XV Online Faculty Development Programme on Research Methodology

(Comprehensive Research Techniques)

Jointly Organized By

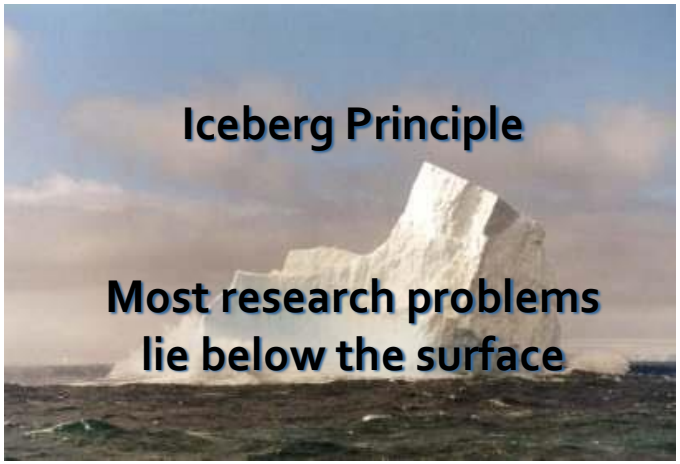
Dr C.V. Raman University, Ballarpur, Chhattisgarh & Grand Academic Portal

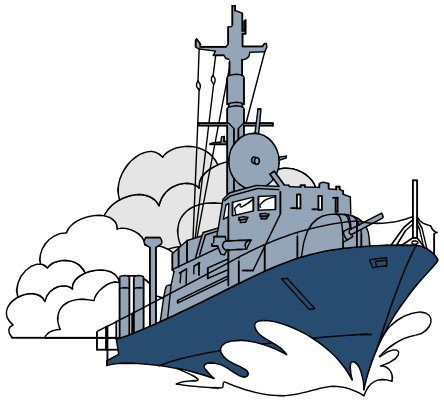
Introduction to Multivariate Analysis

Dr. Hemal Pandya
Professor, S.D. School of Commerce, Gujarat University, Ahmedabad

The Iceberg Principle

- The principle indicating that the dangerous part of many research problems is neither visible to nor understood by researchers.





Bounded Rationality

- This is the behavior that people construct simplified models that extract the essential features from problems without capturing all of their complexities in order to decide rationally.

What is Research?

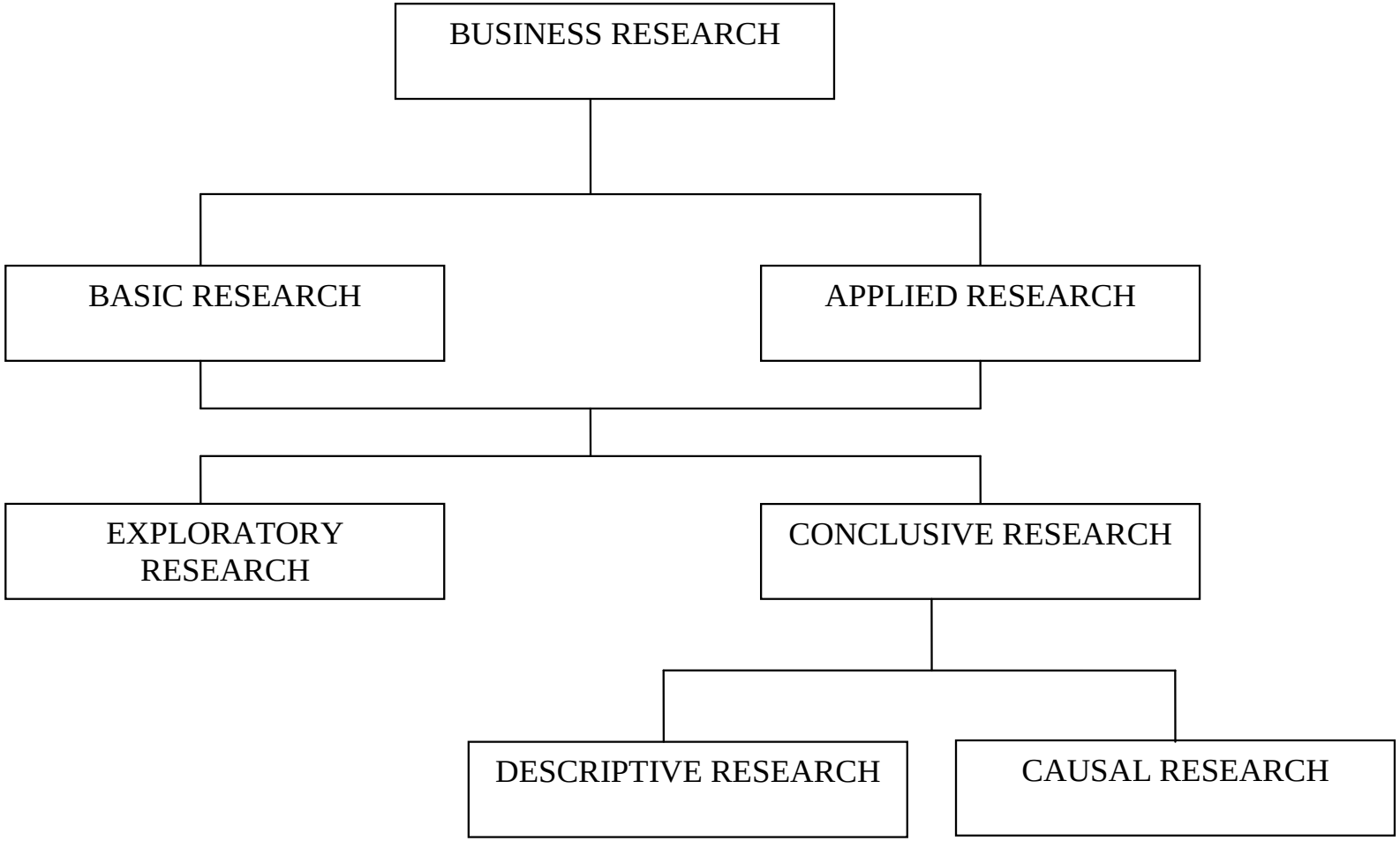
- Research in Social Sciences and Research Process

Eight Step Business Research Process

Types of Research

- Basic or Pure Research
- Applied Research
- Diagnostic Study
- Evaluation Study
- Action Research
- Historical Research
- Experimental Research
- Literature Review Research
- Library Research
- Paradigm Research
- Pre-Paradigm Research

Research Designs



Sampling Design

Probability Sampling Design - Probability sampling designs are used in conclusive research. In a probability sampling design, each and every element of the population has a known chance of being selected in the sample.

Types of Probability Sampling Design

- Simple random sampling with replacement
- Simple random sampling without replacement
- Systematic sampling
- Stratified random sampling
- Cluster sampling
- Two Stage Sampling
- Multi-stage Sampling
- Probability Proportional to size sampling
- Area Sampling

Sampling Design

Non-probability Sampling Designs - In case of non-probability sampling design, the elements of the population do not have any known chance of being selected in the sample.

Types of Non-Probability Sampling Design

- Convenience sampling
- Judgmental sampling or Purposive Sampling
- Snowball sampling
- Quota sampling

Determination of Sample Size

Sample size for estimating population mean - The formula for determining sample size is given as:

$$n = \frac{Z^2 \sigma^2}{e^2}$$

Where

n = Sample size

σ = Population standard deviation

e = Margin of error

Z = The value for the given confidence interval

Determination of Sample Size

Sample size for estimating population proportion –

1. When population proportion p is known

$$n = \frac{Z^2 pq}{e^2}$$

2. When population proportion p is not known

$$n = \frac{1}{4}\frac{Z^2}{e^2}$$

Initial Step- Nature of Variables

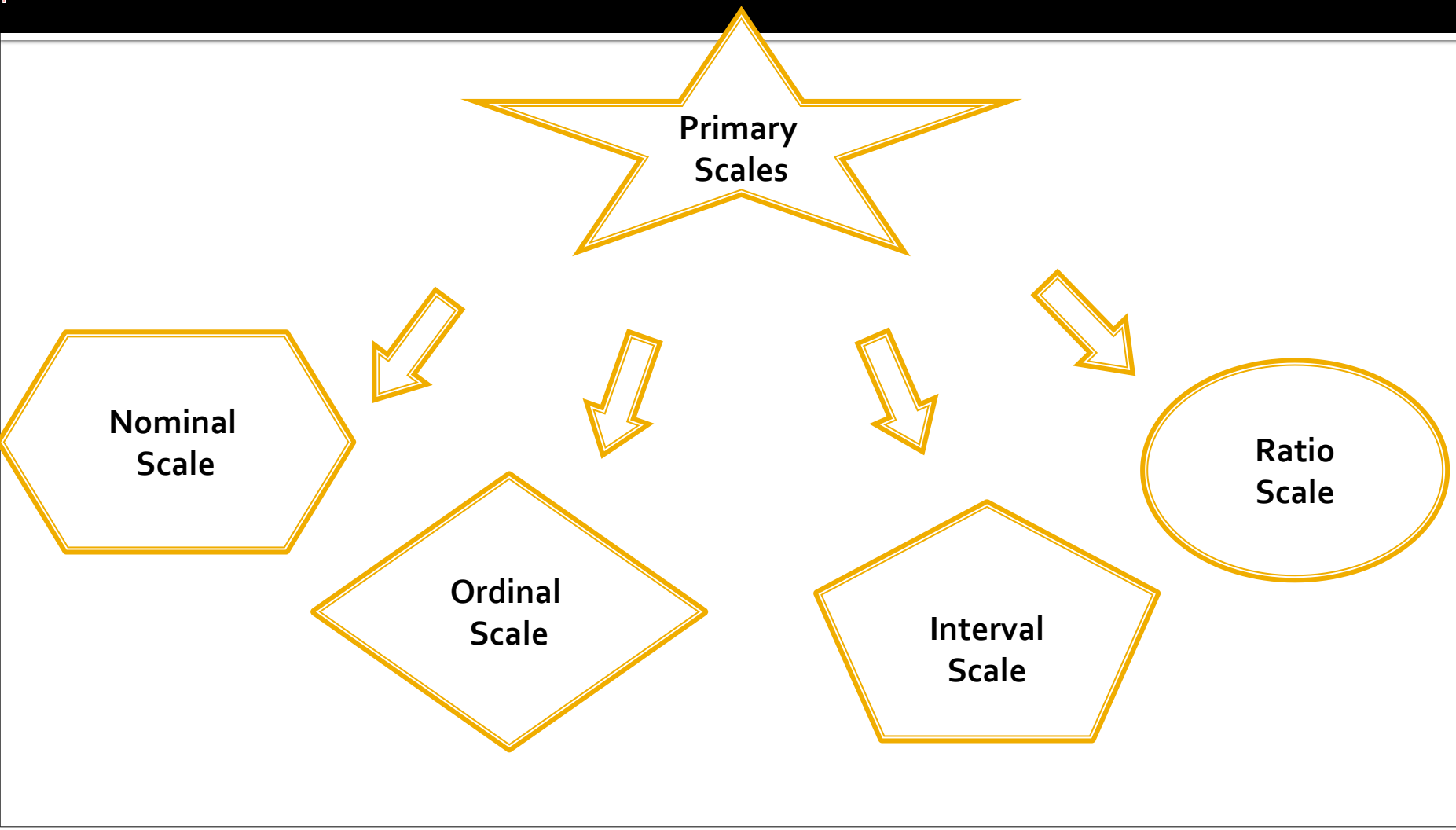
- In order to understand multivariate analysis, it is important to understand some of the terminology. A variate is a weighted combination of variables. The purpose of the analysis is to find the best combination of weights. Nonmetric data refers to data that are either qualitative or categorical in nature. Metric data refers to data that are quantitative, and interval or ratio in nature.

