LECTURE 01

21CSE295T

Statistical Data Analytics And Integration

Unit 1 Introduction to Statistical Analysis

1.1	Mathematical Statistics
1.2	Statistical Pattern Recognition
1.3	Nonparametric Regression
1.4	Exploratory Data Analysis
1.5	Monte Carlo Methods for Inferential Statistics
1.6	Basic Concepts of Statistics
1.7	Types of Data (Categorical, Numerical)
1.8	Populations and Samples
1.9	Measures of Central Tendency (Mean, Median, Mode)
1.1O	Measures of Variability (Range, Variance, Standard Deviation)
1.11	Descriptive vs Inferential Statistics

Virupakshan K

+91-9840847872 | VIRU@KART.LA

SYLLABUS
21CSE295T Statistical Data Analytics And Integration

Unit 1 Introduction to Statistical Analysis **9 hours**

Computational Statistics- Statistical Pattern Recognition, Nonparametric Regression, Exploratory Data Analysis, Monte Carlo Methods for Inferential Statistics-Basic concepts of statistics, Types of data (categorical, numerical), Populations and samples, Measures of central tendency (mean, median, mode), Measures of variability (range, variance, standard deviation), Descriptive vs. inferential statistics.

Tutorial:

- SAS Statements, Structure, Libraries, Rules, Datasets, Representation of missing values, Assigning Libraries, PROC CONTENTS, visualize the descriptor and datasets attributes

Unit 2 Data Analysis Using SAS **9 hours**

Basic Reports – Selecting Variables – Identifying observations – Subsetting the Data – Limiting the observations-SORT Procedure – Generating Column totals – Specifying titles and footnotes in procedure output – Assigning Descriptive labels and permanent label -How SAS Processes Programs – Compilation Phase – PDV – Descriptor Portion-Execution Phase – Data Portion – Debugging a Data Step – Testing Programs-BY-Group Processing Definitions – Preprocessing Data – Determine Whether the Data Requires Preprocessing – Sorting Observations for BY-Group Processing-FIRST. and LAST. DATA Step Variables – How the DATA Step Identifies BY Groups – Assignment Statements – Operators in SAS Expressions

Tutorial:

- Introduction of report and PRINT Procedure-Hands on practice of VAR, ID, WHERE statement, where expression[eq, ne, gt, lt, ge, le, AND, OR, CONTAINS]

Unit 3 Advanced Statistical Techniques in SAS **9 hours**

Determining the Structure and Contents of Data Sets – Testing Program – Methods of Combining SAS Data Sets-One-to-One Reading – Concatenating – Match-Merging – Renaming Variables – Excluding Unmatched Observations -Selecting Matching Observations -Basics of DO Loops – Executing DO Loops -OUTPUT Statements – Nesting DO Loops-Iteratively Processing Observations from a Data Set -DO UNTIL Statement – DO WHILE Statement Defining and Processing Arrays –Arrays -Applying SAS Formats and Informats – Specifying SAS Formats

Tutorial:

- One-to-One Reading to Combine Data Sets-Concatenating to Combine Data Sets-Match-Merging to Combine Data Sets-Renaming Variables-Processing Iterative DO Loops-Indenting and Nesting DO Groups

Unit 4 SAS Programming and Automation **9 hours**

Definitions – Reading Date and Time with Informats – Displaying Date and Time Values with Formats-Basics of SAS Functions - SAS Functions Syntax – Converting Data with Functions- The MEANS Procedure – Specifying Descriptive Statistics – Limiting Decimal Places – Specifying Variable Using VAR Statement – Summarized Data Set Using OUTPUT Statement – The FREQ Procedure-Specifying Variable Using TABLES Statements - Two-Way and N-Way Tables – LIST Options-The Output Delivery System (ODS) – HTML Outputs – PDF Outputs-RTF Outputs – EXCEL Outputs – The EXPORT Procedure

Tutorial:

- Reading Date and Time with Formats and Informats- Execution of Data Values Using Numeric Functions-Creating a Descriptive Statistics by Using PROC MEANS-Creating a Descriptive Statistics by Using PROC FREQ-Creating Outputs with HTML, PDF, RTF, EXCEL Formats-Exporting a External File

Unit 5 Advanced SAS Topics **9 hours**

Creating a Default List Report – Windowing/Non Windowing Mode - Define Statement-Defining Variables – Column Statement – Selecting Observations – Defining the Variables-Group Processing the Variables by Using Different Options – Computing the Variable-Bar Chart – Box Plot – Mean Measurement Over the Time using Line Plot – Spaghetti Plot – Survival Plot – Waterfall Plot – Forest Plot-Introduction to Macro Facility – SAS Programs and Macro Processing – Macro Variables – Macro Processing – Scopes of Macro Variables – Macro Expressions-Scopes of Macro Variables – Macro Expressions-Macro Quoting – Interfaces with Macro Facility -Storing and Reusing Macros.

Tutorial:

- Creating Graphs by Using Bar Chart -Creating Graphs Using Various Types of Plots-Generating SAS Code Using Macros-Defining User-Defined Macro Variables-Creating a Tables Using PROC SQL with different Clauses- Generating a Cartesian Product -Combining Tables Using Multiple Joins

LECTURE 01 TOPICS

Unit 1 Introduction to Statistical Analysis — **9 hours**

Computational Statistics- Statistical Pattern Recognition, Nonparametric Regression, Exploratory Data Analysis, Monte Carlo Methods for Inferential Statistics-Basic concepts of statistics, Types of data (categorical, numerical), Populations and samples, Measures of central tendency (mean, median, mode), Measures of variability (range, variance, standard deviation), Descriptive vs. inferential statistics.

Tutorial:

- SAS Statements, Structure, Libraries, Rules, Datasets, Representation of missing values, Assigning Libraries, PROC CONTENTS, visualize the descriptor and datasets attributes

Unit 2 Data Analysis Using SAS — **9 hours**

Basic Reports – Selecting Variables – Identifying observations – Subsetting the Data – Limiting the observations-SORT Procedure – Generating Column totals – Specifying titles and footnotes in procedure output – Assigning Descriptive labels and permanent label -How SAS Processes Programs – Compilation Phase – PDV – Descriptor Portion-Execution Phase – Data Portion – Debugging a Data Step – Testing Programs-BY-Group Processing Definitions – Preprocessing Data – Determine Whether the Data Requires Preprocessing – Sorting Observations for BY-Group Processing-FIRST. and LAST. DATA Step Variables – How the DATA Step Identifies BY Groups – Assignment Statements – Operators in SAS Expressions

Tutorial:

- Introduction of report and PRINT Procedure-Hands on practice of VAR, ID, WHERE statement, where expression[eq, ne, gt, lt, ge, le, AND, OR, CONTAINS]

Unit 3 Advanced Statistical Techniques in SAS — **9 hours**

Determining the Structure and Contents of Data Sets – Testing Program – Methods of Combining SAS Data Sets-One-to-One Reading – Concatenating – Match-Merging – Renaming Variables – Excluding Unmatched Observations -Selecting Matching Observations -Basics of DO Loops – Executing DO Loops -OUTPUT Statements – Nesting DO Loops-Iteratively Processing Observations from a Data Set -DO UNTIL Statement – DO WHILE Statement Defining and Processing Arrays –Arrays -Applying SAS Formats and Informats – Specifying SAS Formats

Tutorial:

- One-to-One Reading to Combine Data Sets-Concatenating to Combine Data Sets-Match-Merging to Combine Data Sets-Renaming Variables-Processing Iterative DO Loops-Indenting and Nesting DO Groups

Unit 4 SAS Programming and Automation — **9 hours**

Definitions – Reading Date and Time with Informats – Displaying Date and Time Values with Formats-Basics of SAS Functions - SAS Functions Syntax – Converting Data with Functions- The MEANS Procedure – Specifying Descriptive Statistics – Limiting Decimal Places – Specifying Variable Using VAR Statement – Summarized Data Set Using OUTPUT Statement – The FREQ Procedure-Specifying Variable Using TABLES Statements - Two-Way and N-Way Tables – LIST Options-The Output Delivery System (ODS) – HTML Outputs – PDF Outputs-RTF Outputs – EXCEL Outputs – The EXPORT Procedure

Tutorial:

- Reading Date and Time with Formats and Informats- Execution of Data Values Using Numeric Functions-Creating a Descriptive Statistics by Using PROC MEANS-Creating a Descriptive Statistics by Using PROC FREQ-Creating Outputs with HTML, PDF, RTF, EXCEL Formats-Exporting a External File

Unit 5 Advanced SAS Topics — **9 hours**

Creating a Default List Report – Windowing/Non Windowing Mode - Define Statement-Defining Variables – Column Statement – Selecting Observations – Defining the Variables-Group Processing the Variables by Using Different Options – Computing the Variable-Bar Chart – Box Plot – Mean Measurement Over the Time using Line Plot – Spaghetti Plot – Survival Plot – Waterfall Plot – Forest Plot-Introduction to Macro Facility – SAS Programs and Macro Processing – Macro Variables – Macro Processing – Scopes of Macro Variables – Macro Expressions-Scopes of Macro Variables – Macro Expressions-Macro Quoting – Interfaces with Macro Facility -Storing and Reusing Macros.

Tutorial:

- Creating Graphs by Using Bar Chart -Creating Graphs Using Various Types of Plots-Generating SAS Code Using Macros-Defining User-Defined Macro Variables-Creating a Tables Using PROC SQL with different Clauses- Generating a Cartesian Product -Combining Tables Using Multiple Joins

Unit 1: Overview

Statistical analysis is the process of collecting, organizing, summarizing, analyzing, and interpreting data to make decisions.

In data analytics, statistics provides the mathematical foundation for understanding patterns, uncertainty, relationships, and predictions.

