Data Analytics

BY… SAKEEB SHEIKH

What is Data Analysis ?

- Analysis of data refers to the critical examination of the assembled and grouped data for studying the characteristics of the object under study and for determining the patterns of relationships among the variables relating to it.

Purpose of Statistical Analysis?

- It summarizes data into understandable and meaningful forms.
- Statistics makes exact descriptions possible.
- Statistical analysis help in the
  - Identification of the causal factors.
  - Underlying complex phenomena.
  - In drawing of reliable inference from observed data
  - In making estimations or generalizations from the result of sample surveys.

Types of statistical analysis

- Analysis may be broadly classified into
  - Descriptive and
  - Inferential statistics.

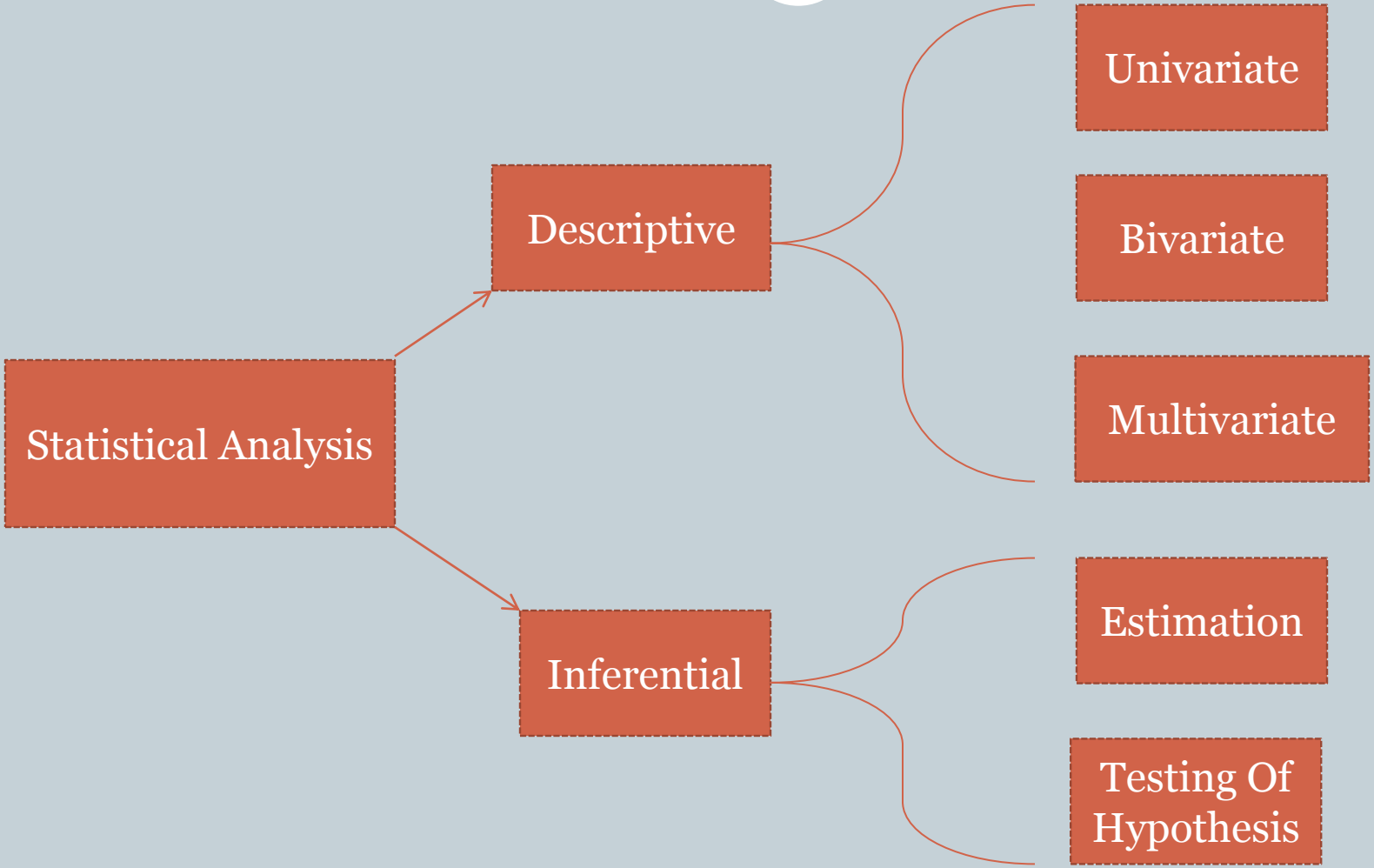
Descriptive Statistics

- Descriptive statistics describes the nature of an object under study. It may describe data on one, two or more than two variables and is accordingly called Univariate, Bivariate, and Multivariate analysis respectively.

Inferential Statistics

- Inference means a conclusion reached on the basis of evidence and reasoning.
- Inferential statistics is concerned with drawing inferences and conclusions from the findings of the research study.
- There are two areas of statistical inferences, i.e
  - Statistical estimation
  - Testing of hypothesis

Descriptive Vs Inferential Statistics



Univariate Analysis

- It is method for analyzing data on a single variable at a time, where we're observing only one aspect of phenomenon at a time.
- With Single-variable data, put all our observations into a list of numbers.
- Answers to statistical problems by collecting and analyzing data on one variable are known as **univariate analysis.**
- **Sometimes data is collected to analyze just one variable.**

Univariate Analysis…

- **For example**, if a researcher records the income of all employed residents of a particular area and tabulates that data, it would depict just one variable, the income of employed people in that area.

Representation and analysis of one variable