SOME BASIC CONCEPTS OF MVA

- **MEASUREMENT SCALE**
- Data can be classified in to two categories:

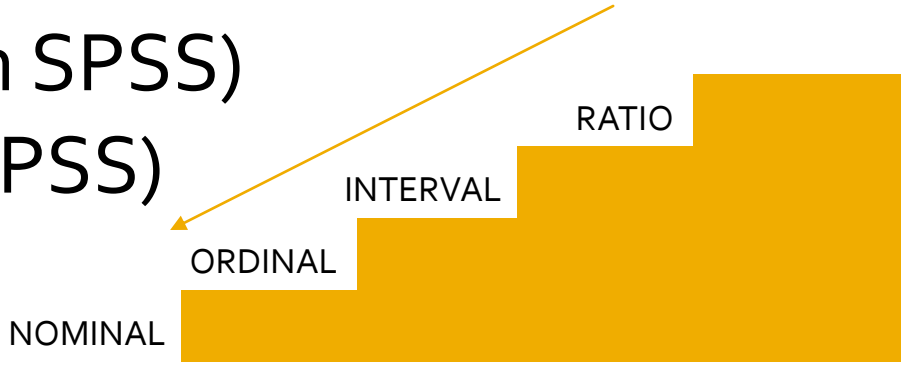
a) Non-metric (qualitative)
b) Metric (quantitative)

Measurement Scales



LEVELS OF DATA

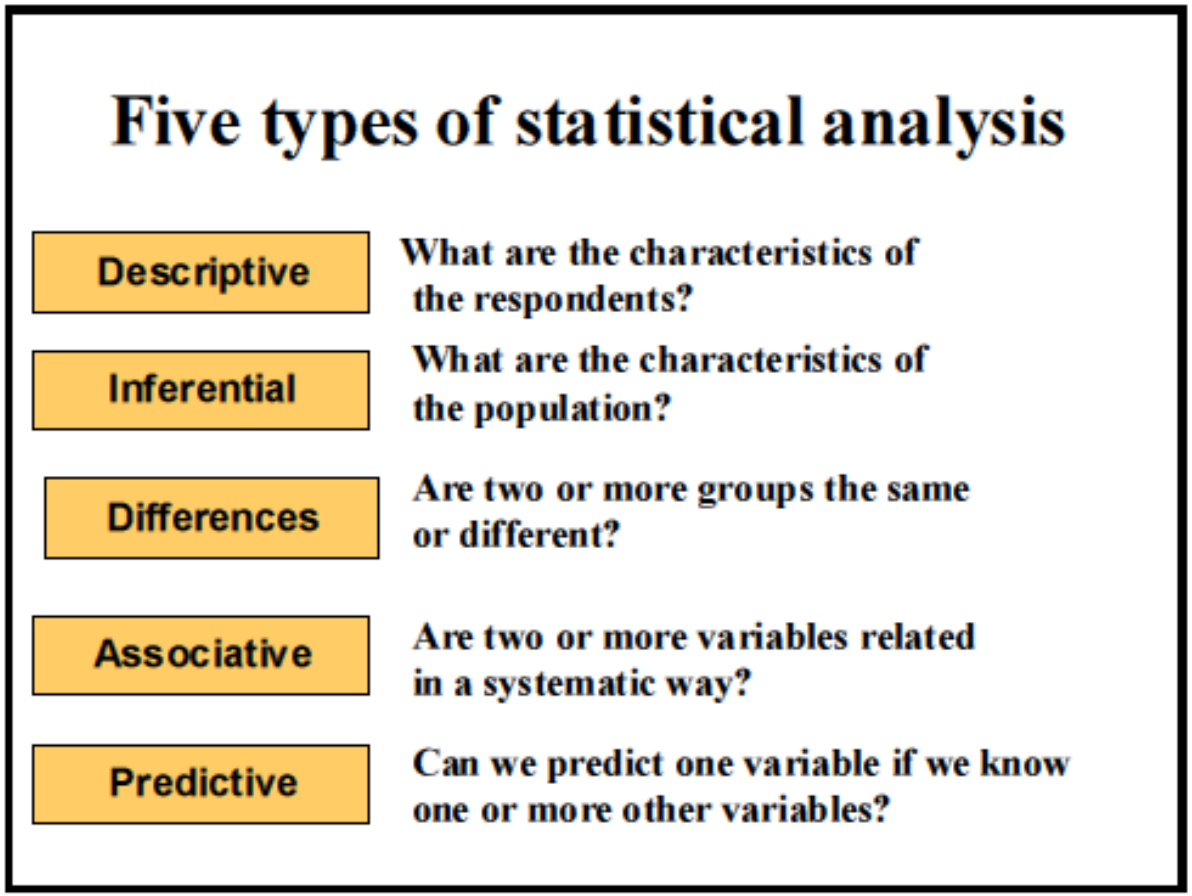
- Nominal
- Ordinal
- Interval (Scale in SPSS)
- Ratio (Scale in SPSS)



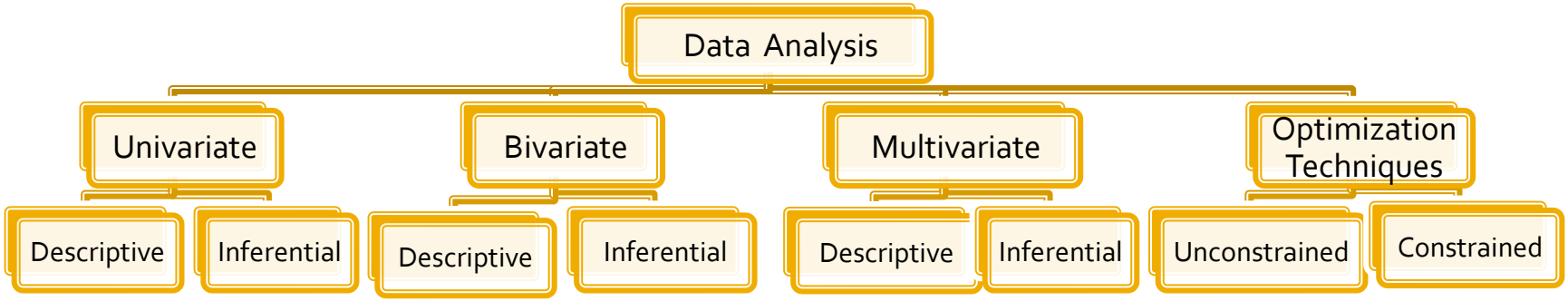
The impact of choice of measurement scale

- The researcher must identify scale of each variable used.
- So that non-metric data are not incorrectly used as metric data and vice versa.
- The measurement scale is very important in determining which multivariate applications are the most applicable to the data.

