
3. **Data Visualization**
   - ggplot2, Tableau, Power BI, Matplotlib, Seaborn for creating insightful visualizations.
4. **Version Control**
   - Git and GitHub for collaborative coding and version control.
5. **Integrated Development Environment (IDE)**
   - Jupyter Notebooks, Spyder, or Visual studio Code for coding and analysis.
6. **Statistical Package for Social Sciences**
7. **STATA**

**Hardware Equipment**

1. **Computers with Sufficient Processing Power**
   - Laptops or desktops with a reasonably powerful CPU and sufficient RAM.
2. **External Storage**
   - An external hard drive for backing up large datasets and projects.
3. **Cloud Computing Services**
   - Consider using cloud platforms like AWS, Google Cloud, or Microsoft Azure for more extensive computing needs.
4. **Smart screen**
   - ICT equipment for recording and video documentation and online training

## 1.10. Model of Delivery

**Training and Workshops**

Training sessions and workshops on statistical analysis techniques and skills to understand and act upon analytics insights.

**Reports and Dashboards**

Data analytics insights are delivered through interactive dashboards and static reports to visualize key metrics, trends, and performance indicators.

### Collaboration Platforms

Collaborative platforms like Microsoft Teams, google meet and KENET

## 1.11. Course Venue

Data Analytics Training Center at Chuka University

## 1.12. Statistical Training Course outlines

## EXCEL Training Outline

| | Topic | Description | Detailed Description | Estimated Hours |
|---|---|---|---|---|
| INTRODUCTION | Introduction | An introduction to Microsoft Excel and its environment | Microsoft Excel & spreadsheet software used for data management, analysis, and visualization | 1 |
| | Data Analysis Tool Kit Activation | Data analytics in Microsoft Excel is made possible through Data Analysis ToolPak | Data Analysis ToolPak in Excel, which provides additional data analysis tools such as regression analysis, histograms, and t-tests | 1 |
| | Data Entry in Excel | Manual data entry and importation into Excel will be done | Manual data entry, importing data from external sources, and copy-pasting data from other applications | 1 |
| DESCRIPTIVE STATISTICS | Descriptive Statistics | Data summarization and presentation | Descriptive statistics in Excel, which involves summarizing and presenting data in a meaningful way | 1 |
| | Measures of Central Tendency | Mean<br>Median<br>Mode<br>Deciles<br>Quartiles<br>Percentiles | This section covers measures of central tendency which include the mean, median mode, deciles, percentiles and Quartiles | 1 |
| | Measures of Dispersion | Range,<br>Variance<br>Standard Deviation<br>Interquartile Range | This section of this course discusses measures of dispersion and variability to measure how the data is dispersed. Some of these measures include range, variance, standard deviation and interquartile range. | 1.5 |
| | Frequency Distributions | Frequency curves<br>Frequency polygon<br>Ogives | This section covers distribution methods such as summarizing categorical or continuous data. Besides, this section introduces concepts of frequency distribution and creation of frequency distribution | 1.5 |
| | Measures of Shape | Skewness<br>Kurtosis, | This section of this course discusses the skewness and kurtosis statistics which are measure of distribution of the data which shows whether the data is positively, negatively or normal distribution. Kurtosis on the other shows whether the data has a | 1 |

| | | | | |
|---|---|---|---|---|
| | | | mesokurtic, leptokurtic and platykurtic distribution. | |
| **DATA VISUALIZATION** | Data Visualization | Introduction to data visualization | Introduction to data visualization in Excel is done in this section to obtain meaning insight from the data | 1 |
| | Histogram | Distribution of the data | Graphical representations of the frequency distribution of a dataset | 1 |
| | Scatter Plot | Two-way relationship between continuous variables | Display the relationship between two continuous variables through individual data points | 1 |
| | Pie Charts | Showing the distribution of both categorical and continuous data | Covers pie charts as visualizations for displaying categorical data and comparing proportions or percentages. | 1 |
| | Bar Graphs | Showing the distribution of both categorical and continuous data | Bars chart to compare values across different categories or groups as well continuous variables | 1 |
| | Box Plot | Box plots in Excel, also known as box-and-whisker plots, | Box plot provides descriptive statistics such as median, first quartile, third quartile, minimum and maximum | 1 |
| **STUDEN T T-** | Student t-test | Hypothesis testing using Student T-test | One Sample T-Test<br>Two Sample T-Test<br>Paired T-Test<br>Independent T-Test | 1 |
| **ANOVA AND OTHER TESTS** | ANOVA | Introduction to Analysis of Variance (ANOVA) | ANOVA (One-Way ANOVA, Two-Way ANOVA) | 1.5 |
| | F-Test | Hypothesis testing using F-test | Explanation of the F-test in Excel, which is used in ANOVA to assess whether the means of two or more groups are significantly different | 1 |
| | Chi-Square | Hypothesis testing for categorical data | Introduction to hypothesis testing using the Chi-Square test in Excel, which is used to determine whether there is a significant association between categorical variables | 1 |