- As a **frequency Distribution** : it shows you the number of times an event occurs within the topic being researched.
- For example if one were to ask students about the mode of transport they take to come to college and the answers can be tabulated as follows:

Mode of transport	Frequency
Train	60
Bus	20
Bike	10
Walk	10

First Way Using Graphs

- As a Graph :
  - Histogram
  - Frequency Polygon (Line Chart)
  - Pie Chart
  - Box Plot
  - Scattered Plot

Second Way (Measure of Central Tendency)

- As a measure of Central tendency
  - Mean (avg)
  - Median (center most value)
  - Mode (Most repeated value)

Third Way (Spread of the data)
Measures of Variability

- **Range** (Difference between minimum and maximum value.)
- **Variance.**
- **Standard Deviation**
- **Quartiles and the Interquartile Range**

Continue…

- **Skewness** : Lack of Symmetry (Effect Of Outliers)
- **Kurtosis** : Sharpness of the peak of the frequency distribution

Bivariate Analysis

- Gathering Information about more than one variable.
- **Example** : To Obtain the literacy rate of a population we also consider other variable like age, sex, family income, availability of institutions.
- When only two variables are under consideration we say **Bivariate**.

Types of variables/ Level of Measurement

- Nominal Variables. (Labeling)
- Ordinal Variables. (Frequency Distribution, %)
- Interval Variables. (regression analysis)
- Ratio Measures. (regression analysis)

Association

- Testing the association between the two variables and causality.
- Measure of association is called Correlation between variables.

Summary Statistics that decribe a variable's numeric values

- address = 'datasets/mtcars.csv'
- cars = pd.read_csv(address)
- cars.columns = ['car_names','mpg','cyl','disp', 'hp', 'drat', 'wt', 'qsec', 'vs', 'am', 'gear', 'carb']

- cars.sum()
- cars.sum(axis=1)
- cars.median()
- cars.mean()
- cars.max()

- mpg = cars.mpg

#Gives index position of the maximum value in mpg column

- mpg.idxmax()

Describe variable Distribution

- cars.std()
- cars.var()
- gear = cars.gear
- gear.value_counts()
- cars.describe()

Outliers

- Univariate Method
- Bivariate Method