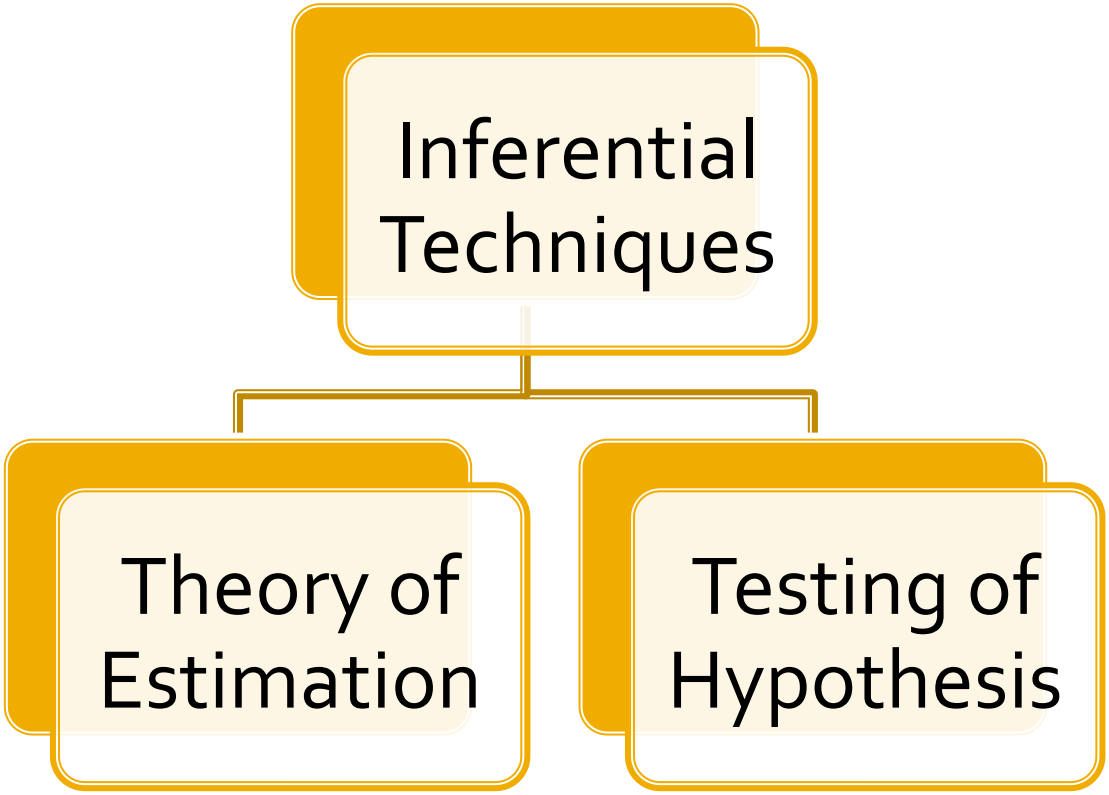
Types of Statistical Data Analysis



Techniques of Data Analysis



Inferential Techniques

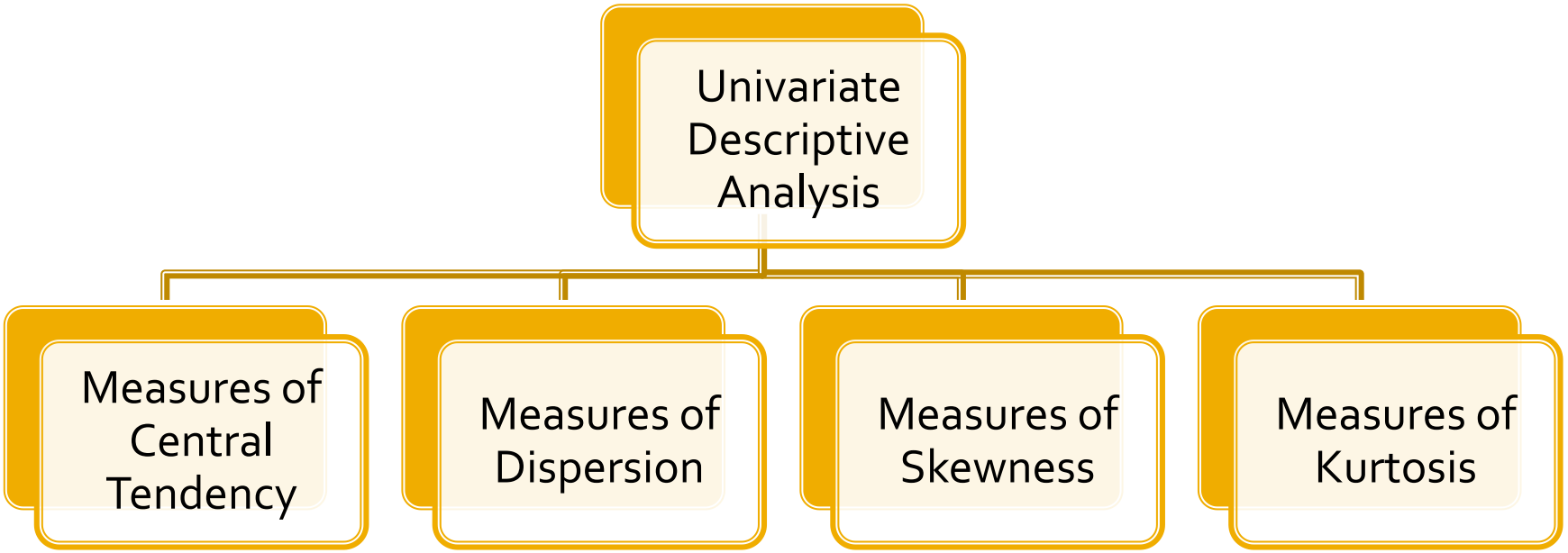


UNIVARIATE ANALYSIS

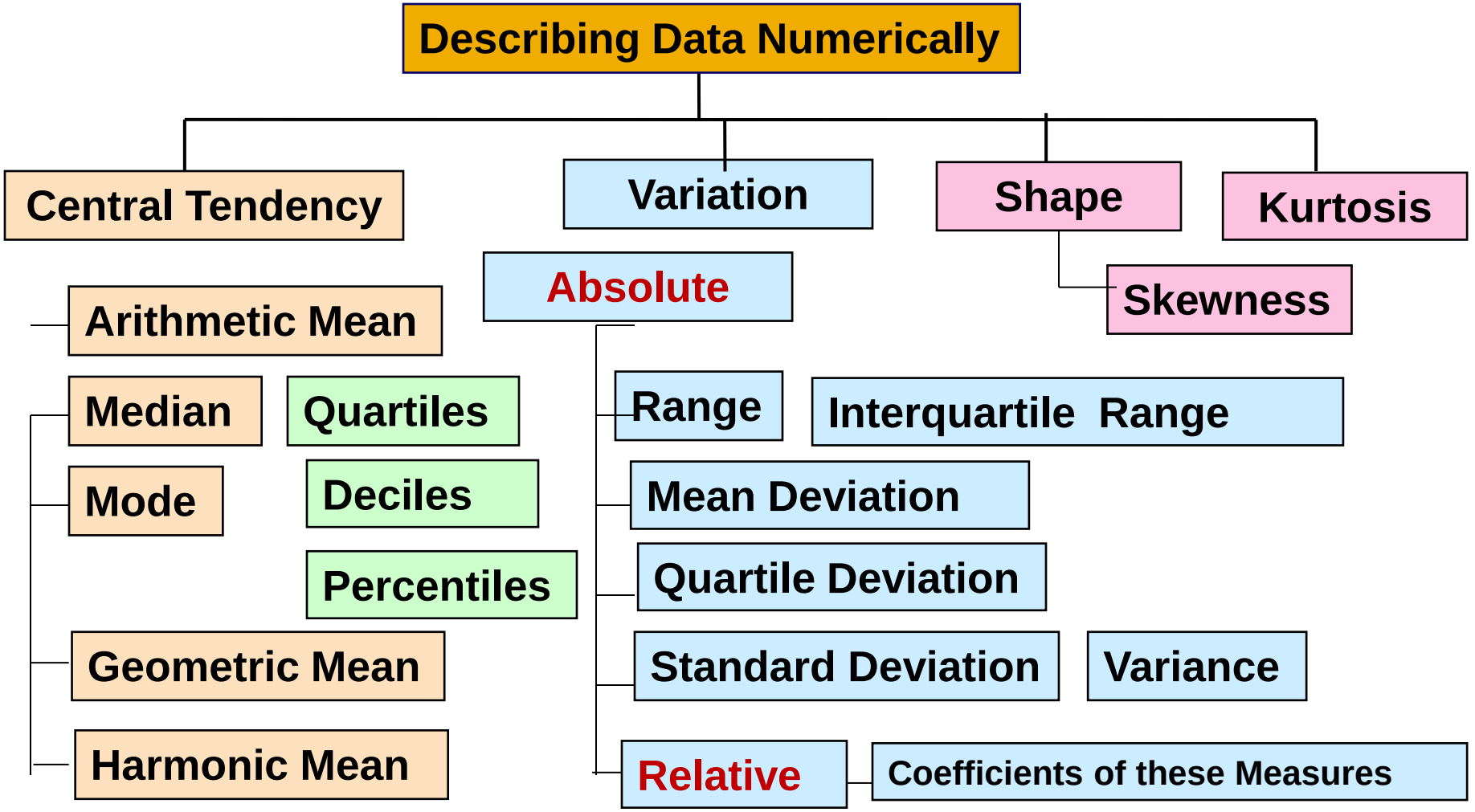
- Univariate data is used for the simplest form of analysis.
- It is the type of data in which analysis are made only based on one variable.
- Purpose: Mainly description
- Thus, it takes the data, summarizes that data and find the patterns in the data.
- There are some ways to describe patterns found in univariate data which include:

a) Graphical methods, (bar & pie charts, histograms)
b) Measures of central tendency (mean, median, mode)
c) Measures of dispersion, Skewness and Kurtosis (range, variance, SD)

Univariate Descriptive Statistics



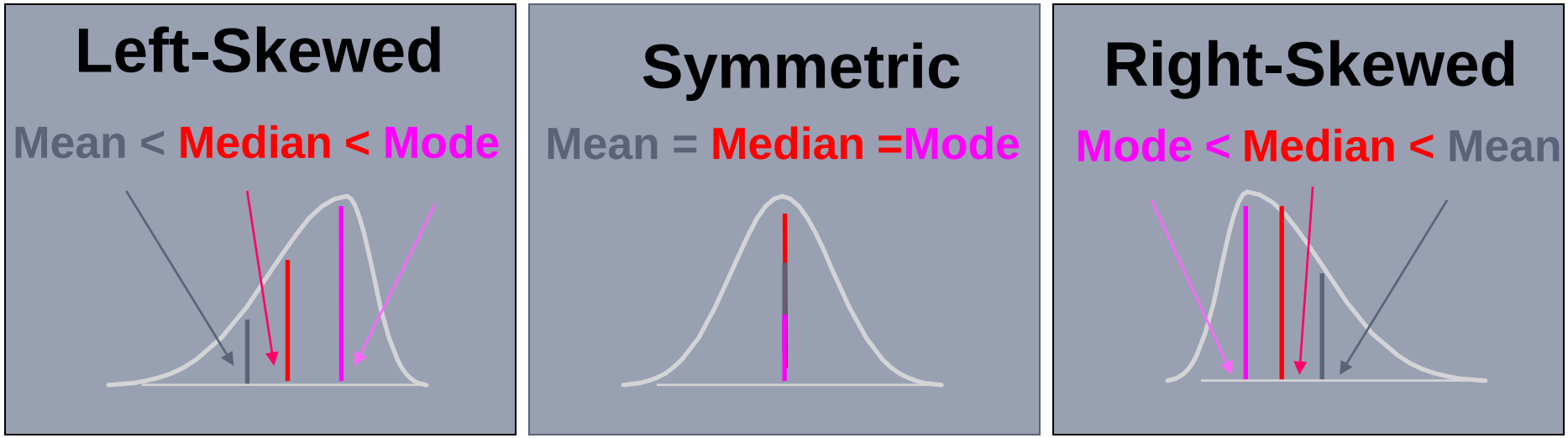
Univariate Analysis: Summary Measures



Descriptive Analysis of Univariate Data: Skewness

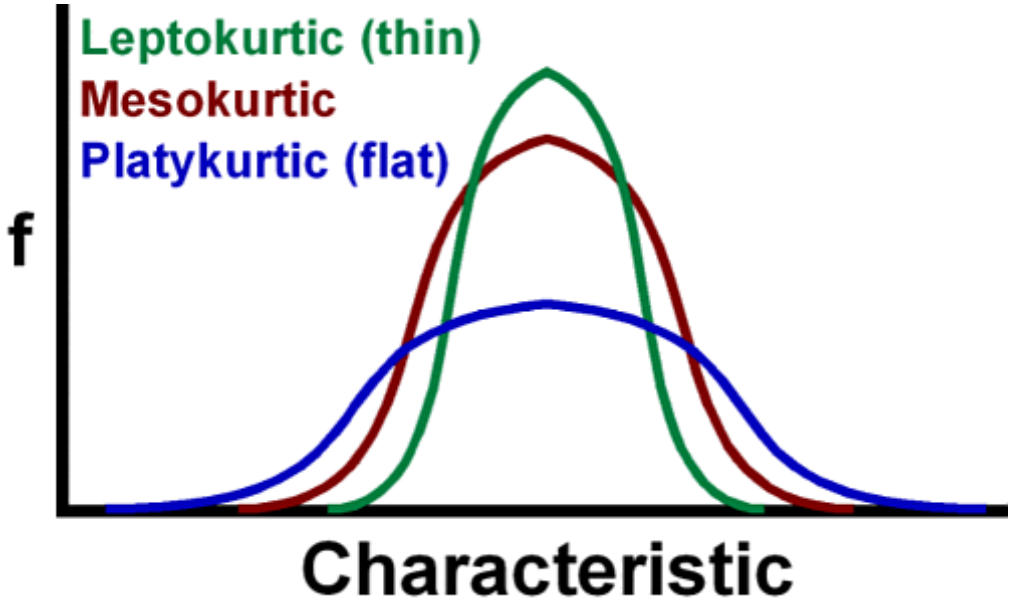
Shape of a Distribution: Symmetric or skewed

- Describes how data is distributed
- Measures of shape



Descriptive Analysis of Univariate Data: KURTOSIS

- The height and sharpness of the peak relative to the rest of the data or the slope of the curve and the frequency height is called Kurtosis. Balanda and MacGillivary say that increasing Kurtosis is associated with the movement of the probability mass from the shoulders of a probability distribution into its center and tails

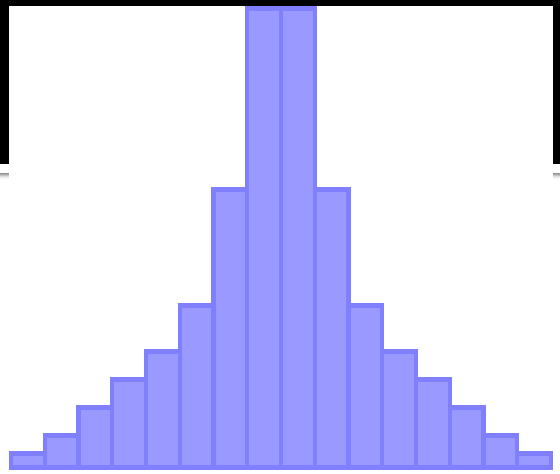


Descriptive Analysis of Univariate Data: Kurtosis

The reference standard is a normal distribution, which has a kurtosis of 3. In token of this, often the **excess kurtosis** is presented: excess kurtosis is simply **kurtosis–3**. For example, the "kurtosis" reported by Excel is actually the excess kurtosis.