|  |  |  |  |  |
|---|---|---|---|---|
| **LINEAR REGRESSION ANALYSIS AND CORRELATION ANALYSIS** | Linear Regression Analysis and Correlation Analysis | Introduction to Regression analysis. | Introduction to linear regression analysis and correlation analysis in Excel, which involve examining the relationships between variables and making predictions based on these relationships | 1.5 |
|  | Regression Analysis: Simple Linear Regression Analysis | Introduction Simple Linear Regression analysis |  | 1 |
|  | Regression Analysis: Multiple Linear Regression Analysis | An introduction to multiple linear regression analysis | Explanation of simple linear regression analysis in Excel which assess the linear effect of multiple independent variables used to predict the dependent variable | 1.5 |
|  | Correlation Analysis | The measure of relatedness between two or more categorical variables | Introduction to correlation analysis, a measure of relatedness of two continuous variables. | 1 |
|  | Data Interpretation and Reporting | Interpretation of the data to draw conclusion. | Explanation of data interpretation and reporting in Excel, which involves summarizing findings, drawing conclusions, and communicating results effectively | 1 |

## SPSS Training Outline

| | Topic | Description | Detailed Description | Estimated Hours |
|---|---|---|---|---|
| INTRODUCTION | Introduction to SPSS | Introduction to SPSS environment | Introduction to SPSS (Statistical Package for the Social Sciences) | 1 |
| | SPSS Window | Introduction to SPSS Window | Explanation of the main windows and interface elements in SPSS, including the Data Editor, Variable View, and Output Viewer | 1 |
| | Data Entry and Coding | Introduction to data entry and coding categorical variables | Methods for entering and coding data in SPSS, including manual data entry, importing data from external sources, and assigning variable labels and values. | 1.5 |
| | Descriptive Statistics | Introduction to descriptive statistics | Introduction to descriptive statistics in SPSS, including frequency distributions, cross-tabulations, and summary measures. | 1.5 |
| INFERENTIA STATISTICS | Inferential Statistics | Introduction to inferential statistics and hypothesis testing | Introduction to inferential statistics in SPSS, which involves making inferences or predictions about a population based on sample data | 1 |
| | Chi Square | Explanation of the chi-square test in SPSS, used to determine whether there is a significant association between categorical variables | Covers the chi-square test as a non-parametric test for analyzing categorical data. Discusses how to conduct chi-square tests in SPSS, interpret the results, and assess the strength of associations between variables. Provides examples and practical guidance for performing chi-square tests in SPSS. | 1.5 |
| | T-test | Introduction to T-test for hypothesis testing | Overview of the t-test in SPSS, including both the paired sample t-test and the independent sample t-test | 1.5 |
| DETAILED OUTLINE FOR T-TEST | T-test in SPSS | Introduction to the T-test in SPSS | This section of this course introduces T-test in SPSS and how to test the hypothesis using one sample, or two sample t-test | 1.5 |
| | One-Sample T-Test | Hypothesis testing using one sample t-test. | This section introduction to one sample t-test. One sample t-test is used to compares the hypothesized mean to the true population mean | 1 |