1.0 Introduction

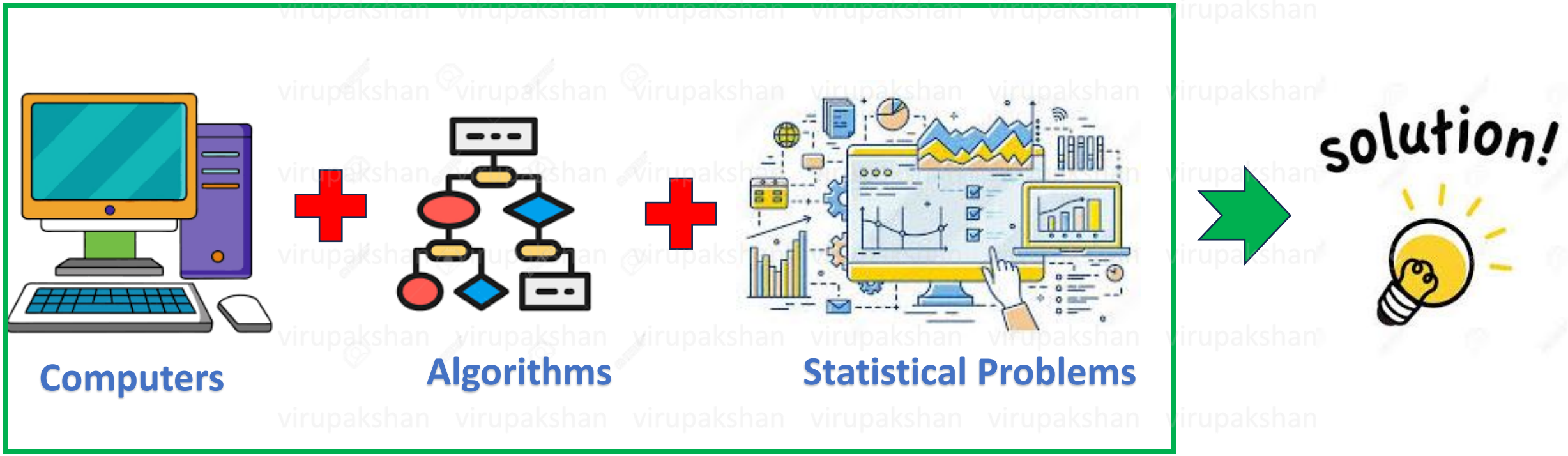
This unit introduces:

- Basic statistical concepts
- Types of data
- Population and sample concepts
- Descriptive and inferential statistics
- Measures of central tendency and variability
- Exploratory Data Analysis, EDA
- Statistical pattern recognition
- Nonparametric regression
- Monte Carlo methods
- Basic SAS concepts including libraries, datasets, missing values, PROC PRINT, and PROC CONTENTS

1.1.0 Computational Statistics: **DEFINITION**

Q) What is Computational Statistics ?

A) **Computational Statistics** is a branch of statistics that uses computers and algorithms to solve statistical problems.



1.1.1 Computational Statistics: APPLICATIONS

IT COMBINES

- Statistics
- Computer Science
- Mathematics
- Data Analysis

Computational statistics is very useful when calculations become too large or complex for manual methods.

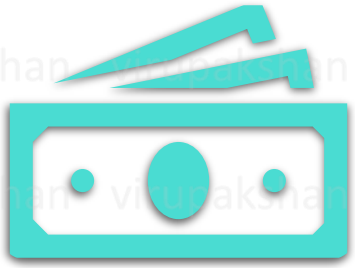
APPLICATIONS

Machine Learning

Data Mining

Artificial Intelligence

Medical Diagnosis



Financial Forecasting

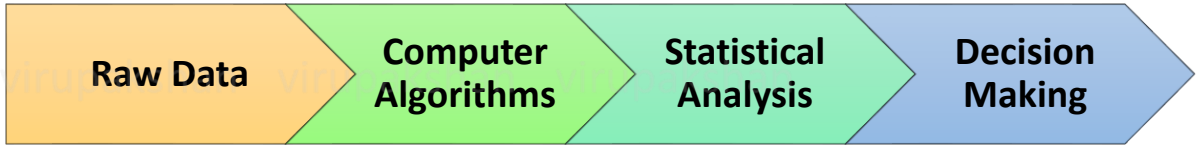
Weather Prediction

1.1.2 Computational Statistics: TECHNIQUES & STEPS

IMPORTANT TECHNIQUES

- Simulation
- Bootstrap Methods
- Monte Carlo Methods
- Numerical Optimization
- Statistical Modeling

STEPS



1.1.3 Computational Statistics: Q&A

Q1. What is Computational Statistics?

Answer: Computational Statistics is the use of computer-based methods and algorithms to solve statistical problems.

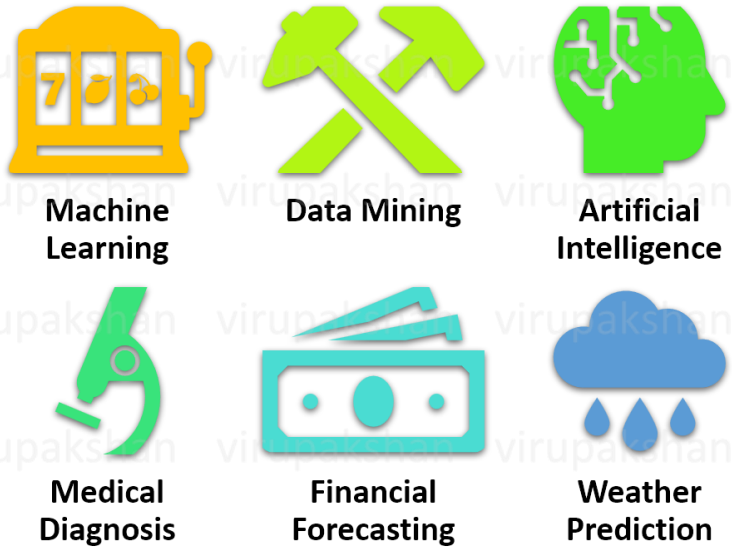
- Uses computers for statistical analysis
- Handles large datasets
- Uses simulations and algorithms
- Important in AI and Machine Learning

Q2. Name two applications of Computational Statistics.

Answer:

1.Machine Learning

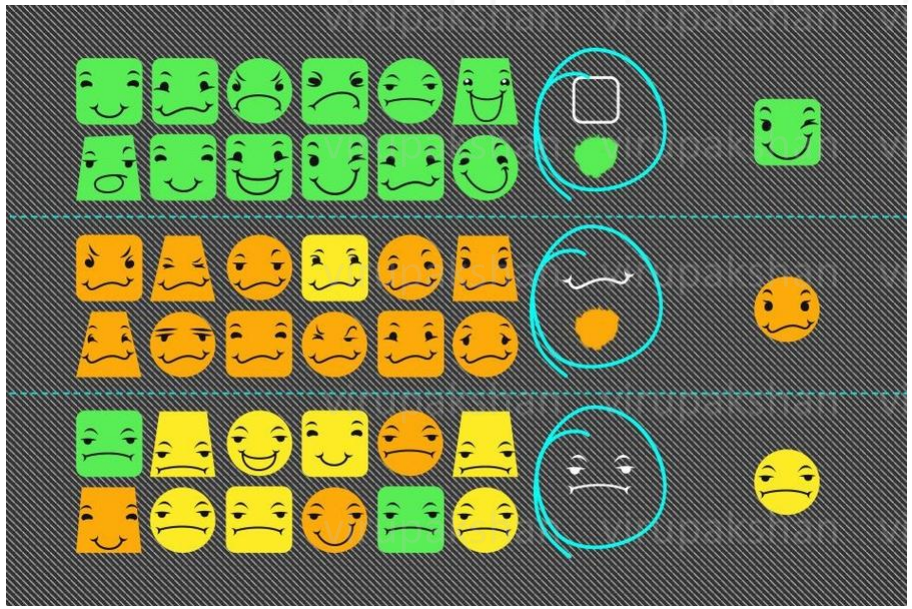
2.Financial Forecasting



1.2.0 Pattern Recognition – BASIC TERM

Q) What is Pattern Recognition?

A) Pattern recognition is the **cognitive process or automated method** of identifying and interpreting recurring structures, regularities, or relationships in data.



1.2.1 Statistical Pattern Recognition: DEFINITION & USES

Q) Define Statistical Pattern Recognition

A) Statistical Pattern Recognition is the process of **identifying patterns** in **data** using statistical methods.

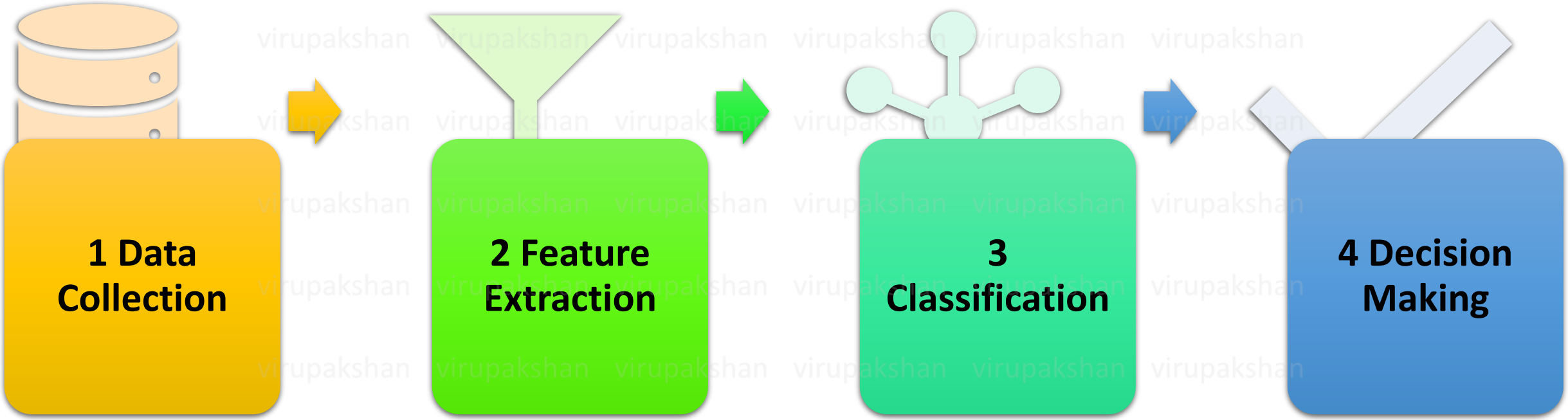
Q) How does it help?

A) It helps **computers** recognize:

- Images
- Speech
- Text
- Human behaviour

1.2.2 Statistical Pattern Recognition: **STEPS**

Q) What are the steps in Pattern Recognition?