- A normal distribution has kurtosis exactly 3 (excess kurtosis exactly 0). Any distribution with kurtosis ≈3 (excess ≈0) is called **mesokurtic**.
- A distribution with kurtosis <3 (excess kurtosis <0) is called **platykurtic**. Compared to a normal distribution, its central peak is lower and broader, and its tails are shorter and thinner.
- A distribution with kurtosis >3 (excess kurtosis >0) is called **leptokurtic**. Compared to a normal distribution, its central peak is higher and sharper, and its tails are longer and fatter.

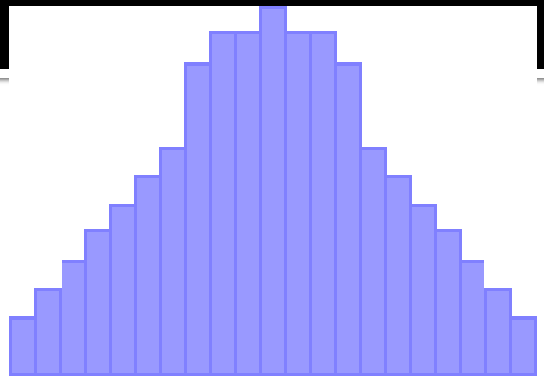
Kurtosis



Leptokurtic

(high peak)

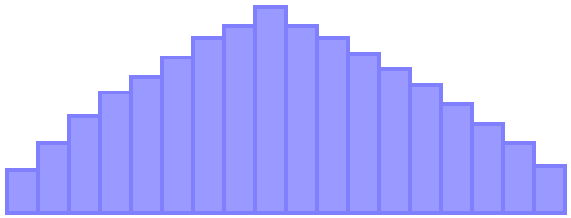
(+ve kurtosis)



Mesokurtic

(normal)

(zero kurtosis)



Platykurtic

(low peak)

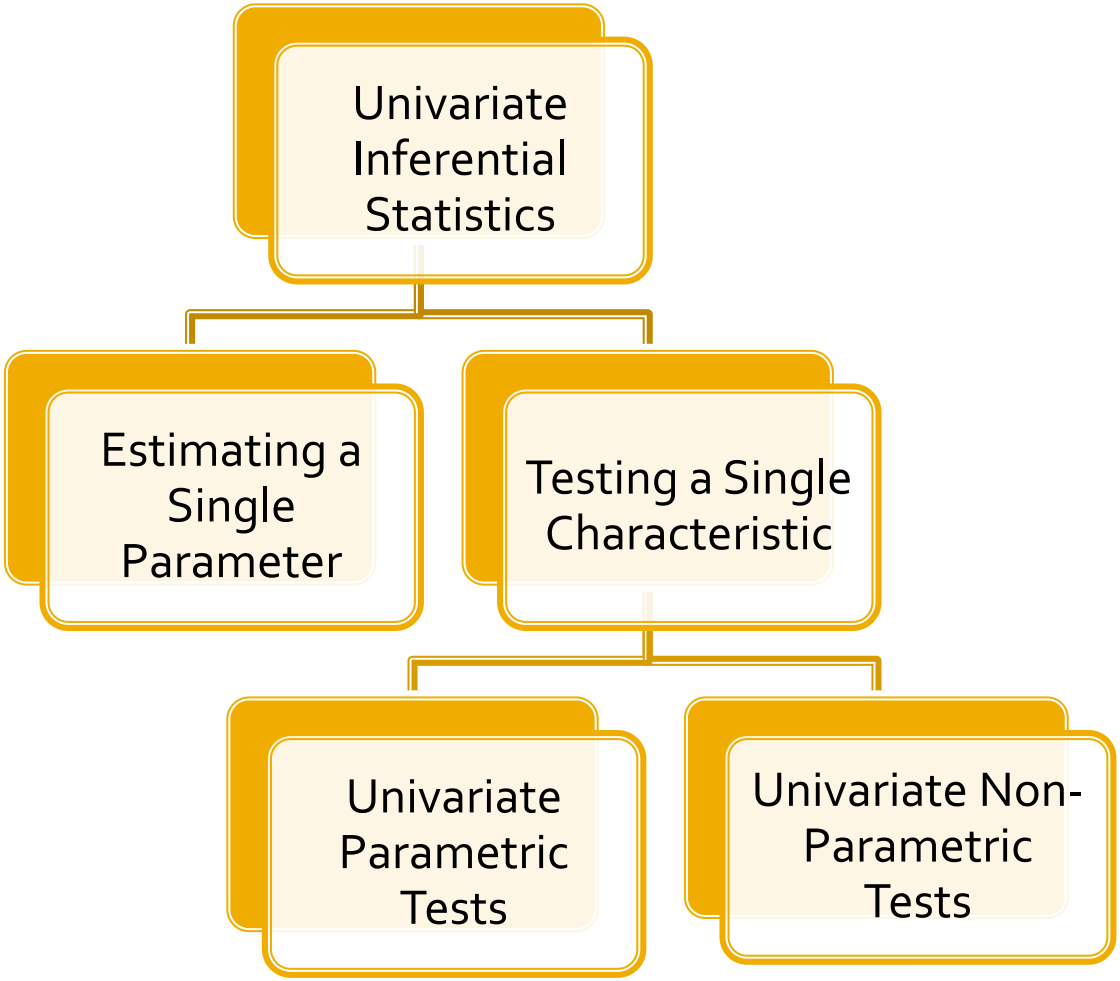
(–ve kurtosis)