| | | | | |
|---|---|---|---|---|
| | Paired Sample T-Test | Hypothesis testing using paired sample | Paired sample t-test is used to compared the means of two related sample, e.g., the mean weight before and after intervention | 1.5 |
| | Independent Sample T-Test | Hypothesis testing using Unpaired sample | This technique is called independent t-test where the means of two independent sample (unpaired) are compared to determine if there exist a significant difference | 1.5 |
| **LINEAR REGRESSION AND CORRELATION ANALYSIS** | Linear Regression and Correlation Analysis | Introduction to linear regression analysis | This section of this course introduces the history and concepts regression and correlational analysis. Linear regression analysis measures the linear effect of one or more predictors on the predicted variable. | 1.5 |
| | Simple Linear Regression Analysis | Simple Linear Regression Analysis | Simple linear regression analysis measures the linear effect of one independent variable on one predicted variable. | 1 |
| | Multiple Linear Regression Analysis | Introduction to multiple linear regression analysis | Multiple linear regression analysis is an extension of simple linear regression analysis. In this technique the linear effect of multiple independent variables on dependent variable. | 1.5 |
| | Model Diagnostic | Model diagnostic to measure the efficacy and usefulness of the regression model | Explanation of model diagnostics in linear regression analysis, including assessing assumptions, detecting outliers, and evaluating model fit | 1 |
| | Pearson Correlation Coefficient | This is a measure of how two continuous variables are related. | The Pearson correlation coefficient is a measure of the strength and direction of the linear relationship between two continuous variables | 1 |
| **DESIGN AND ANALYSIS OFEXPERIMENT** | Design and Analysis of Experiment | Introduction to design and analysis of experiments | Introduction to the design and analysis of experiments in SPSS, which involves planning and conducting experiments to test hypotheses and make inferences about the effects of one or more factors on a response variable | 1.5 |
| | Complete Randomized Design (CRD) | This is also known as one factor ANOVA | Explanation of the complete randomized design (CRD) in SPSS, where treatments | 1.5 |

| | | | | |
|---|---|---|---|---|
| | | | are randomly assigned to experimental units | |
| | Randomized Complete Block Design (RCBD) | This technique is known as two-factor ANOVA | Randomized complete block design (RCBD) involves grouping experimental units into blocks and randomizing treatments within blocks | 1.5 |
| | Factorial Design of Experiment | This is an extension of two factor ANOVA | Explanation of factorial design of experiments is an extension of two-factor ANOVA, where multiple factors and their interactions are studied simultaneously | 1.5 |
| | Repeated Measure ANOVA | An introduction to repeated measure ANOVA | The repeated measure ANOVA is used when the same participants are measured under different conditions or at multiple time points | 1.5 |
| | Multivariate Analysis of Variance (MANOVA) | Introduction to MANOVA | Multivariate analysis of variance (MANOVA) allows for the simultaneous analysis of multiple dependent variables | 1.5 |
| NON-PARAMETRIC TESTS | Non-parametric Tests | Introduction to non-parametric analysis | Introduction to non-parametric tests in SPSS, which are used when the assumptions of parametric tests are violated or data are not normally distributed | 1.5 |
| | Wilcoxon Signed-Rank Test | Compares the distribution and the median of two related sample | Explanation of the Wilcoxon signed-rank test in SPSS, a non-parametric test used to compare two related samples or to test the median difference between paired observations | 1 |
| | Mann-Whitney U Test | Compares the distribution and the median of two independent sample | The Mann-Whitney U test, which is a non-parametric test used to compare the distributions of two independent samples | 1 |
| | Kruskal-Wallis Test | This is an extension of Mann-Whitney U-test and compares the distribution and the medians of three or mores independent sample | The Kruskal-Wallis test is a non-parametric test used to compare the distributions of three or more independent samples | 1 |