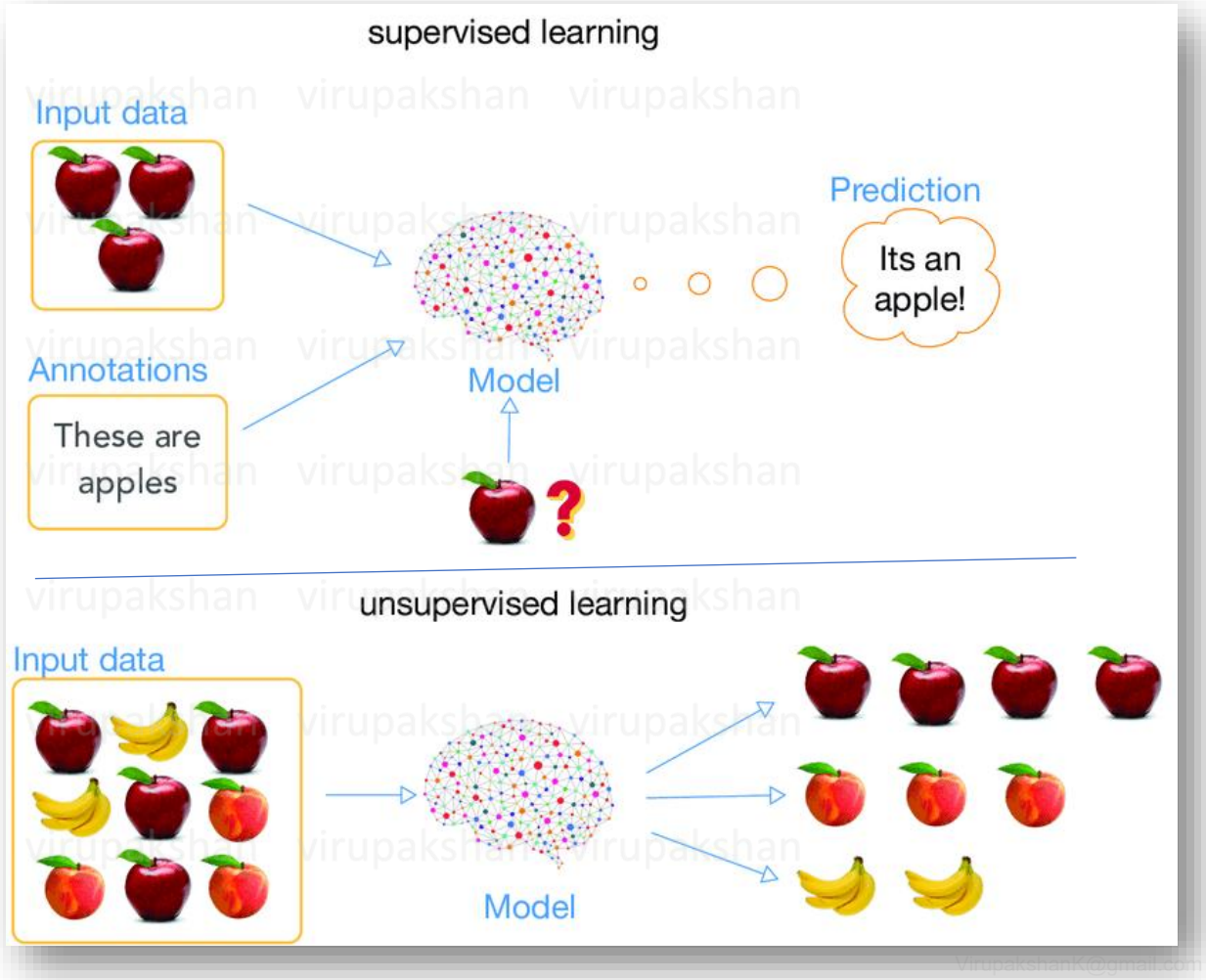
1.2.3 Statistical Pattern Recognition: **TYPES**

Q) What are the types in Pattern Recognition?

A) Two types are:

1. Supervised Learning
Uses labelled data.

2. Unsupervised Learning
Uses unlabelled data.



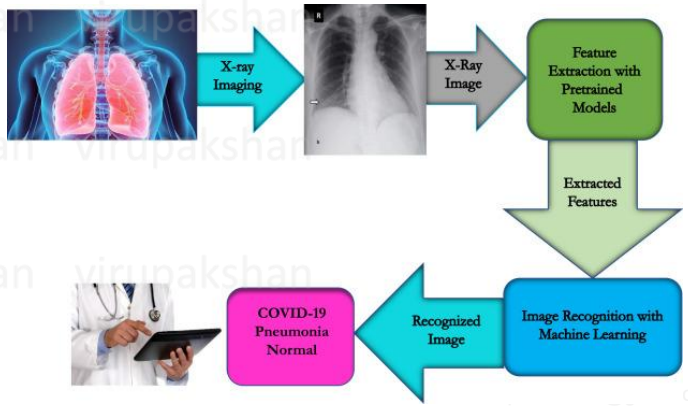
1.2.4 Statistical Pattern Recognition: APPLICATIONS

Q) What are the Applications of Statistical Pattern Recognition?

A) Few applications are:

- Face recognition
- Fingerprint identification
- Market segmentation
- Credit card fraud detection
- Handwriting recognition
- Disease diagnosis

Credit Card Fraud Detection



1.3.0 Nonparametric Regression: What is "regression"?

In statistics and machine learning, **regression** is a method used to estimate the relationship between:

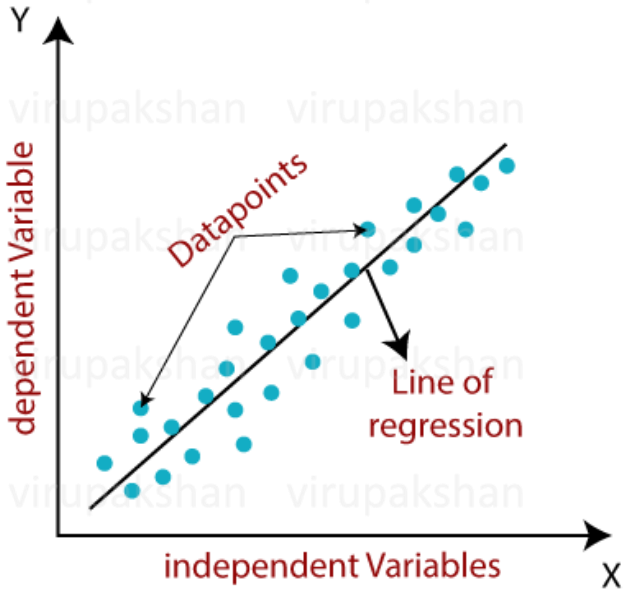
- **a dependent (target) variable** and
- **one or more independent (predictor) variables**

In ordinary **linear regression**, we assume a relationship such as:

$$Y = a + bX$$

Dependent Variable

Independent Variable



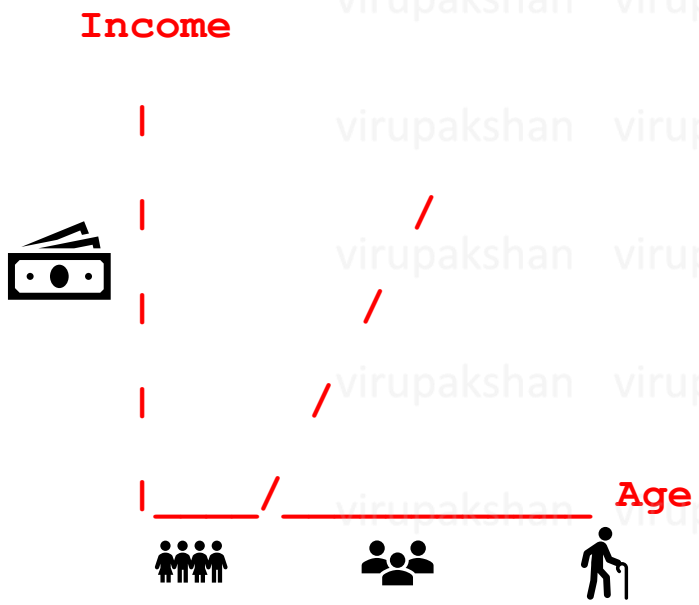
1.3.1 Nonparametric Regression: PARAMETRIC vs NON-PARAMETRIC

Feature	Parametric Regression	Nonparametric Regression
Assumed form	Fixed equation	No fixed form
Example	Linear regression	Kernel / KNN regression
Flexibility	Lower	Higher
Interpretability	Usually easier	Sometimes harder
Data need	Lower	Higher

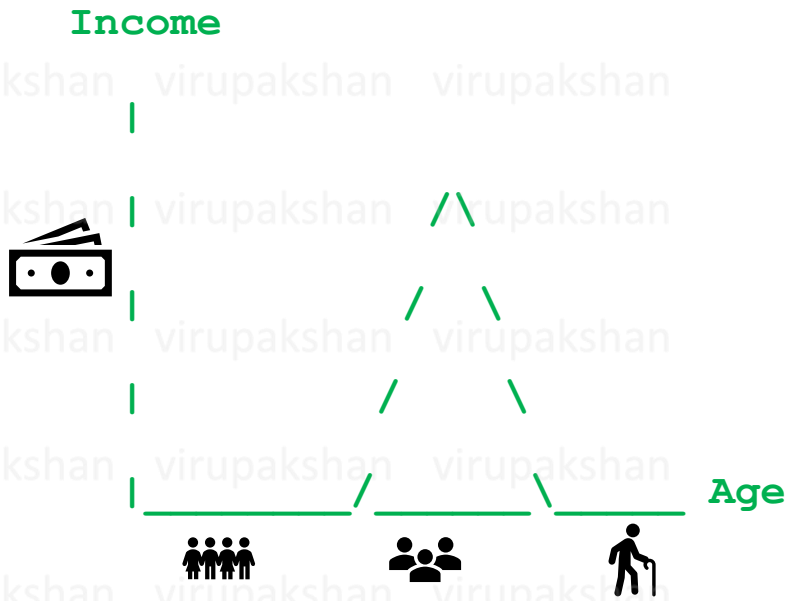
1.3.2 Nonparametric Regression: PARAMETRIC vs NON-PARAMETRIC

Suppose we study the relationship between age and income.

A **parametric model** may assume **income** increases linearly with **age**:



But the real relationship may be curved.



Nonparametric regression can capture this pattern better.

1.3.3 Nonparametric Regression: COMMON METHODS

Common Nonparametric Regression Methods

1. Kernel Regression

Uses nearby data points to estimate the value of a function.

2. Local Polynomial Regression

Fits small polynomial models in local regions of the data.

3. Smoothing Splines

Uses smooth curves to fit the data.

4. k-Nearest Neighbor Regression

Predicts a value based on the average of nearby observations.

1.3.4 Nonparametric Regression : k-NN Regression Example

Example: k-Nearest Neighbor Regression

Data: study hours and marks: (2,40), (3,50), (4,60), (5,70).

To predict marks for 3.5 hours, use nearby observations 3 and 4 hours.

Predicted marks = (50 + 60) / 2 = 55.

This is a local, nonparametric prediction method.

1.3.5 Nonparametric Regression: **Adv and Disadv**

Advantages

- Flexible
- Useful for complex patterns
- Does not require strong assumptions

Disadvantages

- Requires large datasets
- Can be computationally expensive
- May overfit if not controlled

1.4.1 Exploratory Data Analysis (EDA): MEANING AND GOALS

Exploratory Data Analysis, or **EDA**, is the process of examining data to understand its structure, patterns, anomalies, and relationships before applying formal statistical models.

EDA is often the first step in data analytics.