Mesokurtic distribution…kurtosis = 3

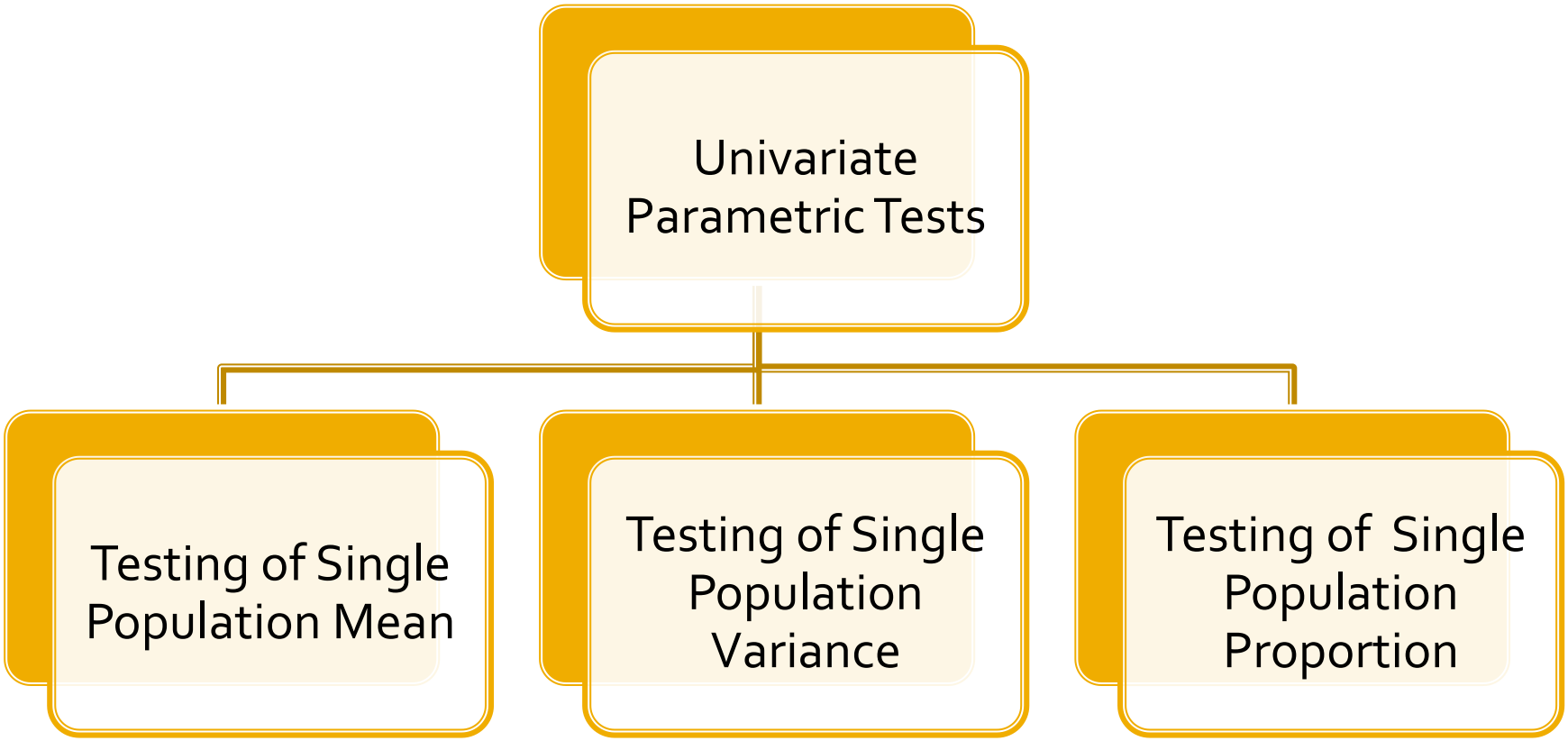
Platykurtic distribution…kurtosis < 3

Leptokurtic distribution…kurtosis > 3

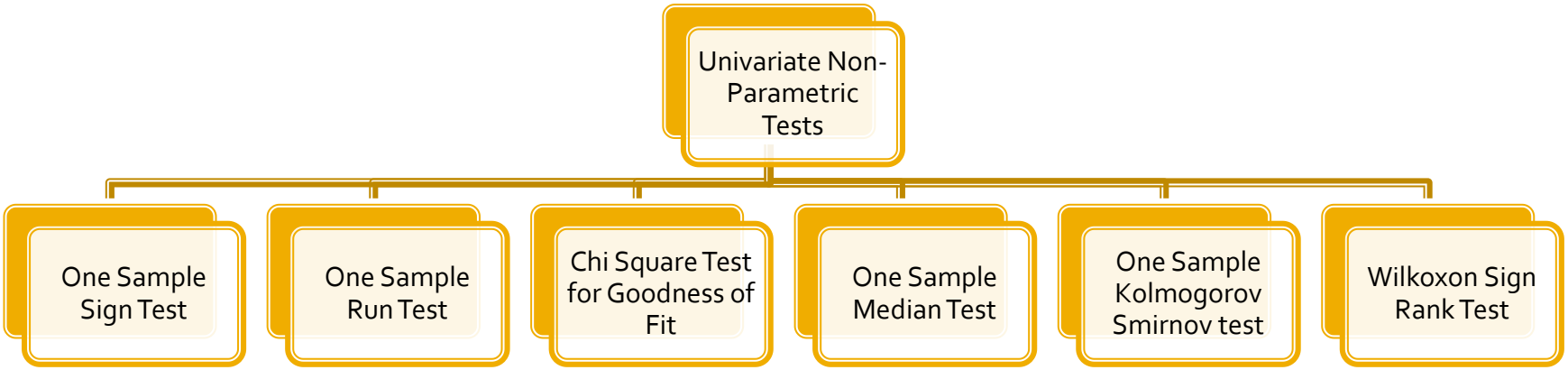
Univariate Inferential Statistics



Univariate Inferential Statistics



Univariate Inferential Statistics



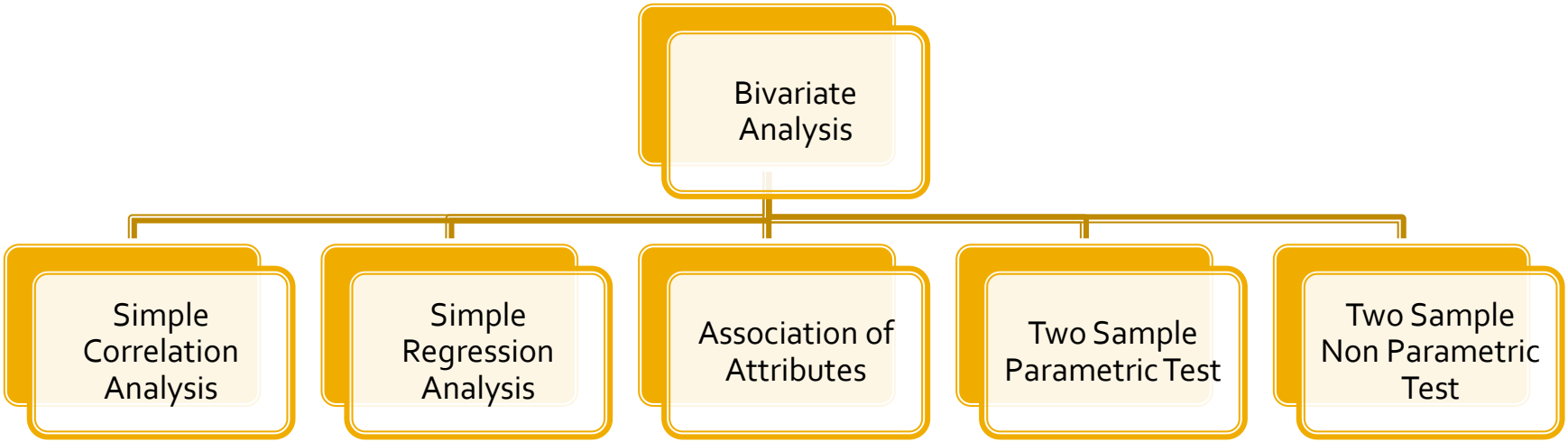
BIVARIATE ANALYSIS

- Bivariate data is used for little complex analysis than as compared with univariate data.
- Bivariate data is the data in which analysis are based on two variables per observation simultaneously.
- Purpose: Determining empirical relationship between two variables
- Bivariate analysis is a simple (two variable) special case of multivariate analysis .

BIVARIATE ANALYSIS

- Bivariate analysis can be helpful in testing simple hypotheses.
- It can help to determine at what extent it becomes easier to know and predict a value for one variable (possibly a dependent variable) if we know the value of the other variable (possibly the independent variable).
- Cross classification, correlation, analysis of variance, simple regression etc are some applications of bivariate analysis.

Tools and Techniques for Bivariate Analysis



Statistics for assessing an association between two Variables, unpaired data

Risk factor (independent variable, exposure, group assignment)	Outcome (dependent variable)					
	Dichotomous	Nominal	Interval, normal distribution	Interval non-normal	Ordinal	Time to event, censored data
Dichotomous	Chi-squared, Fisher's exact test, risk ratio, odds ratio	Chi-squared	*t*-test	Mann-Whitney test	Chi-squared for trend, Mann-Whitney test	Log-rank, Wilcoxon, rate ratio
Nominal	Chi-squared, exact test	Chi-squared	ANOVA	Kruskal–Wallis test	Kruskal–Wallis test	Log-rank, Wilcoxon
Interval, normal distribution	*t*-test	ANOVA	Linear regression, Pearson's correlation coefficient	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	–
Interval, non-normal	Mann-Whitney test	Kruskal–Wallis test	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	–
Ordinal	Chi-squared for trend, Mann-Whitney test	Kruskal–Wallis test	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	Spearman's rank correlation coefficient	–

Comparison of Bivariate tests for unpaired and paired data

	Independent observations (2 groups)	Paired observations (2 observations)	Independent observations (≥3 groups)	Repeated observations (≥3 observations)
Dichotomous variable	Chi-squared Fisher's exact	McNemar's test	Chi-squared	Cochran's *Q*
Normally distributed interval variable	*t*-test	Paired *t*-test	ANOVA	Repeated-measures ANOVA
Non-normally distributed interval variable	Mann-Whitney test	Wilcoxon signed rank test	Kruskal–Wallis test	Friedman statistic
Ordinal variable	Mann-Whitney test	Wilcoxon signed rank test	Kruskal–Wallis test	Friedman statistic

MULTIVARIATE ANALYSIS:
Introduction

MULTIVARIATE ANALYSIS

- Multivariate analysis techniques are popular because they enable organization to create knowledge and thereby improve their decision making.
- MVA refers to all statistical techniques that simultaneously analyze multiple measurements on individuals or objects under investigation.
- Thus, any simultaneous analysis of more than two variables can be loosely considered MULTIVARIATE ANALYSIS.
- Purpose: Determining empirical relationship among multiple variables

MULTIVARIATE ANALYSIS

- Many multivariate techniques are extension of univariate analysis and bivariate analysis. for example: MANOVA, Multiple regression etc.

A STRUCTURED APPROACH TO MULTIVARIATE MODEL BULIDING

- STAGE- 1
- **Define the research problem, Objectives and Multivariate techniques to be used**
  - A conceptual model need not be complex and detailed; it can be just a simple representation of the relationship to be studied.
  - Model specified means → the researcher has to choose an appropriate multivariate technique based on measurement characteristics of the dependent and independent variables.

A STRUCTURED APPROACH TO MULTIVARIATE MODEL BULIDING

- Stage – 2
- **Develop the Analysis Plan**
  - At this stage attention turns to the implementation issues.
  - Minimum or required sample size, Allowable or required type of variables,
  - Data collection method etc.

A STRUCTURED APPROACH TO MULTIVARIATE MODEL BULIDING

- Stage – 3
- **Evaluate the assumptions underlying the multivariate technique**
  - With the data collected, the first task is to evaluate its underlying assumptions, both statistical and conceptual.
  - For the techniques based on statistical inference the assumptions of multivariate normality, linearity, independence of error terms, and equality of variance must all be met.
  - Each technique also involves a series of conceptual assumptions dealing with such issues as model formulation and the types of relationships to be represented.

A STRUCTURED APPROACH TO MULTIVARIATE MODEL BULIDING

- Stage – 4
- **Estimate the Multivariate Model and assess the overall model fit**
  - Actual estimation of the multivariate model and assessment of overall model fit.
  - After the model estimated to ascertain whether it achieves acceptable levels on statistical criteria(i.e. level of significance), identified the proposed relationships and achieves practical significance.
  - Many times the model will be respecified in an attempt to achieve better levels of overall fit and/ or explanations.
  - Overall fitting may be identified as outliers, influential observations or the other disparate results(e.g.; single member clustering)

A STRUCTURED APPROACH TO MULTIVARIATE MODEL BULIDING

- Stage- 5
- **Interpret the variates**
  - Interpreting the variates revels the nature of multivariate relationships.
  - The objective is to identify empirical evidence multivariate relationship in sample data that can be generalized to the total population.

A STRUCTURED APPROACH TO MULTIVARIATE MODEL BULIDING

- Stage-6
- **Validate the multivariate model**
  - i.e. final approval to generalization to the total population

Data Preparation

MULTIVARIATE ANALYSIS

Initial Step—Data Quality

- Before launching into an analysis technique, it is important to have a clear understanding of the form and quality of the data.
- The form of the data refers to whether the data are nonmetric or metric.
- The quality of the data refers to how normally distributed the data are.
- The first few techniques discussed are sensitive to the linearity, normality, and equal variance assumptions of the data.
- Examinations of distribution, skewness, and kurtosis are helpful in examining distribution.
- Also, it is important to understand the magnitude of missing values in observations and to determine whether to ignore them or impute values to the missing observations.
- Another data quality measure is outliers, and it is important to determine whether the outliers should be removed. If they are kept, they may cause a distortion to the data; if they are eliminated, they may help with the assumptions of normality. The key is to attempt to understand what the outliers represent.

Data Preparation for multivariate analysis

- Once data is collected, process of analysis begins…
  but, data has to be translated in an appropriate form.
- This process is known as Data Preparation or Data Processing.