## STATA Training Outline

| | Topic | Description | Detailed Description | Estimated Hours |
|---|---|---|---|---|
| INTRODUCTION | Introduction | Overview of stata and its application | Introduction to Stata software and its applications in data analysis | 1 |
| | Stata Windows | Introduction to stata windows | Explanation of the main windows and interface elements in Stata, including the Command window, Results window, and Review and Property window | 0.5 |
| | Creating a Log File | Procedure and steps for log file creation and | Steps for creating a log file in Stata to record commands and results for reproducibility and documentation purposes | 0.5 |
| | Do-file | Introduction to log file creation | Introduction to the concept of a do-file in Stata, a script file containing Stata commands for automating analyses and ensuring reproducibility | 1 |
| | Data Entry | Data entry and importation into data | Methods for entering data into Stata, including importing from Excel files and direct data entry | 1.5 |
| Basic STATA Computation | String Variable | Creating string variables in Stata | Steps for creating a new string variable in Stata, typically used for storing text data such as names, categories, or labels | 0.5 |
| | Creating Variable | Creating another variable in Stata | Instructions for creating a new numeric or string variable based on existing variables in the dataset | 0.5 |
| | Value Labels | Creating value label in Stata | Guidance on assigning value labels to categorical variables in Stata, facilitating easier interpretation and analysis of the data | 0.5 |
| | New Labels | Adding new label to the categorical variable in Stata | Steps for adding new labels to existing value labels in Stata, updating the labels associated with specific numeric values | 0.5 |
| | Variables Data Types | Changing variable type in Stata | Instructions for converting variables from string to numeric data type or vice versa, adjusting the variable type to match its intended use | 0.5 |
| | Data Sorting | Sorting data in Stata | Demonstration of sorting observations in a dataset based on one or more variables, arranging data in ascending or descending order for analysis | 0.5 |

| | | | | |
|---|---|---|---|---|
| Summary statistics | Summary Statistics | Overview of summary statistics | Introduction to summary statistics in data analysis, providing insights into the central tendency, dispersion, and distribution of variables. | 1.5 |
| | Descriptive Statistics | Introduction to descriptive statistics | Explanation of descriptive statistics, which summarize and describe the basic features of a dataset | 1 |
| | Frequencies | Frequencies tables and distribution | Introduction to frequency distributions, which show the number of occurrences of different values in a dataset | 1 |
| | Cross Tabulation | Cross tabulation for two-way categorical data | Explanation of cross-tabulation, a method for summarizing and analyzing the relationship between two categorical variables | 1 |
| | Correlation and Causation | Correlations and causation and their differences | Differentiation between correlation and causation, emphasizing the importance of correlation analysis in identifying relationships between variables | 1.5 |
| Student t-test | Student t-test | Introduction to Student t-test in Stata | The student’s t-test, a statistical method used to determine if there is a significant difference between the means of two groups | 1.5 |
| | Independent T-Test | Independent sample T-test | Detailed description of the independent t-test, also known as the unpaired t-test, used to compare the means of two independent groups | 1 |
| | Dependent T-Test | Dependent sample t-test | Overview of the dependent t-test, also called the paired t-test, used to compare the means of two related groups | 1 |
| | One-Sample T-Test | One sample t-test comparing the hypothesized mean true population mean | Introduction to the one-sample t-test, used to determine if the true population mean of a single sample differs significantly from a known or hypothesized value | 1 |
| Non-parametric | Non-parametric Tests | Overview of Non-parametric tests | Introduction to non-parametric tests, which are used when data do not meet the assumptions of parametric tests | 1.5 |

| | | | | |
|---|---|---|---|---|
| | Rank Sum Test | Wilcoxon and Mann-Whitney U-test | The rank sum test, which includes the Wilcoxon signed-rank test and the Mann-Whitney U test | 1 |
| | - Wilcoxon Test | Compares the distribution and the median of paired sample | The Wilcoxon signed-rank test, a non-parametric test used to compare paired samples | 0.5 |
| | - Mann-Whitney Test | Compared the distribution and the median of unpaired sample | Introduction to the Mann-Whitney U test, a non-parametric test used to compare independent samples | 0.5 |
| | The H-Test | Compares the distribution of more than three independent sample or groups | The H-test, also known as the Kruskal-Wallis test is an extension of Mann-Whitney U-test, used to compare the distributions of three or more independent groups | 1 |
| | Chi-Square Test | Association between categorical data | The chi-square test, a non-parametric test used to determine whether there is a significant association between two categorical variables | 1 |
| Regression analysis | Regression Analysis | Introduction to regression analysis | Regression analysis is a statistical method used to model the relationship between one or more independent variables and a dependent variable | 2 |
| | Simple Linear Regression | The basic linear model; SLM | In a simple linear regression, the linear effect of one independent variable is tested on the dependent variables | 1.5 |
| | Multiple Linear Regression | Extension of Simple Linear Model (SML) | In a multiple linear regression model, which is an extension of Simple Linear Model, the linear effect of multiple predictors on the predicted variable is estimated. | 2 |
| | Output Interpretation | Model interpretation and reporting | Interpretation of regression analysis output, including regression coefficients, t-test statistics, F-test statistics, coefficient of determination, and adjusted R-squared (R^2) | 2 |