Goals of EDA

- Understand the data
- Detect missing values
- Identify outliers
- Find patterns and relationships
- Check data quality
- Select suitable statistical methods

1.4.2 Exploratory Data Analysis (EDA): EDA TECHNIQUES

1. Summary Statistics

Examples:

- Mean
- Median
- Mode
- Standard deviation
- Minimum
- Maximum
- Quartiles

2. Data Visualization

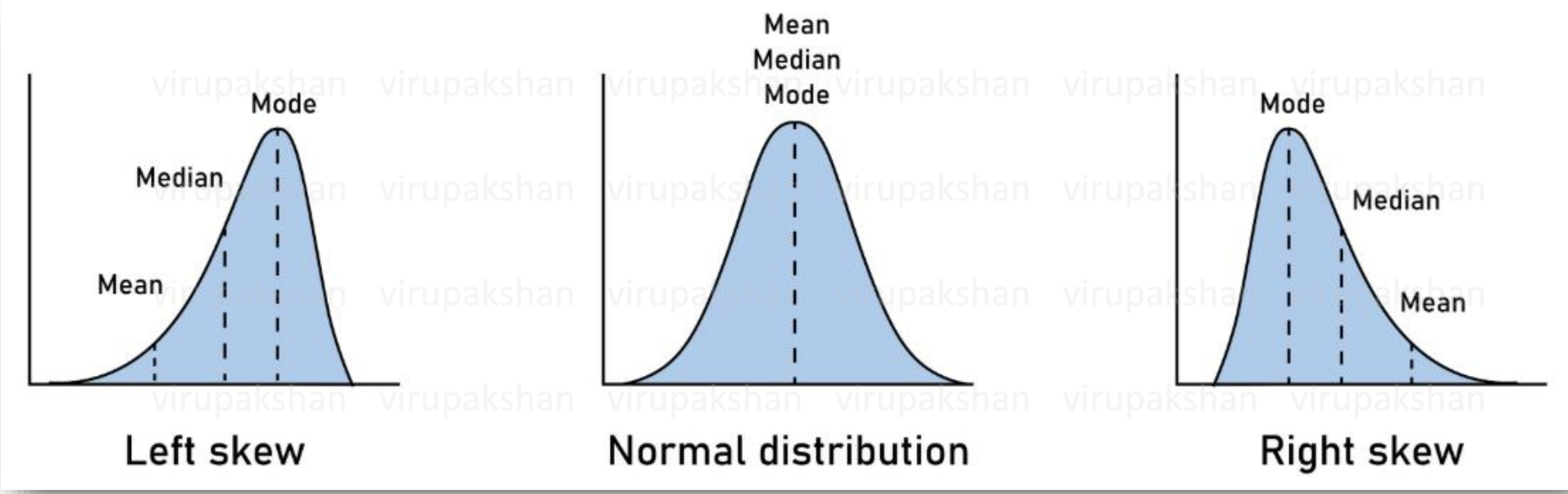
Common charts:

Chart	Purpose
Histogram	Shows distribution
Box plot	Detects outliers
Scatter plot	Shows relationship between variables
Bar chart	Compares categories
Line chart	Shows trends over time

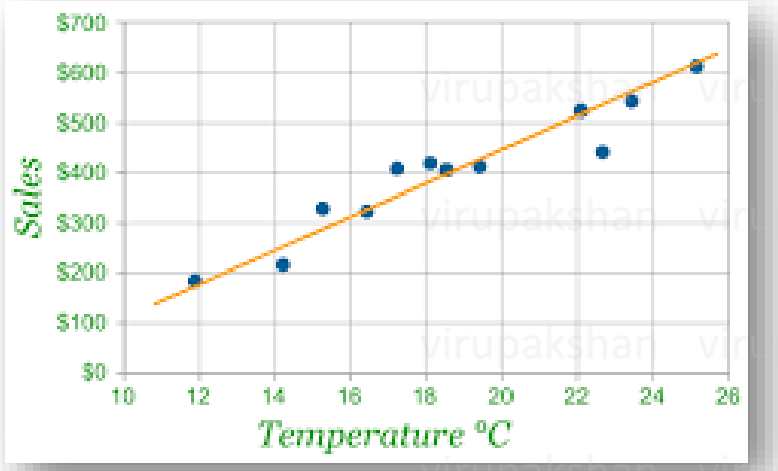
1.4.3 Exploratory Data Analysis (EDA): **1 SUMMARY STATISTICS**

Measures of Central Tendency (The Middle)These numbers show you exactly where the "center of gravity" of your dataset lies:

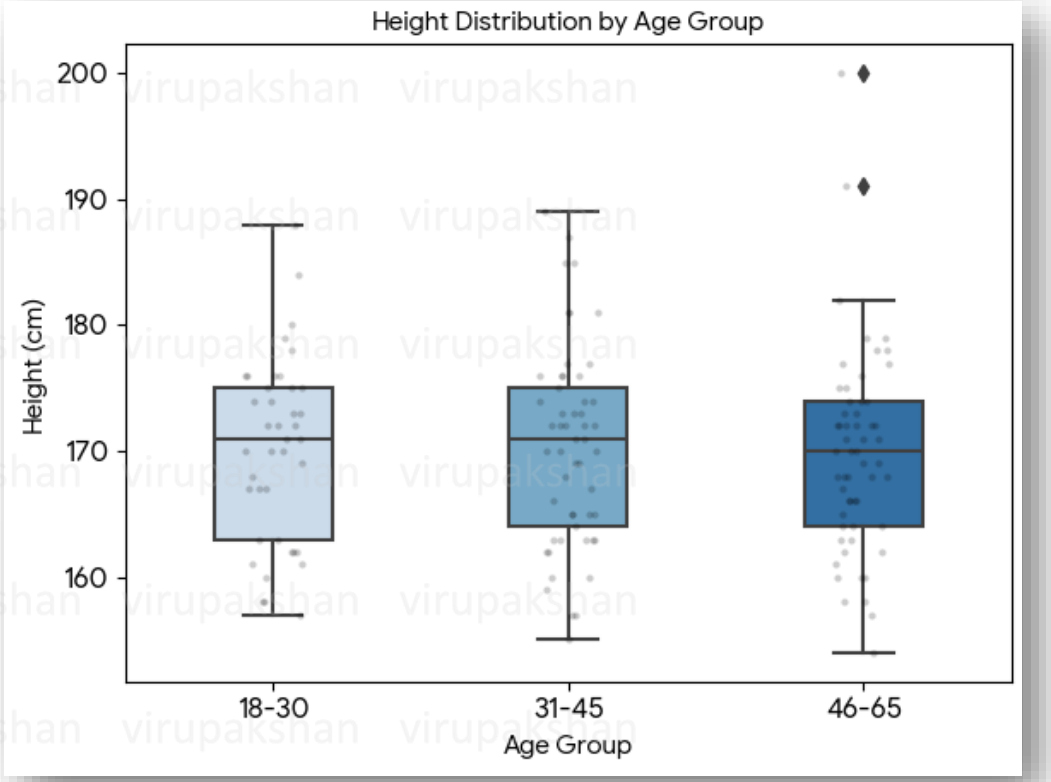
1. **Mean**: The standard mathematical average. You add all values together and divide by the total number of points.
2. **Median**: The absolute middle value when the data is sorted from smallest to largest. It splits your data perfectly into two halves.
3. **Mode**: The most frequently occurring value in your dataset.



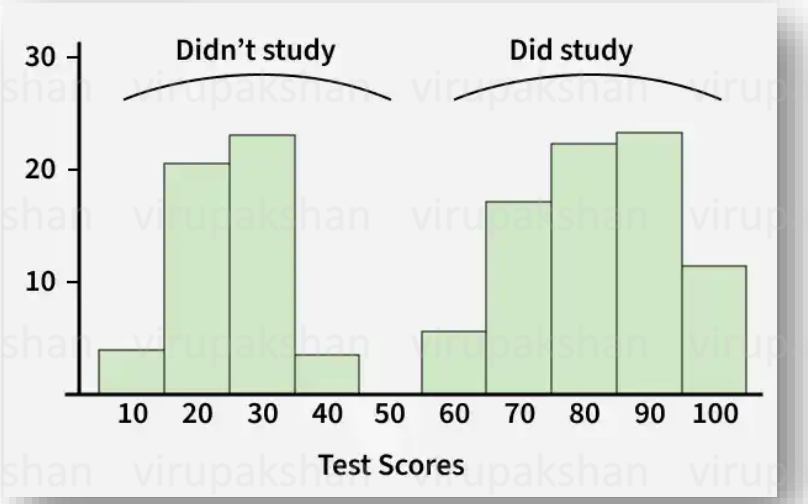
1.4.4 Exploratory Data Analysis (EDA): **2 DATA VISUALIZATION**



Scatter Plots



Boxplots



Histograms

1.4.5 Exploratory Data Analysis (EDA): **EDA WORKFLOW**

Collect Data → Clean Data → Summarize Data → Visualize Data → Identify Patterns → Prepare for Modeling

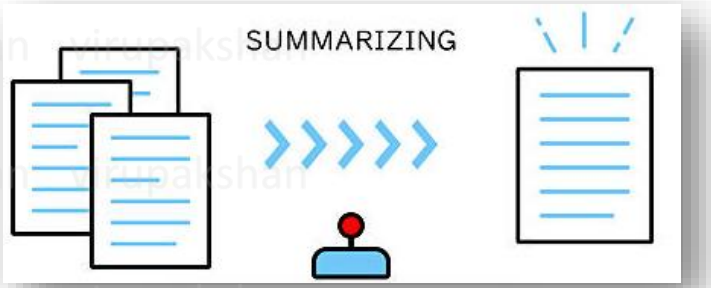
Collect & Clean Data

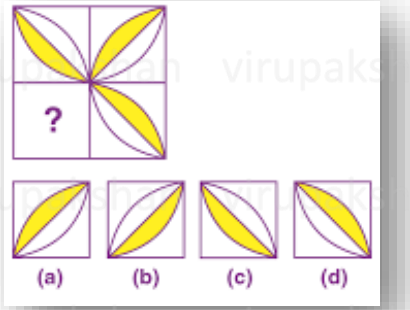
Summarize Data

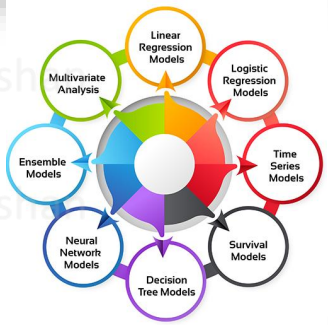
Visualize Data

Identify Patterns

Prepare for Modeling

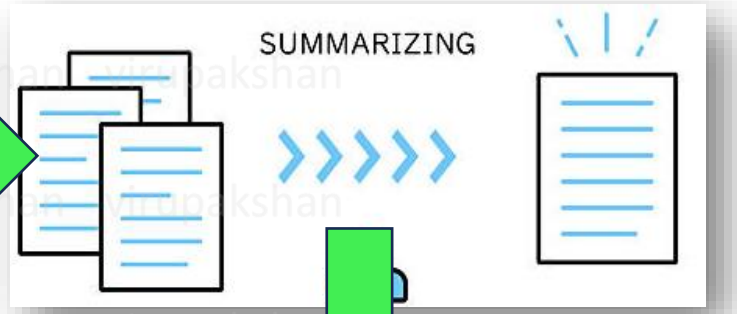




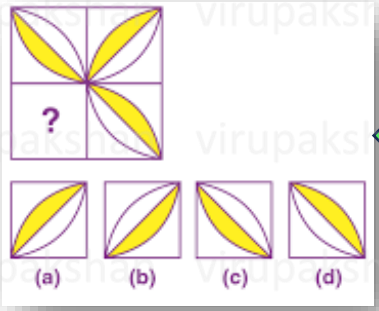


1.4.5 Exploratory Data Analysis (EDA): **EDA WORKFLOW**

Collect Data → Clean Data → Summarize Data → Visualize Data → Identify Patterns → Prepare for Modeling