Steps in Data Preparation

- Authenticity, Reliability and Validity of data
- Questionnaire checking
- Edit acceptable questionnaires
- Code the questionnaires
- Data Cleaning: Clean the data set: Consistency Check, Checking for extreme values, Statistically adjust the data, Missing Values and Imputation, Outliers

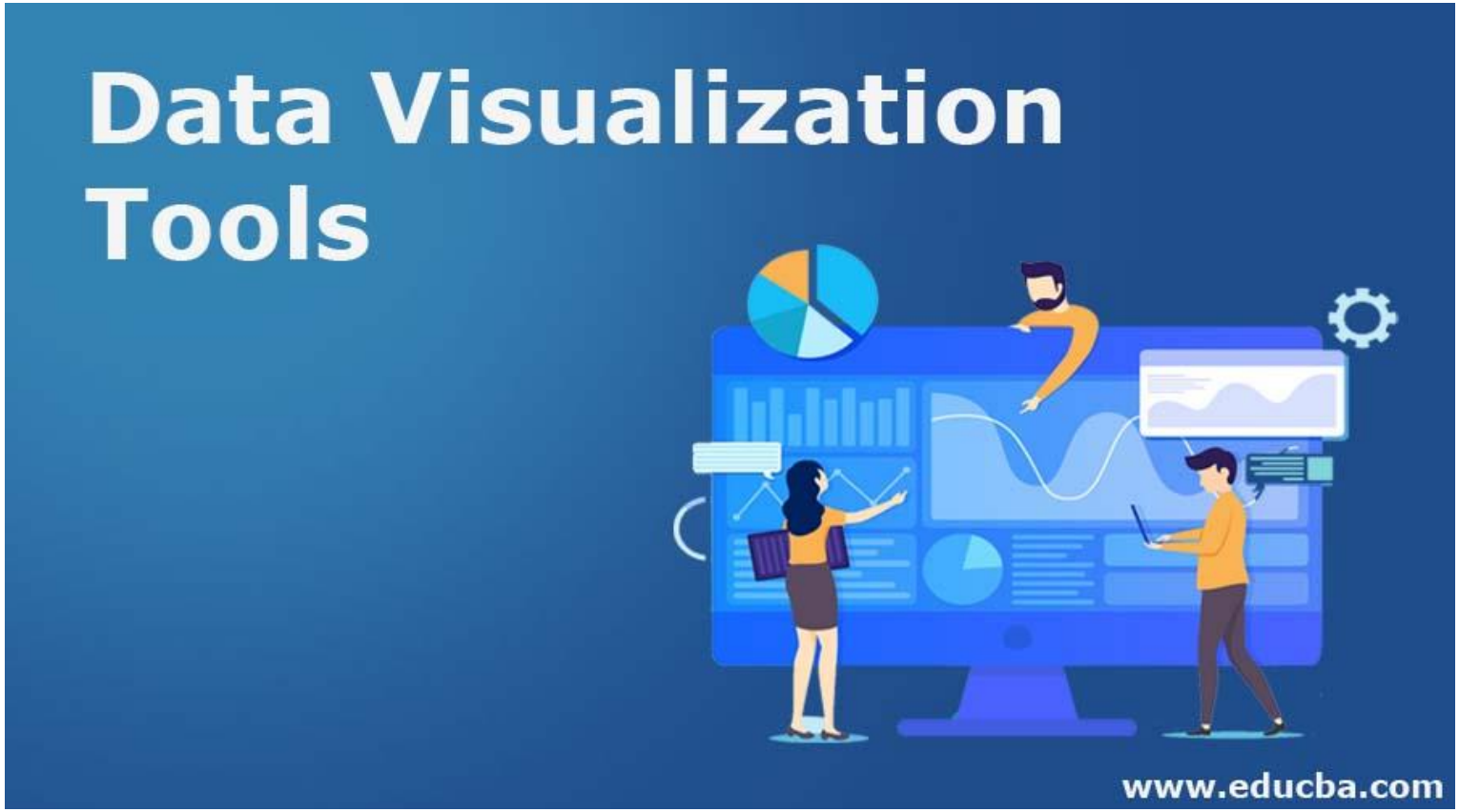
Statistically adjusting the data

- Weighting: each cases/respondent in database is assigned a weight to reflect its importance
- Weighting is most widely used to make the sample data more representative of target population on specific characteristics
- Variable re-specification : involves transformation of data to create new variables or to modify existing ones.

SOME BASIC CONCEPTS OF MVA

- **THE VARIATE**:
- The building block of multivariate analysis is the variate.
- Variate is a linear combinations of Variables with empirically determined weights.
- Variate value = $w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$
- Single value represents a combination of the entire set of variables that best archives the objective of the specific multivariate analysis.
- Here, variables are specified by researcher and weights are determined by multivariate technique.

Data Visualization

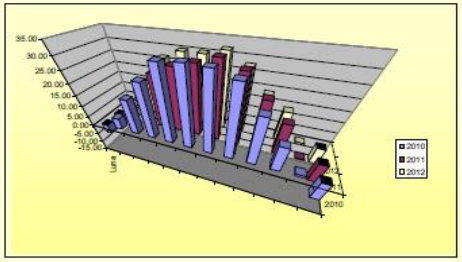


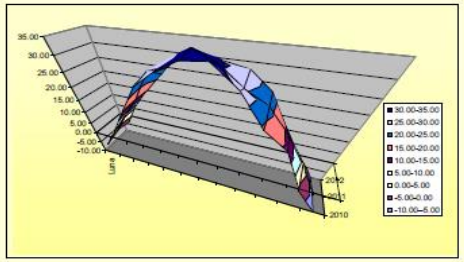
Data Visualization

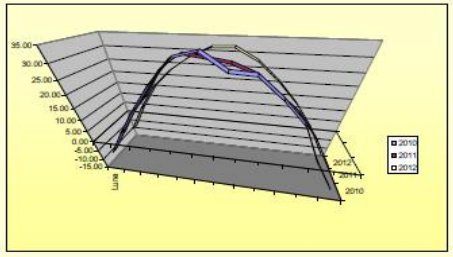
Data Visualization

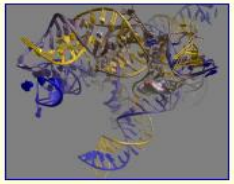
Visual technologies make decision support application more attractive and understandable to users.

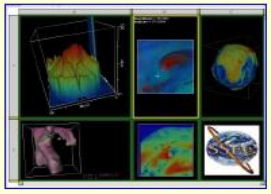
Data visualization refers to technologies that support visualization and interpretation of data and information. It includes digital images, GIS, Graphical user interface, graphs, virtual reality, dimensional presentations, videos and animation. Visual tools can help idenftify relationships such as trends.







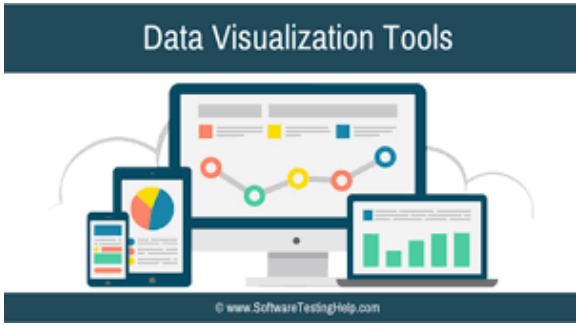




Lumina DECISION SYSTEMS *Bringing clarity to difficult decisions*

Commonly used Charts for Data Visualization

- Bar Chart
- Line Chart
- Scatterplot
- Sparkline
- Pie Chart
- Gauge
- Waterfall Chart
- Funnel Chart
- Heat Map
- Histogram
- Box Plot
- Maps
- Tables
- Indicators
- Area Chart
- Radar or Spider Chart
- Tree Map



Multivariate Analysis

Statistical Assumptions

All statistical procedures have underlying assumptions, some more stringent than others. In some cases, violation of these assumptions will not change substantive research conclusions. In other cases, violation of assumptions will undermine meaningful research. Establishing that one's data meet the assumptions of the procedure one is using, is an expected component of all quantitatively-based journal articles, theses, and dissertations.

Statistical Assumptions

- Data Related Assumptions
- Model Related Assumptions

Data Related Statistical Assumptions

- SOUND MEASUREMENT
- All forms of statistical analysis assume sound measurement, relatively free of coding errors. It is good practice to run descriptive statistics on one's data so that one is confident that data are generally as expected in terms of means and standard deviations, and there are no out-of-bounds entries beyond the expected range.
- Avoiding Attenuation: When the range of the data is reduced artificially, as by classifying or dichotomizing a continuous variable, correlation is attenuated, often leading to underestimation of the effect size of that variables
- Sample Size Adequacy
- Randomness
- Descriptive Statistics: Central Tendency, Dispersion, Skewness and Kurtosis

Model Related Statistical Assumptions for Multivariate Analysis

- Normality of Variables
- Zero Expected Residuals
- Normality of Residuals
- No Multicollinearity
- Homoscedasticity
- No Autocorrelation
- Linearity
- Model Specification (Inclusion and Omission of Variables)
- Model Identification (Appropriate Functional Form for the Model)

Tests of Normality

- There are several methods of assessing whether the data are normally distributed or not. They fall under two categories:
- GRAPHICAL METHODS
  - Histogram (SPSS)
  - BOX PLOT (SPSS)
  - Q-Q Probability Plots (SPSS)
  - P-P (Cumulative Probability (or Frequency) Plots (SPSS)
- STATISTICAL TESTS
  - W/S Test (SPSS)
  - Jarque Bera Test (EVIEWS)
  - Shapiro-Wilk's Test (SPSS)
  - Kolmogorov-Smirnov Test (SPSS)
  - D' Agostino's Test (NCSS)
  - Chi-Square Test (SPSS)
  - Anderson Darling Test (MINITAB) (EXCEL AD Calculator)

Testing for Homogeneity of variances/Homoscedasticity

- Brown & Forsythe's Test (SPSS)
- Welch Test (SPSS, SAS, MINITAB)
- Bartlett's Test of Sphericity (for Multivariate Normality) (SPSS)
- F-max test
- Box M test (SPSS)
- ANCOVA & MANCOVA (SPSS)
- Parallelism Tests (Test of Proportional Odds) (Running Ordinal Regression in SPSS)
- Leven's Test (SPSS)

Tests of Linearity

- Graphical Method (SPSS)
- Curve Fitting with R-Squared Difference Test (SPSS)
- ANOVA (SPSS)
- Eta the Correlation Ratio (SPSS)
- Adding Non-Linear Terms to the Model
- Ramsey's Reset Test (STATA)

Testing for Multicollinearity

- Tolerance (SPSS)
- Variance Inflation Factor (SPSS)
- Condition Indices (SPSS)
- Multicollinearity in Structural Equations Modelling (SEM)
- Intra Class Correlation

Tests for Autocorrelation

- Graphical Method (SPSS)
- Runs Test (SPSS)
- Durbin Watson d-test (SPSS)
- Breusch- Godfrey Test (Lagrange's Multiplier Test) (EXCEL)
- Akaike-Schwartz Information Criteria (EVIEWS)
- Autocorrelation Function (ACF) & Partial Autocorrelation Function (PACF) (EVIEWS)