| | | | | |
|---|---|---|---|---|
| | Model Diagnostics | Assessing the usefulness and efficacy of the estimated model | Model diagnostic involves conducting several statistical tests to determine the usefulness and efficacy of the regression model. Some of the statistical tests conducted in this stage aims at identifying econometric problems such as multicollinearity, autocorrelation and heteroscedasticity if present. | 2 |
| Normality Tests | Normality Tests | Normality of the residuals | One of the assumptions of the classical linear models is that the residuals are normally distributed with a mean of zero and a constant variance. | 1.5 |
| | Skewness and Kurtosis | Assessing the distribution of the data | The skewness and kurtosis statistics measure the distribution of the data. Skewness tells whether the data is positively, negatively or normally distributed. On the other hand, kurtosis tells whether the data has mesokurtic, leptokurtic and platykurtic distribution. | 1 |
| | Shapiro-Wilk Test | Assessing the normality of the data | Shapiro-Wilks's statistics shows whether the data is normally distributed or deviates from normality. | 1 |
| Design and Analysis of Experiment | Design and Analysis of Experiment | Introduction to design and analysis of experiments | This section discusses more about the importance of scientific research and to minimize bias as we try to establish causality. Besides, the section will illustrate on how to plan and carry out scientific experimental research | 1.5 |
| | Definition | Definition of design and analysis of experiment. | This section defined design and analysis of experiment as a scientific inquiry. Further, this section emphasizes on what to put into consideration when designing a scientific experiment to avoid or minimize bias in a scientific inquiry. | 1 |
| | Principles of Experimental Design | Fundamental principles of scientific inquiry | Several principles of design of experiment and scientific experiment are discussed in this section, which include randomization, replication, blocking and control among others. | 2 |
| | - Randomization | Randomly assigning experimental units to treatment group | Randomization principle is key and very important in design and analysis of experiment. This principle helps in minimizing or avoiding bias in assigning experimental units into the treatment group. | 0.5 |
| | - Replication | Replicability of experiment. | One of the assumptions of the scientific inquiry if replicability. It's assumed that an experiment can be replicated using the same methodology several times and obtain similar. This is made possible by repeating the experiment. This is done to ensure consistency and minimize variability. | 0.5 |

| | | | | |
|---|---|---|---|---|
| | - Blocking | Grouping of the experimental units. | This is done to ensure that the experimental units are grouped into homogenous blocks. This is specifically done to account for variability. | 0.5 |
| | - Control | Isolation of the effect of the variable of interest | This section discusses the effect of creating the control variable in an experimental design. This process to hold the effect of some variable constant, especially the variable of interest | 0.5 |
| **Experimental Designs** | Experimental Designs | Introduction to experimental designs in Stata | This section discusses several experimental designs including but not limited Completely Randomized Design (CRD), Randomized Complete Block Design (RCBD) and Multivariate Analysis of Variance (MANOVA) among other deigns | 1.5 |
| | Completely Randomized Design (CRD) | Also known as single factor ANOVA | This is the simplest experimental design where the effect of one factor is evaluated. In this case the factor under consideration has three or more levels | 1 |
| | Randomized Complete Block Design (RCBD) | Also known as two-factor ANOVA | This is an extension of RCD, where another factor is added into the experiment. In this case, the effect of two factors are studies simultaneously, however, the effect of their interaction is not of interest in this case. | 1.5 |
| | Factorial Experimental Design | Introduction to Factorial Experimental design. | This is an extension of RCBD with the effect of two factors on the response variables studies simultaneously together with their interaction of effect. | |
| | Multivariate Analysis of ANOVA (MANOVA) | Introduction to multivariate analysis of variance (MANOVA) | This experiment puts into consideration a case of multiple dependent variables unlike in the other experiment where we had only one dependent variable. Besides, several assumptions are discussed in this section | 1.5 |

## R & Rstudio Training Outline

### Introduction

- ❖ Introduction to Rstudio

R, a powerful and open-source programming language, has become a cornerstone in the realm of data analysis and statistical computing. Developed by statisticians and data scientists, R offers a versatile environment for handling, exploring, and visualizing data. Its extensive collection of packages and libraries caters to a wide array of statistical and machine learning techniques, making it an invaluable tool for researchers, analysts, and postgraduate students alike. Whether you are cleaning and manipulating datasets, conducting exploratory data analysis, or building sophisticated predictive models, R provides a seamless and reproducible workflow. With its syntax designed for ease of use and a vibrant community that contributes to its growth, R serves as a fundamental skill for those seeking to unravel the insights hidden within the vast landscape of data. As we embark on this journey of learning, the versatility and efficiency of R will empower you to navigate the intricacies of data analysis and derive meaningful conclusions from complex datasets.