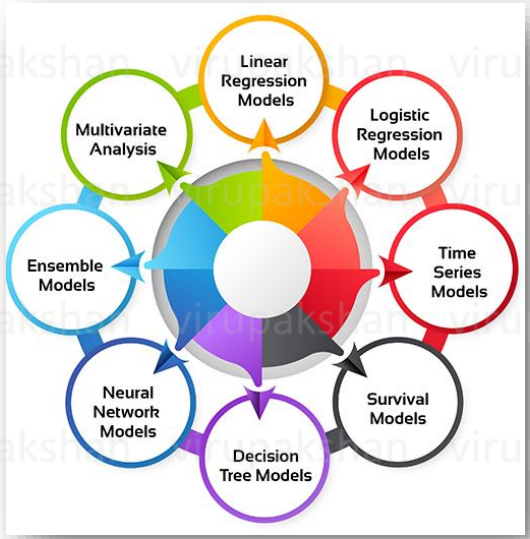


Visualize Data



Identify Patterns

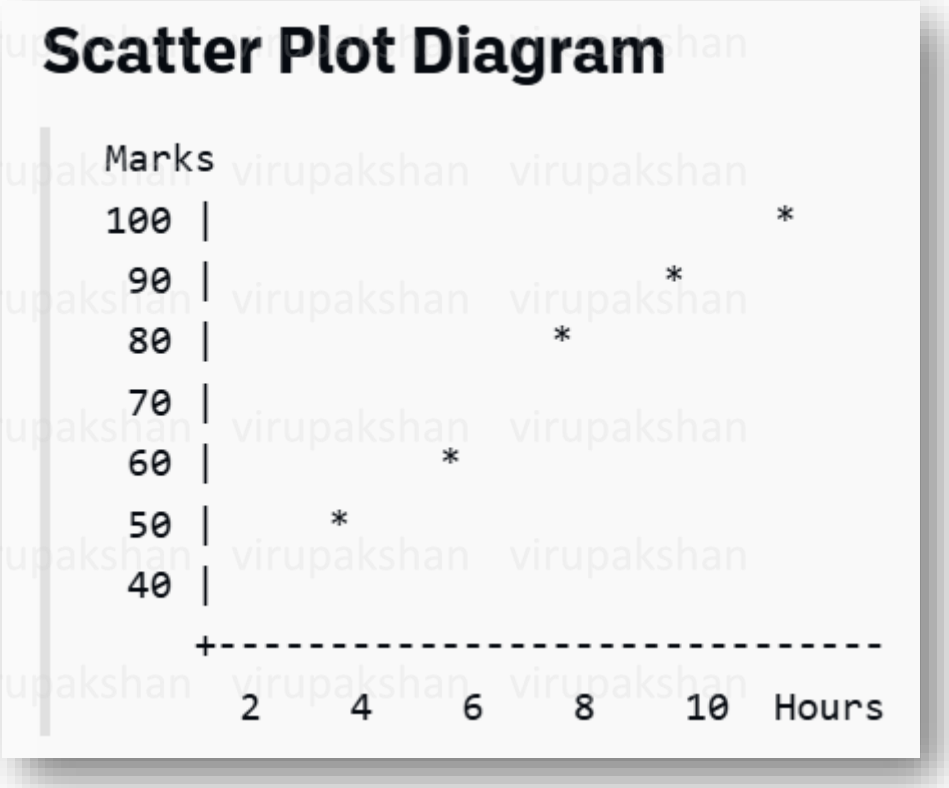
Prepare for Modeling



1.4.6 Exploratory Data Analysis (EDA): **Scatter Plot Example**

EXAMPLE:

Student	Hours Studied	Marks
A	2	45
B	4	60
C	6	75
D	8	85
E	10	95



This shows a **positive** relationship between study hours and marks.

1.4.7 Exploratory Data Analysis (EDA): **Q&A**

Q1. What is EDA?

Answer: EDA is the process of exploring and understanding data using graphs and statistical methods.

Q2. Name two EDA tools.

Answer:

1.Histogram

2.Boxplot

1.5.1 Monte Carlo Methods for Inferential Statistics

MEANING: Monte Carlo methods use repeated random sampling to estimate unknown quantities, probabilities, or statistical properties.

They are useful when exact mathematical solutions are difficult.

BASIC IDEA: Instead of solving a problem directly, we simulate it many times and observe the results.

1.5.2 Monte Carlo Methods for Inferential Statistics

Monte Carlo Process

```
Define Problem
    |
    v
Generate Random Samples
    |
    v
Run Simulation Many Times
    |
    v
Summarize Results
    |
    v
Make Statistical Inference
```

OR in simple terms…

1. Generate random samples
2. Perform calculations repeatedly
3. Estimate final result

Applications

- Risk Analysis
- Simulation
- Probability Estimation
- Financial Modeling

1.5.3 Monte Carlo Methods for Inferential Statistics

Use in Inferential Statistics

Monte Carlo methods are used for:

- Estimating sampling distributions
- Calculating p-values
- Confidence intervals
- Bayesian statistics
- Bootstrapping
- Risk analysis

1.5.4 Monte Carlo Methods for Inferential Statistics

Example: Estimating Probability

Suppose we want to estimate the probability of getting heads when tossing a coin.

Instead of calculating theoretically, we simulate 1,000 coin tosses.

If heads appears 510 times:

$$P(H) \approx \frac{510}{1000} = 0.51$$

The theoretical value is 0.5, and simulation gives an approximate value.

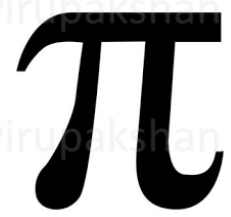
1.5.5 Monte Carlo Methods for Inferential Statistics

Example: Estimating Pi

Monte Carlo can estimate π using random points inside a square.

Imagine a circle inside a square:

Random points are generated in the square.



Formula:

$$\pi \approx 4 \times \frac{\text{Number of points inside circle}}{\text{Total number of points}}$$

1.5.6 Monte Carlo Methods for Inferential Statistics

Example: Bootstrap Mean

Suppose we have marks:

60, 70, 75, 80, 90

Bootstrap repeatedly samples with replacement:

Bootstrap Sample	Values
Original	60, 70, 75, 80, 90
Sample 1	60, 70, 70, 90, 80
Sample 2	75, 75, 60, 90, 90

Each sample **mean** is calculated.

The distribution of these **means** helps estimate uncertainty.

1.5.7 Monte Carlo Methods for Inferential Statistics

Q1. What is Monte Carlo Method?

Answer: It is a statistical method that uses random sampling for estimation.

Q2. Mention one application of Monte Carlo methods.

Answer: Risk analysis.

1.6.1 Basic Concepts of Statistics

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data.

Two Meanings of Statistics

1. Statistics as Data

Example:

- Population of India
- Average rainfall
- Literacy rate
- Sales figures

2. Statistics as a Method

Example:

Using formulas and methods to analyze exam marks or business sales.

Branches of Statistics

1. Descriptive Statistics

Describes data.

2. Inferential Statistics

Draws conclusions from samples.

1.6.2 Basic Concepts of Statistics: **Important Terms**

Variable

A variable is a characteristic that can take different values.

Example:

- Age
- Height
- Gender
- Income
- Marks

Data

Data are raw facts and figures.

Example:

```
45, 60, 70, 85, 90
```

Parameter

A parameter describes a population.

Example:

Average income of all citizens in a country.

Statistic

A statistic describes a sample.

Example:

Average income of 1,000 selected citizens.

Observation

An observation is a single data record.

Example:

Student	Age	Marks
Ravi	20	85

Diagram

```
Data
 |
 |-- Organized into variables
 |
 |-- Analyzed using statistics
 |
 |-- Converted into information
 |
 |-- Used for decision-making
```

1.7.1 Types of Data (Categorical, Numerical) / (Qualitative, Quantitative)

Data Types Diagram

```
Data
 |
 |-- Qualitative
 |       |
 |       |-- Nominal
 |       |-- Ordinal
 |
 |-- Quantitative
         |
         |-- Discrete
         |-- Continuous
```

1. Qualitative Data

Qualitative data describes categories or qualities.

Examples:

- Gender
- Blood group
- Marital status
- Product type
- City name

2. Quantitative Data

Quantitative data is numerical.

Examples:

- Age
- Height
- Income
- Weight
- Marks

1.7.2 Types of Data (Categorical, Numerical)

Types of Qualitative (Categorical) Data

a. Nominal Data

Nominal data has categories with no natural order.

Examples:

- Gender: Male, Female
- Blood group: A, B, AB, O
- Color: Red, Blue, Green

b. Ordinal Data

Ordinal data has categories with a meaningful order.

Examples:

- Education level: School, Bachelor's, Master's, PhD
- Satisfaction level: Poor, Average, Good, Excellent

1.7.3 Types of Data (Categorical, Numerical)

Types of Quantitative (Numerical) Data

a. Discrete Data

Discrete data takes countable values.

Examples:

- Number of students
- Number of cars
- Number of defects
- Number of children

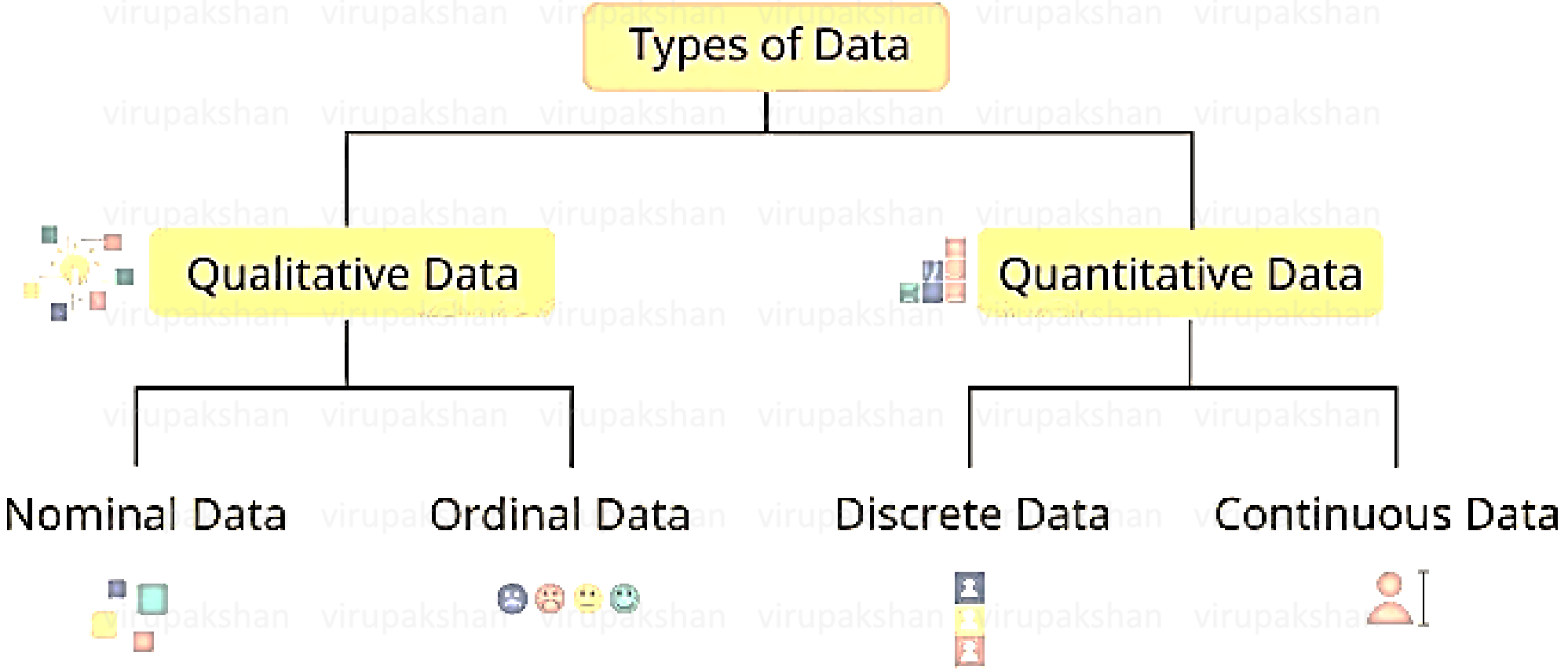
b. Continuous Data

Continuous data can take any value within a range.

Examples:

- Height
- Weight
- Temperature
- Time
- Distance

1.7.4 Types of Data (Categorical, Numerical) / (Qualitative, Quantitative)



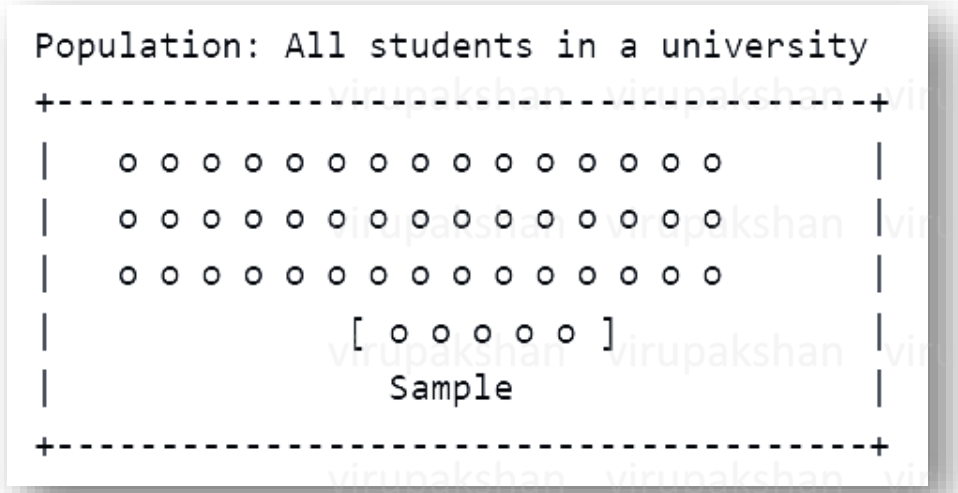
1.8.1 Populations and Samples: **DEFINITION**

Population

A **population** is the complete set of individuals, items, or observations under study.

Example:

All students in a university.



Sample

A **sample** is a subset of the population selected for analysis.

Example:

100 students selected from a university.

1.8.2 Populations and Samples: WHY USE SAMPLES?

Why Use Samples?

Studying the whole population may be:

- Expensive
- Time-consuming
- Impossible in some cases
- Impractical

Example

A company wants to know customer satisfaction for 1 million customers.

Instead of asking all customers, it surveys 2,000 customers.

Population: 1 million customers
Sample: 2,000 customers

1.8.3 Populations and Samples: **TYPES OF SAMPLING**

1. Simple Random Sampling
Every individual has an equal chance of selection.

2. Stratified Sampling
Population is divided into groups, and samples are taken from each group.

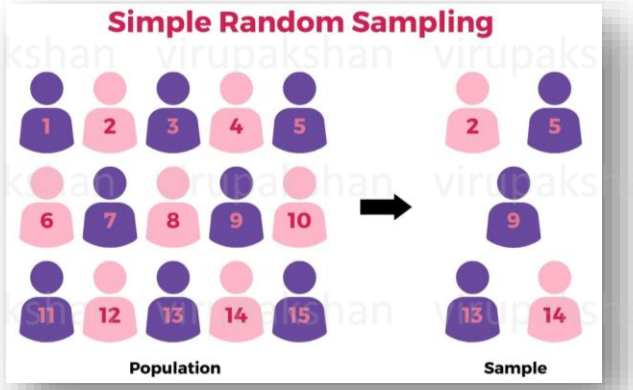
Example: Sampling students from each department.

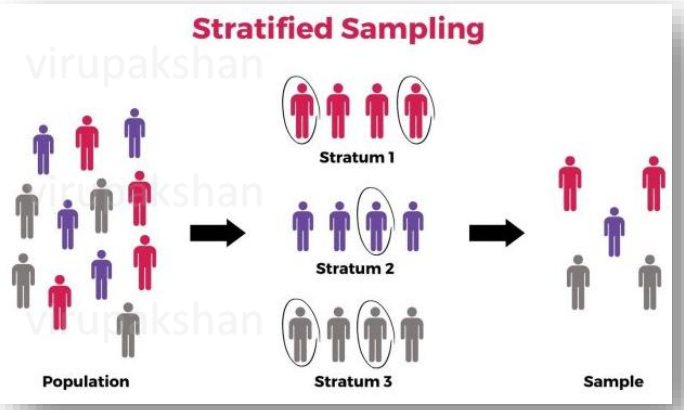
3. Systematic Sampling
Every kth item is selected.

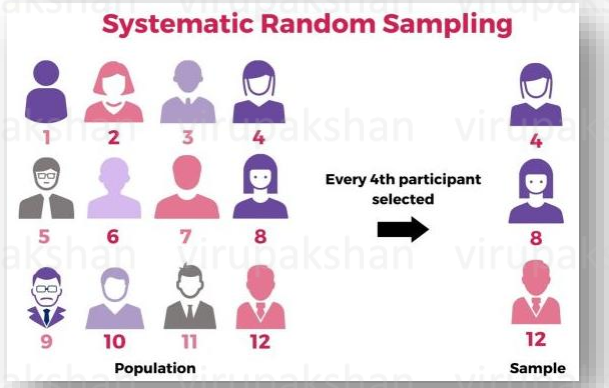
Example: Selecting every 4th customer.

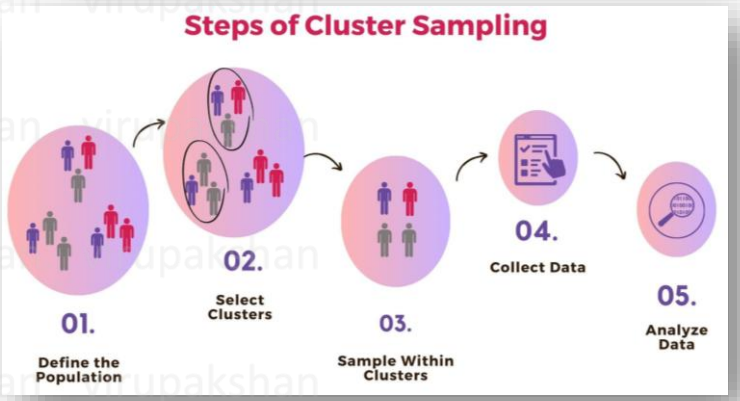
4. Cluster Sampling
Population is divided into clusters, and some clusters are selected.

Example: Selecting a few schools from a city and surveying all students in those schools.