Guidelines for Multivariate Analysis and Interpretation

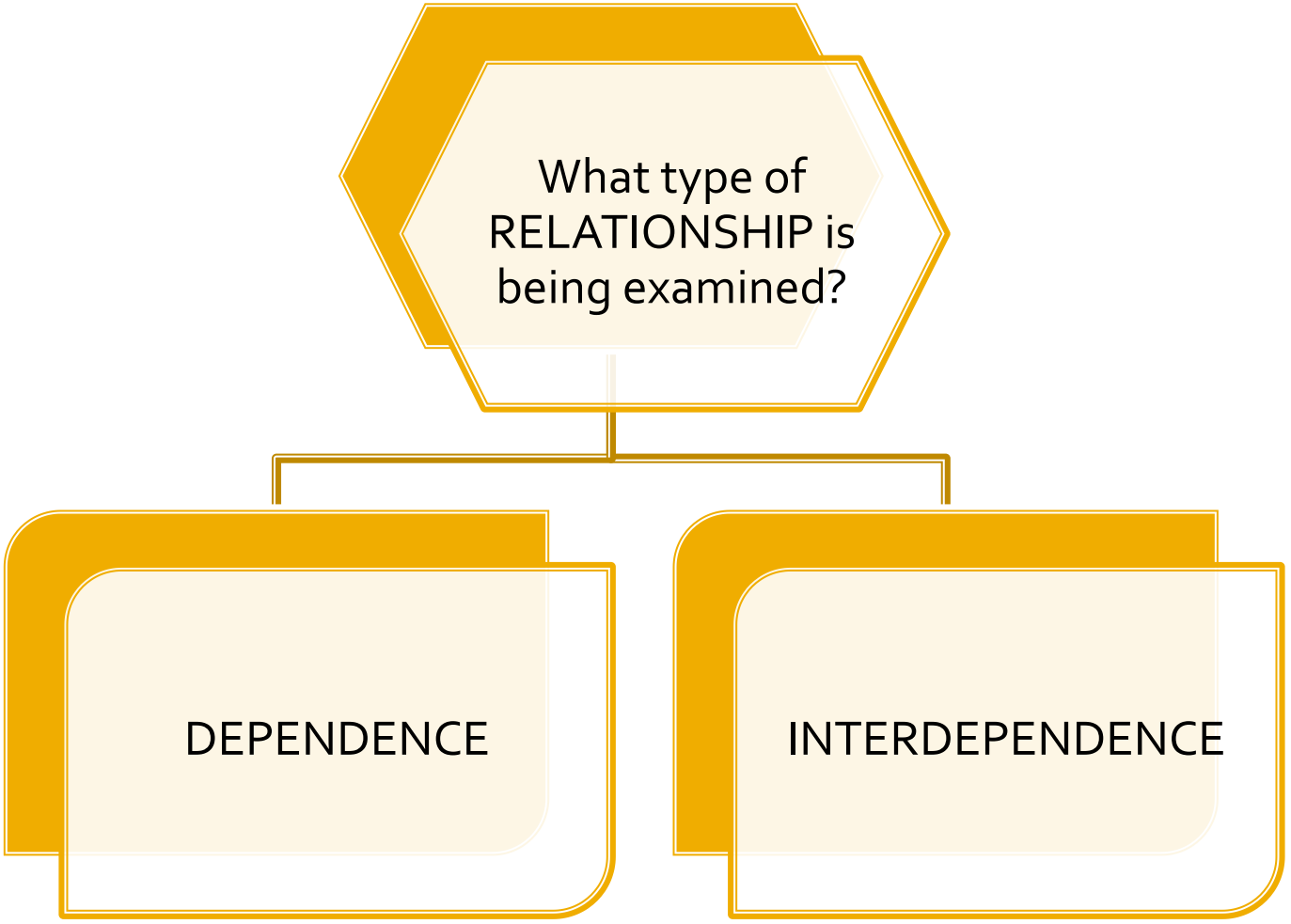
- **Establish practical significance as well as statistical significance**
- **Recognize that sample size affects all results**
- **Know your data**
  - Influence of outliers
  - Violation of assumptions
  - Missing data
- **Strive for model parsimony**
  - Specification error
  - Multicollinearity
- **Look at your errors**
- **Validate your results**
  - Splitting samples
  - Gathering a separate sample
    - **Bootstrapping**
    - (Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples. This process allows you to calculate standard errors, construct confidence intervals, and perform hypothesis testing for numerous types of sample statistics. Bootstrap methods are alternative approaches to traditional hypothesis testing and are notable for being easier to understand and valid for more conditions.)

A CLASSIFICATION OF MULTIVARIATE TECHNIQUES

- The classification is based on three judgments the researcher must make about the research objective and nature of the data.

1) Can the variables be divided into dependent and independent classification based on some theory?
2) If they can, how many variables are treated as dependent in a single analysis?
3) How are the variables, both dependent and independent measured?

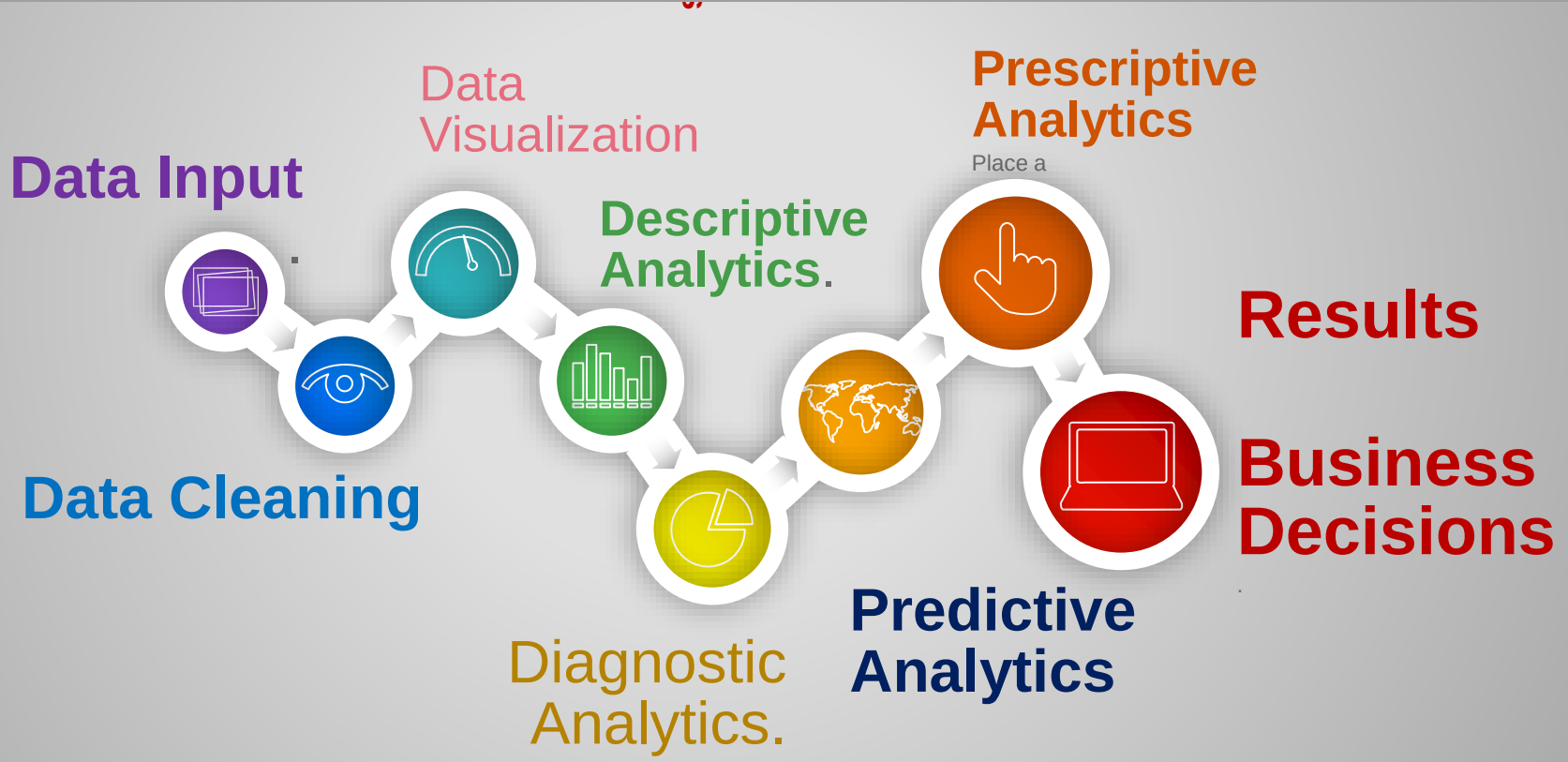
A CLASSIFICATION OF MULTIVARIATE TECHNIQUES



Nomenclature of Variables

Dependent Variable	Independent Variable
Predictand	Predictor
Regressand	Regressor
Explained Variable	Explanatory Variable
Outcome Variable	Covariate
Endogeneous Variable	Exogeneous Variable
Response Variable	Stimulus Variable
Controlled Variable	Control or Controlling Variable

Modern Approach to Classifying Multivariate Techniques: Data Analytics and Predictive Modelling:



Analysis vs. Analytics

- **Analysis** refers to the process of segmenting your problem into easily digestible chunks that you can study individually and examine how they relate to each other.
- **Analytics,** on the other hand, Analytics is the application of logical and computational reasoning to the component parts obtained in an analysis. And in doing that one is looking for patterns and often exploring what she could do with them in the future.
- So instead of *Business* and *Data*, we should better use *Business Analytics* and *Data Analytics*.

Time Before going any further, let's introduce a timeline as it turns out to be crucial for subsequent segmentation. We will employ three states – past, present, and future.

There will be a line that crosses the diagram indicating the present moment for any analytics problem. Everything on the left will refer to analytics looking **backward**, to the past that. All that is on the right will refer to **predictive** analytics.

Analysis vs. Analytics

- The word “Analytics” has come into the foreground in last decade or so. The proliferation of the internet and information technology has made analytics very relevant in the current age.

- Analytics is a field which combines data, information technology, statistical analysis, quantitative methods and computer-based models into one. This all are combined to provide decision makers all the possible scenarios to make a well thought and researched decision. The computer-based model ensures that decision makers are able to see performance of decision under various scenarios.

- As a decision-making paradigm, business data analytics is a means for informed decision-making. Through this lens, business data analytics is considered the tool of making decisions using evidence-based problem identification and problem solving.

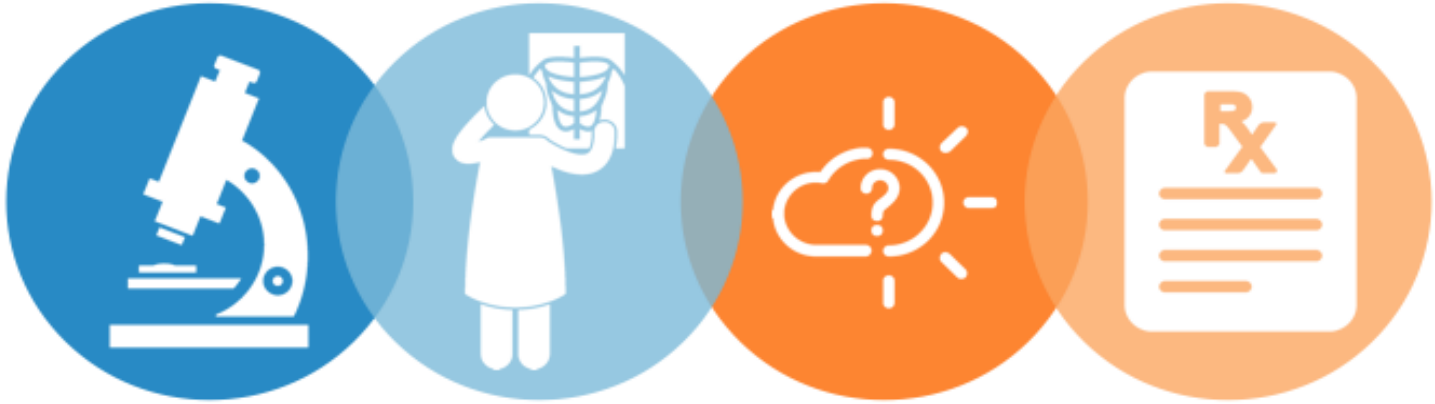
Types of Data Analytics

- ***Business Analytics for Competitive Advantage -> Develops Data-Driven Approaches to Managerial Decisions using these key Analytics capabilities***
  - Descriptive Analytics: What happened?
  - Diagnostic Analytics: Why did it happen?
  - Predictive Analytics: What is likely to happen?
  - Prescriptive Analytics: What should I do about it?
- Today, businesses whether big or small, apply a broad-based business analytics driven culture to achieve excellence and growth, make informed and optimised decisions, improve outcomes and manage risk which sets them apart.