Although R is a programming language, it is unlike most others. It is designed to analyze data. It isn’t too difficult to learn, and is extremely popular. R has the advantage that it is free and open-source, and that thousands of users have contributed “add-on” packages that are readily downloadable by anyone. R is found in all areas of academia that encounter data, and in many private and public organizations. R is great for any job or task that uses data.

**Where to get R**

In this training, we will use R and RStudio. Both are free and open-source. Download and install R first:

https://cran.r-project.org/bin/windows/base/

(for Windows) or

https://cran.r-project.org/bin/macosx/

(for Mac).

Then, download and install RStudio from https://www.rstudio.com/products/rstudio/download/

| Topic | Description | Estimated Hours |
|---|---|---|
| Introduction to RStudio | Overview of RStudio environment and its key components | 1 |
| RStudio Windows | Explanation of Source Code, Console, Environment, and Plots windows | 1 |
| Work Directory | Understanding and setting the working directory in RStudio | 0.5 |
| Data Importation | Methods for importing data from various file formats such as Excel, CSV, SPSS, and Stata | 1.5 |
| Data Visualization | Techniques for visualizing data using Histograms, Scatter plots, Bubble plots, Box plots, Bar charts, Pie charts, and Line plots | 2 |
| Descriptive Statistics | Calculating and interpreting descriptive statistics such as Frequencies, Cross tabs, and Descriptive Statistics | 1.5 |

1. **Introduction to RStudio:** This session provides an overview of RStudio, explaining its interface and how it can be used for data analysis tasks. It's important to familiarize students with the layout and functionality of RStudio before delving into specific tasks.

2. **RStudio Windows:** Break down each window in RStudio (Source Code, Console, Environment, and Plots) and explain their respective purposes. This helps students understand where to write code, view results, manage objects, and visualize data.

3. **Work Directory:** Emphasize the importance of setting the working directory in RStudio for efficient file management and data import/export operations.

4. **Data Importation:** Cover different methods for importing data into RStudio from popular file formats like Excel, CSV, SPSS, and Stata. Also introduce the **choose. file** option for flexibility in selecting files.

5. **Data Visualization:** Discuss various visualization techniques available in RStudio, including Histograms, Scatter plots, Bubble plots, Box plots, Bar charts, Pie charts, and Line plots. Provide examples and demonstrate how to create each type of plot.

6. **Descriptive Statistics:** Teach students how to compute and interpret descriptive statistics using RStudio, including Frequencies, Cross tabs, and other summary statistics. This section should focus on basic statistical analysis to provide insights into the dataset.

| **Topic** | **Description** | **Estimated Hours** |
|---|---|---|
| Inferential Statistics | Introduction to inferential statistics and its importance in making inferences about populations based on sample data | 1.5 |
| Chi Square | Understanding and applying the chi-square test for categorical data analysis | 1 |
| T-test | Introduction to the t-test and its application in hypothesis testing | 1 |
| One-sample T-test | Testing the mean of a single sample against a known value | 0.5 |
| • Left tailed test | Explanation and application of left-tailed hypothesis testing | 0.5 |
| • Right tailed test | Explanation and application of right-tailed hypothesis testing | 0.5 |
| • Two-tailed test | Explanation and application of two-tailed hypothesis testing | 0.5 |
| Paired Sample T-test | Comparing means of two related groups | 0.5 |
| • Left tailed test | Application of left-tailed hypothesis testing for paired samples | 0.5 |
| • Right tailed test | Application of right-tailed hypothesis testing for paired samples | 0.5 |
| • Two-tailed test | Application of two-tailed hypothesis testing for paired samples | 0.5 |
| Unpaired Sample T-test | Comparing means of two independent groups | 0.5 |
| • Left tailed test | Application of left-tailed hypothesis testing for unpaired samples | 0.5 |
| • Right tailed test | Application of right-tailed hypothesis testing for unpaired samples | 0.5 |
| • Two-tailed test | Application of two-tailed hypothesis testing for unpaired samples | 0.5 |

1. **Inferential Statistics:** This section introduces the concept of inferential statistics, which allows us to draw conclusions about populations based on sample data. It sets the foundation for hypothesis testing and making predictions.