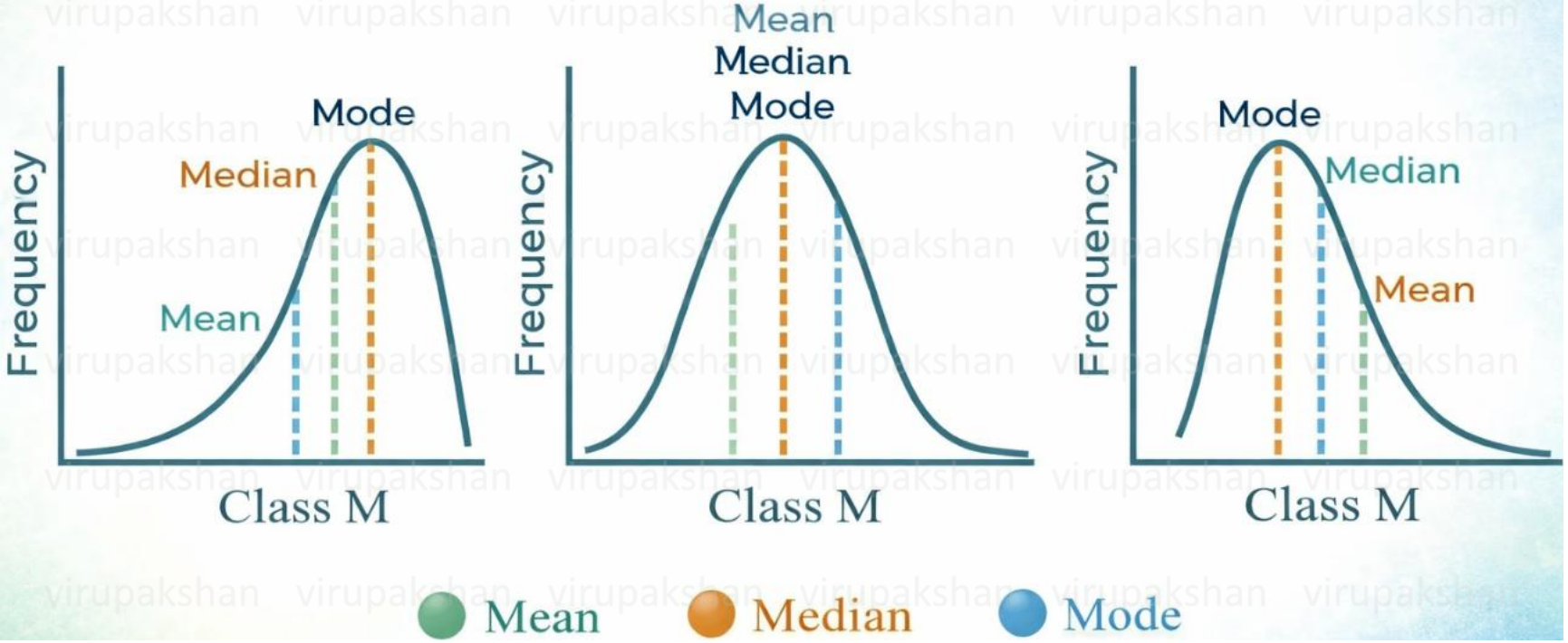






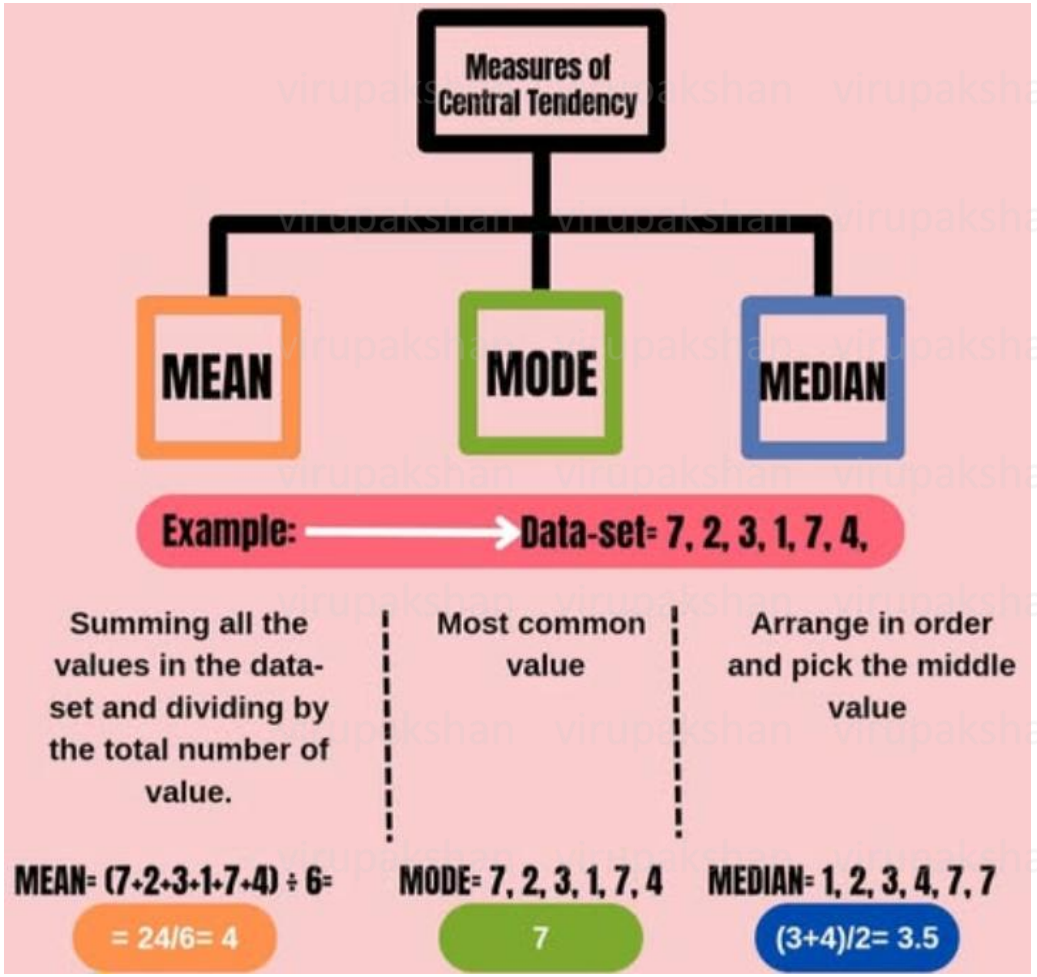
1.9.1 Measures of Central Tendency (**Mean, Median, Mode**)

Measures of central tendency are statistical summary metrics used to represent the center, middle, or most typical value of a dataset.



1.9.2 Measures of Central Tendency (**Mean**)

The **mean** is the arithmetic average.



The **mean** is the arithmetic average.

Formula:

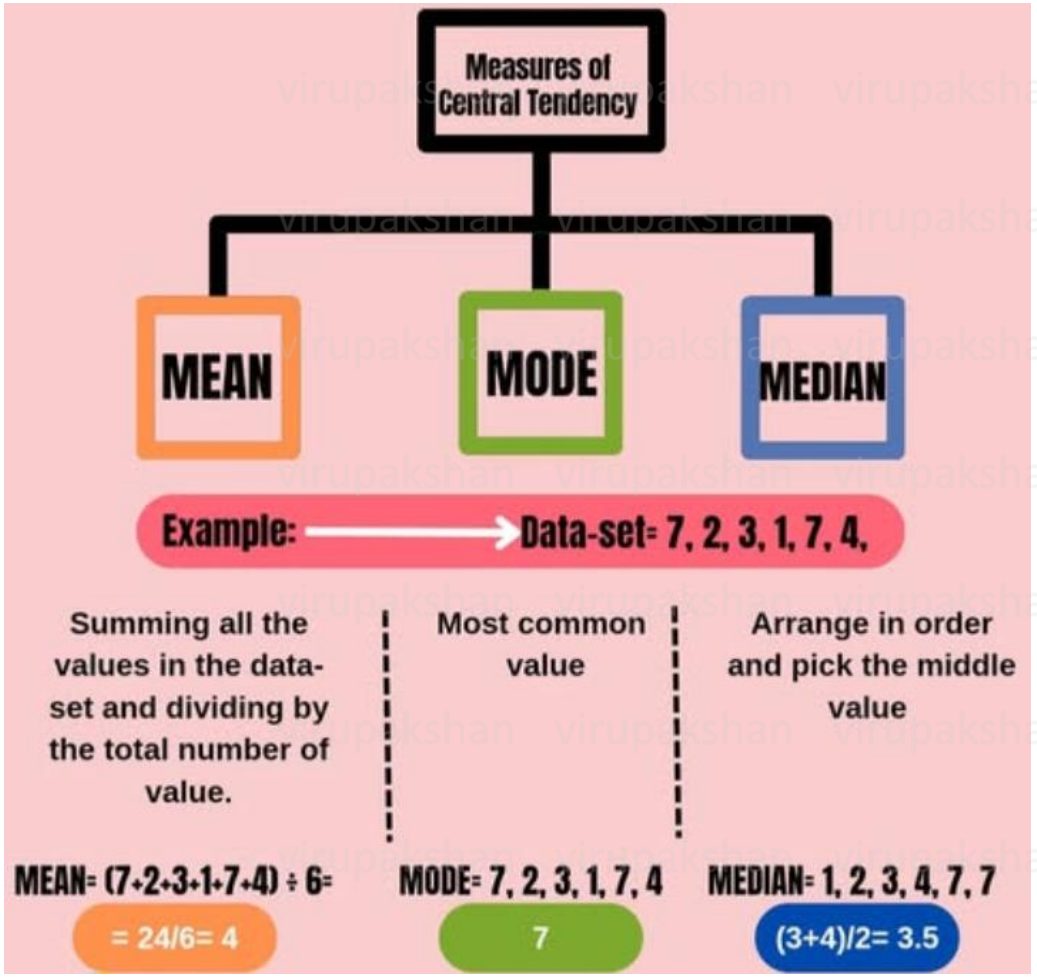
$$\bar{x} = \frac{\sum x}{n}$$

Where:

- $\bar{x}$ = sample mean
- $\sum x$ = sum of all observations
- n = number of observations

1.9.3 Measures of Central Tendency (**Mode**)

The **mode** is the value that appears <u>most frequently</u>.



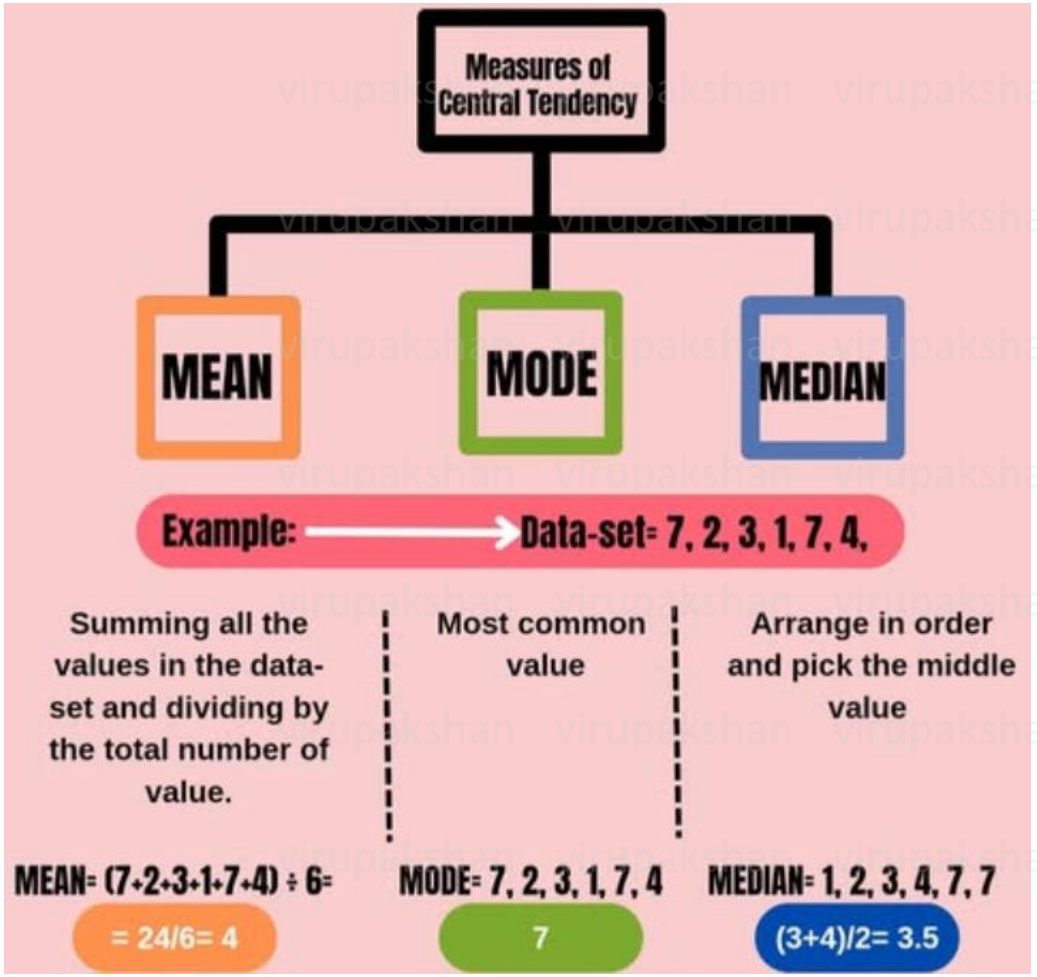
Example:

```
10, 20, 20, 30, 40
```

Mode = 20.

1.9.4 Measures of Central Tendency (**Median**)

The **median** is the middle value when data is arranged in order.



Example 1: Odd Number of Values

Data:

```
40, 50, 60, 70, 80
```

Median = 60.