2. **Chi Square:** Focuses on the chi-square test, which is used for analyzing categorical data to determine if there is a significant association between two categorical variables.

3. **T-test:** Introduces the t-test, a statistical test used to determine if there is a significant difference between the means of two groups. It includes one-sample, paired sample, and unpaired sample t-tests.

4. **One-sample T-test:** Tests whether the mean of a single sample differs from a known value. It's useful for comparing sample data to a population parameter.

5. **Paired Sample T-test:** Compares the means of two related groups, such as before and after measurements or matched pairs.

6. **Unpaired Sample T-test:** Compares the means of two independent groups to determine if there is a significant difference between them. It's commonly used in experimental and observational studies.

| Topic | Description | Estimated Hours |
|---|---|---|
| Linear Regression and Correlation Analysis | Introduction to linear regression and correlation analysis, exploring the relationship between variables and predicting outcomes based on this relationship | 2 |
| Simple Linear Regression Analysis | Explanation of simple linear regression, where one independent variable is used to predict the dependent variable | 1 |
| - Model Construction | Building a linear regression model using a single predictor variable | 0.5 |
| - Interpretation of Coefficients | Understanding the meaning of coefficients in the regression equation and their significance | 0.5 |
| Multiple Linear Regression Analysis | Extension of simple linear regression to multiple predictor variables | 1 |
| - Model Construction | Building a linear regression model using multiple predictor variables | 0.5 |
| - Interpretation of Coefficients | Understanding the interpretation of coefficients in the multiple regression equation | 0.5 |
| Model Diagnostic | Assessment of the assumptions and diagnostics of linear regression models | 1 |
| - Assumption Checking | Checking for assumptions such as linearity, independence, homoscedasticity, and normality | 0.5 |
| - Residual Analysis | Analyzing residuals to assess the goodness-of-fit of the regression model | 0.5 |
| - Influence and Outlier Detection | Identifying influential points and outliers that may impact the regression model | 0.5 |

1. **Linear Regression and Correlation Analysis:** This section introduces the concepts of linear regression and correlation analysis, which are fundamental techniques for exploring relationships between variables and making predictions based on these relationships. It covers both the theoretical background and practical applications of these methods.

2. **Simple Linear Regression Analysis:** Focuses on simple linear regression, where one independent variable is used to predict a dependent variable. It includes constructing the regression model and interpreting the coefficients to understand the relationship between variables.

3. **Multiple Linear Regression Analysis:** Extends simple linear regression to multiple predictor variables, allowing for more complex models that account for multiple factors influencing the dependent variable. It covers model construction and interpretation of coefficients in the context of multiple predictors.

4. **Model Diagnostic**: Covers the diagnostic procedures for assessing the validity and reliability of linear regression models. This includes checking assumptions, analyzing residuals, and identifying influential points and outliers that may affect the model's performance as well as econometric problems such as heteroscedasticity, multicollinearity, autocorrelation among others

| Topic | Description | Estimated Hours |
|---|---|---|
| Logistic Regression Analysis | Introduction to logistic regression, a statistical method used for modeling the probability of a binary or categorical outcome variable | 2 |
| Binary Logistic Regression | Analysis of binary outcomes using logistic regression | 1 |
| - Simple binary logistic regression | Explanation and application of logistic regression with one predictor variable | 0.5 |
| - Multiple binary logistic regression analysis | Extending logistic regression to include multiple predictor variables | 0.5 |
| Multinomial Logistic Regression Model | Extension of logistic regression to handle categorical outcome variables with more than two categories | 1 |