Example 2: Even Number of Values

Data:

```
40, 50, 60, 70
```

Median:

$$\frac{50+60}{2} = 55$$

Median = 55.

1.9. Measures of Central Tendency (Mean, Median, Mode)

Comparison

Measure	Meaning	Best Used When
Mean	Average	Data has no extreme values
Median	Middle value	Data has outliers
Mode	Most frequent value	Categorical or repeated data

1.10.1 Measures of Variability (Range, Variance, Standard Deviation)

1. Range

Range is the difference between maximum and minimum values.

Formula: **Range = Maximum – Minimum**

Example:

20, 30, 40, 50, 60

Range:

60-20 = **40**

2. Variance

Variance measures the average squared deviation from the mean *(sample mean or population mean).*

Sample variance formula:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Population variance formula:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

3. Standard Deviation

Standard deviation is the square root of variance.

Sample standard deviation: $s = \sqrt{s^2}$

Population standard deviation: $\sigma = \sqrt{\sigma^2}$

1.10.2 Measures of Variability (Range, Variance, Standard Deviation)

Example

Data: 2, 4, 6

Mean: $$\bar{x} = \frac{2+4+6}{3} = 4$$

Deviations:

x	x - mean	Squared deviation
	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2	**-2**	**4**
4	**0**	**0**
6	**2**	**4**

Sum of squared deviations = $\sum (x_i - \bar{x})^2$

= 8.

Sample variance: $$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$s^2 = \frac{8}{3-1} = 4$$

Sample standard deviation: $$s = \sqrt{s^2}$$

$$s = \sqrt{4} = 2$$

1.11.1 Descriptive vs Inferential Statistics


```
Data Analysis
 |
 |-- Descriptive Statistics
 |       |
 |       |-- Summarize collected data
 |
 |-- Inferential Statistics
         |
         |-- Use sample to infer population
```

1.11.2 Descriptive vs Inferential Statistics

Descriptive Statistics
summarize and describe data.

Examples:

- Mean
- Median
- Mode
- Standard deviation
- Graphs
- Tables

Example

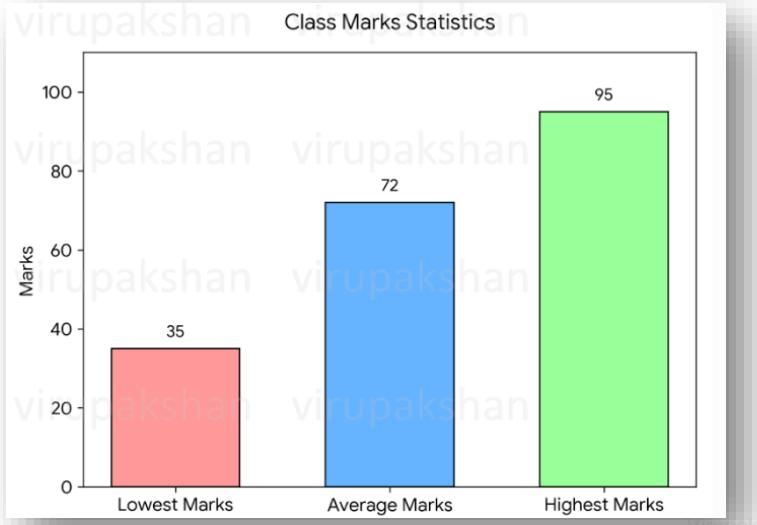
A teacher calculates the average marks of a class.

Average marks = 72

Highest marks = 95

Lowest marks = 35

This describes the class data.



1.11.3 Descriptive vs Inferential Statistics

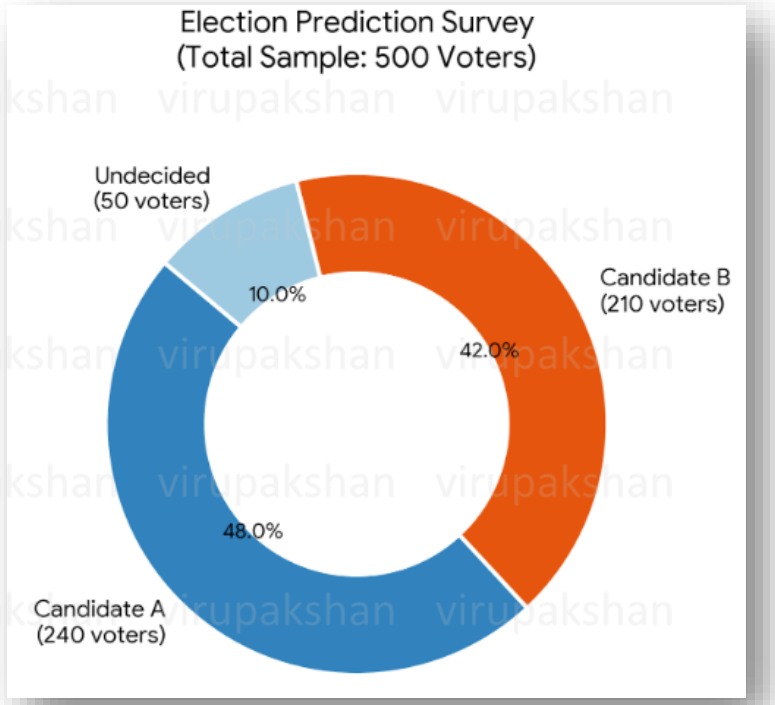
Inferential statistics use sample data to make conclusions about a larger population.

Examples:

- Estimation
- Hypothesis testing
- Confidence intervals
- Regression
- Prediction

Example:

A **survey** of 500 voters is used to **predict** the result of an election.



1.11.4 Descriptive vs Inferential Statistics

Feature	Descriptive Statistics	Inferential Statistics
Purpose	Summarizes data	Makes generalizations
Based on	Available data	Sample data
Output	Charts, averages, tables	Predictions, decisions
Example	Average salary of employees	Estimating average salary of all workers

Lecture 01 Topics

Unit 1 Introduction to Statistical Analysis

1.1	Mathematical Statistics
1.2	Statistical Pattern Recognition
1.3	Nonparametric Regression
1.4	Exploratory Data Analysis
1.5	Monte Carlo Methods for Inferential Statistics
1.6	Basic Concepts of Statistics
1.7	Types of Data (Categorical, Numerical)
1.8	Populations and Samples
1.9	Measures of Central Tendency (Mean, Median, Mode)
1.1O	Measures of Variability (Range, Variance, Standard Deviation)
1.11	Descriptive vs Inferential Statistics

LECTURE 01 TOPICS

Unit 1 Introduction to Statistical Analysis — **9 hours**

Computational Statistics- Statistical Pattern Recognition, Nonparametric Regression, Exploratory Data Analysis, Monte Carlo Methods for Inferential Statistics-Basic concepts of statistics, Types of data (categorical, numerical), Populations and samples, Measures of central tendency (mean, median, mode), Measures of variability (range, variance, standard deviation), Descriptive vs. inferential statistics.

Tutorial:

- SAS Statements, Structure, Libraries, Rules, Datasets, Representation of missing values, Assigning Libraries, PROC CONTENTS, visualize the descriptor and datasets attributes

Unit 2 Data Analysis Using SAS — **9 hours**

Basic Reports – Selecting Variables – Identifying observations – Subsetting the Data – Limiting the observations-SORT Procedure – Generating Column totals – Specifying titles and footnotes in procedure output – Assigning Descriptive labels and permanent label -How SAS Processes Programs – Compilation Phase – PDV – Descriptor Portion-Execution Phase – Data Portion – Debugging a Data Step – Testing Programs-BY-Group Processing Definitions – Preprocessing Data – Determine Whether the Data Requires Preprocessing – Sorting Observations for BY-Group Processing-FIRST. and LAST. DATA Step Variables – How the DATA Step Identifies BY Groups – Assignment Statements – Operators in SAS Expressions

Tutorial:

- Introduction of report and PRINT Procedure-Hands on practice of VAR, ID, WHERE statement, where expression[eq, ne, gt, lt, ge, le, AND, OR, CONTAINS]

Unit 3 Advanced Statistical Techniques in SAS — **9 hours**

Determining the Structure and Contents of Data Sets – Testing Program – Methods of Combining SAS Data Sets-One-to-One Reading – Concatenating – Match-Merging – Renaming Variables – Excluding Unmatched Observations -Selecting Matching Observations -Basics of DO Loops – Executing DO Loops -OUTPUT Statements – Nesting DO Loops-Iteratively Processing Observations from a Data Set -DO UNTIL Statement – DO WHILE Statement Defining and Processing Arrays –Arrays -Applying SAS Formats and Informats – Specifying SAS Formats

Tutorial:

- One-to-One Reading to Combine Data Sets-Concatenating to Combine Data Sets-Match-Merging to Combine Data Sets-Renaming Variables-Processing Iterative DO Loops-Indenting and Nesting DO Groups

Unit 4 SAS Programming and Automation — **9 hours**

Definitions – Reading Date and Time with Informats – Displaying Date and Time Values with Formats-Basics of SAS Functions - SAS Functions Syntax – Converting Data with Functions- The MEANS Procedure – Specifying Descriptive Statistics – Limiting Decimal Places – Specifying Variable Using VAR Statement – Summarized Data Set Using OUTPUT Statement – The FREQ Procedure-Specifying Variable Using TABLES Statements - Two-Way and N-Way Tables – LIST Options-The Output Delivery System (ODS) – HTML Outputs – PDF Outputs-RTF Outputs – EXCEL Outputs – The EXPORT Procedure

Tutorial:

- Reading Date and Time with Formats and Informats- Execution of Data Values Using Numeric Functions-Creating a Descriptive Statistics by Using PROC MEANS-Creating a Descriptive Statistics by Using PROC FREQ-Creating Outputs with HTML, PDF, RTF, EXCEL Formats-Exporting a External File

Unit 5 Advanced SAS Topics — **9 hours**

Creating a Default List Report – Windowing/Non Windowing Mode - Define Statement-Defining Variables – Column Statement – Selecting Observations – Defining the Variables-Group Processing the Variables by Using Different Options – Computing the Variable-Bar Chart – Box Plot – Mean Measurement Over the Time using Line Plot – Spaghetti Plot – Survival Plot – Waterfall Plot – Forest Plot-Introduction to Macro Facility – SAS Programs and Macro Processing – Macro Variables – Macro Processing – Scopes of Macro Variables – Macro Expressions-Scopes of Macro Variables – Macro Expressions-Macro Quoting – Interfaces with Macro Facility -Storing and Reusing Macros.

Tutorial:

- Creating Graphs by Using Bar Chart -Creating Graphs Using Various Types of Plots-Generating SAS Code Using Macros-Defining User-Defined Macro Variables-Creating a Tables Using PROC SQL with different Clauses- Generating a Cartesian Product -Combining Tables Using Multiple Joins

LECTURE 01

21CSE295T

Statistical Data Analytics And Integration

Unit 1 Introduction to Statistical Analysis

Doubts?

THANK YOU!

Virupakshan K

+91-9840847872 | VIRU@KART.LA