1. **Logistic Regression Analysis**: This section introduces logistic regression as a powerful statistical technique for modeling the probability of a binary or categorical outcome variable. It covers both binary and multinomial logistic regression models.
2. **Binary Logistic Regression:** Focuses on analyzing binary outcomes, where the dependent variable has only two possible outcomes. It includes simple binary logistic regression with one predictor variable and multiple binary logistic regression with multiple predictors.
3. **Simple Binary Logistic Regression:** Explains the concept and application of logistic regression when there is only one predictor variable. This model is used to predict the probability of the occurrence of one of the two possible outcomes.
4. **Multiple Binary Logistic Regression Analysis:** Extends logistic regression to cases where there are multiple predictor variables. It allows for more complex models that account for multiple factors influencing the binary outcome variable.
5. **Multinomial Logistic Regression Model:** Expands logistic regression to handle categorical outcome variables with more than two categories. It's used when the dependent variable has three or more unordered categories, and it models the probabilities of each category relative to a reference category.

| **Topic** | **Description** | **Estimated Hours** |
|---|---|---|
| Pearson Correlation Coefficient | Introduction to the Pearson correlation coefficient, a measure of the linear relationship between two continuous variables | 1 |

Pearson Correlation Coefficient:

This section introduces the Pearson correlation coefficient, also known as Pearson's r, which quantifies the strength and direction of the linear relationship between two continuous variables. It covers the calculation of the correlation coefficient, interpretation of its magnitude, and significance testing. Additionally, this topic will include discussions on the assumptions of Pearson correlation, such as linearity and homoscedasticity.

This topic can typically be covered in about 1 hour, including explanations, examples, and practical exercises for students to understand and apply the concept effectively. Adjustments may be necessary based on the pace of the class and the level of prior statistical knowledge.

**DESIGN AND ANALYSIS OF EXPERIMENT**

| **Topic** | **Description** | **Estimated Hours** |
|---|---|---|
| Complete Randomized Design (CRD); One-way ANOVA | Introduces the complete randomized design (CRD) as a basic experimental design where treatments are randomly assigned to experimental units. It also covers the one-way ANOVA for analyzing CRD data. | 2 |
| Randomized Complete Block Design (RCBD); Two-way ANOVA | Explores the randomized complete block design (RCBD), which accounts for variability among blocks or groups of experimental units. It includes the two-way ANOVA for analyzing RCBD data with two factors. | 2.5 |
| Balanced Incomplete Block Design (BIBD) | Discusses the balanced incomplete block design (BIBD), a design used when it is impractical to include all possible treatment combinations in every block. It covers the construction and analysis of BIBD. | 2 |

| | | |
|---|---|---|
| Factorial Experimental Design (FED) | Introduces the factorial experimental design (FED), where multiple factors are manipulated simultaneously to study their individual and interactive effects on the response variable. It includes factorial ANOVA for analyzing FED data. | 2.5 |

**NON-PARAMETRIC TESTS**

| **Topic** | **Description** | **Estimated Hours** |
|---|---|---|
| Wilcoxon | Introduction to the Wilcoxon signed-rank test, a non-parametric test used to compare paired samples when the assumption of normality is violated. It evaluates whether the medians of two paired groups are different. | 1.5 |
| Mann-Whitney U-test | Explanation of the Mann-Whitney U-test, also known as the Wilcoxon rank-sum test, a non-parametric test used to compare independent samples. It assesses whether the distributions of two groups are the same based on their ranks. | 1.5 |
| Rank Sum Test/H-Test/Kruskal Wallis Test | Discusses the Kruskal-Wallis test, a non-parametric alternative to one-way ANOVA, used to determine whether there are statistically significant differences among two or more independent groups. It compares the medians of multiple groups simultaneously. | 2 |

## References

1. *Machine Learning with R, the tidyverse, and mlr by* HEFIN I. RHYS
2. *R For Data Science: Import, Tidy, Transform, Visualize, And Model Data, by* Hadley Wickham & Garrett Grolemund
3. Data Analysis with Stata: *A Comprehensive Guide for Data Analysis and Interpretation of Outputs* First Edition Mohammad Tajul Islam, Russell Kabir, Monjura Nisha
4. Islam M.T, Kabir. R and Nisha. M. (2022). Data analysis with STATA, first edition by ASa Publishers, Dhaka, Bangladesh
5. SPSS Manual
6. R and Rstudio Manual
7. Stata Manual
8. Python Manual