Modern Approach to Classifying Multivariate Techniques: Types of Data Analytics

- Before diving deeper into each of these, let's define the four types of analytics:
- 1) **Descriptive Analytics**: Describing or summarizing the existing data using existing business intelligence tools to better understand what is going on or what has happened.
- 2) **Diagnostic Analytics**: Focus on past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.
- 3) **Predictive Analytics**: Emphasizes on predicting the possible outcome using statistical models and machine learning techniques.
- 4) **Prescriptive Analytics**: It is a type of predictive analytics that is used to recommend one or more course of action on analyzing the data.

Types of Data Analytics



Descriptive
Explains what happened.

Diagnostic
Explains why it happened.

Predictive
Forecasts what might happen.

Prescriptive
Recommends an action based on the forecast.

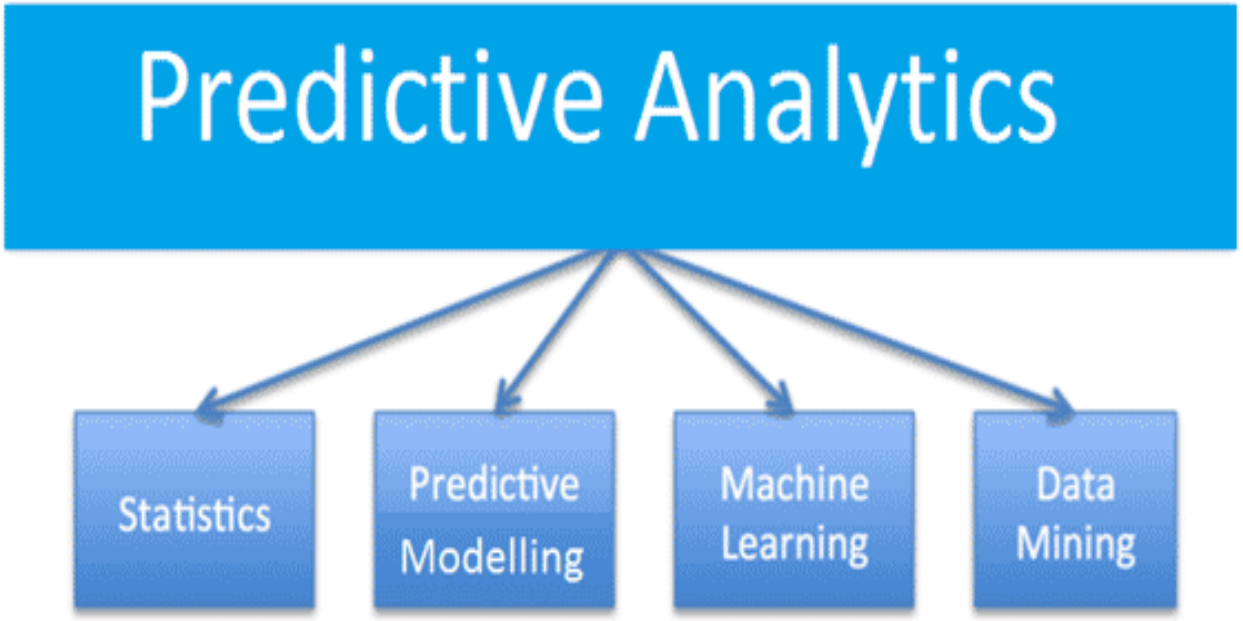
Predictive analytics vs. prescriptive analytics

- Predictive analytics: estimation
- Prescriptive analytics: estimation + optimization

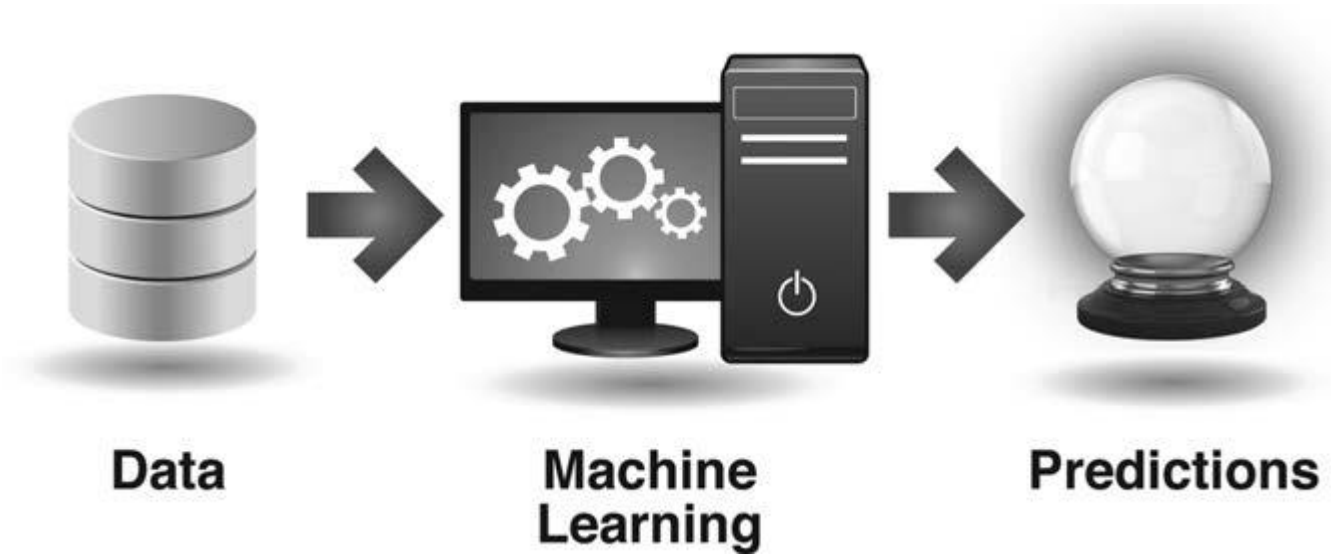
Types of Data Analytics: Comprehensive Approach

Text Analytics	method to discover a pattern in large data sets using databases or data mining tools.
Statistical Analytics	shows "What happened?" by using past data in the form of dashboards. Descriptive Analysis and Inferential Analysis.
Descriptive Analytics	• analyses complete data or a sample of summarized numerical data.
Inferential Analytics	find different conclusions from the same data by selecting different samples.
Diagnostic Analytics	• shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis.
Predictive Analytics	• shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis.
Prescriptive Analytics	combines the insight from all previous Analysis to determine which action to take in a current problem or decision.

Applications of Predictive Analytics?



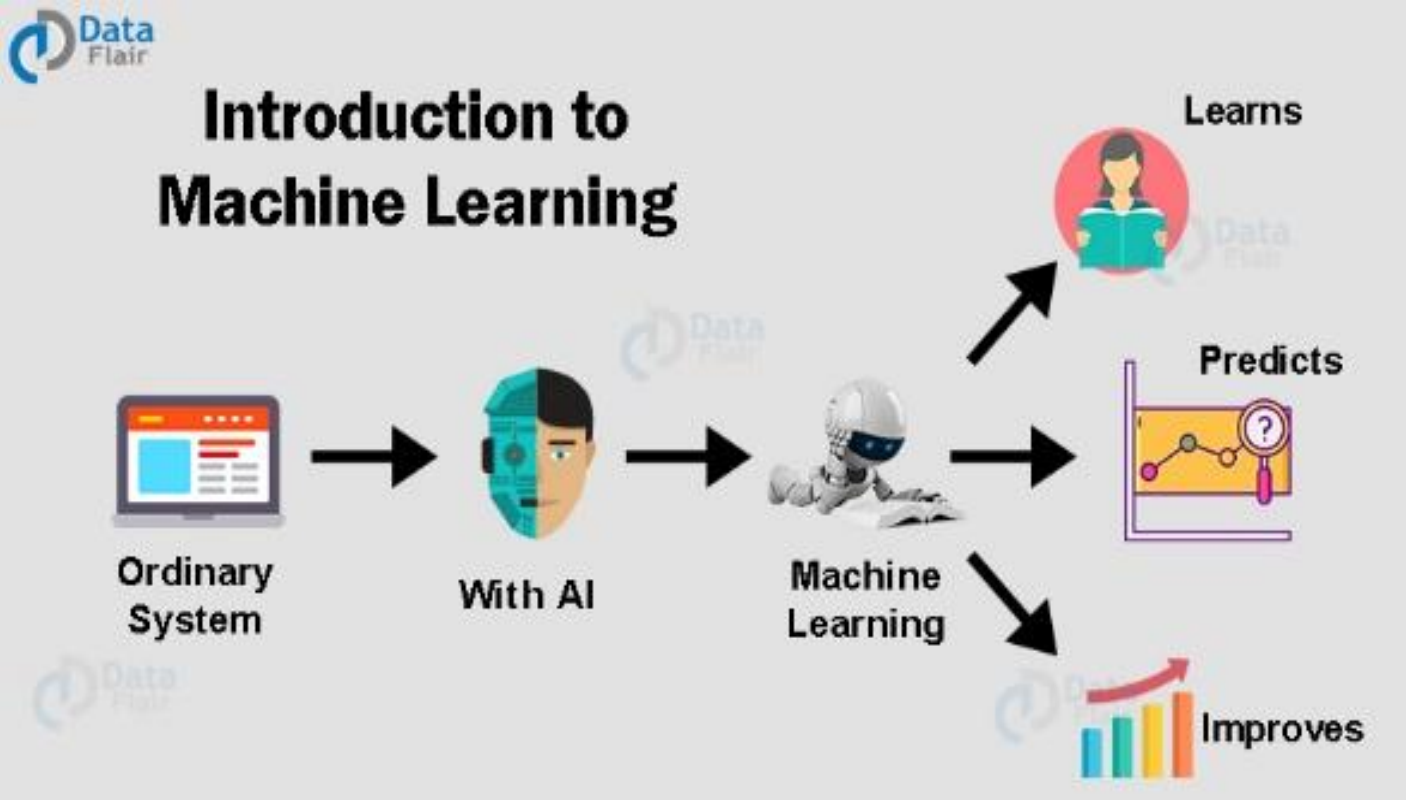
Machine Learning



Machine Learning

- **Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can access data and use it learn for themselves.
- The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.
- Machine learning can seem daunting to beginners. To put it simply, machine learning is the idea that computers can learn from examples and past experience. Machine learning beginners should realize that math plays an important part in helping machines to understand and learn. Beginning machine learning engineers should have a grasp of linear algebra, statistics, calculus and complex algorithms.

Machine Learning



Machine Learning

Machine Learning

- **TYPES OF MACHINE LEARNING**
- There are three kinds of machine learning algorithms:
- a. Supervised learning
- b. Unsupervised learning
- c. Reinforcement learning

Machine Learning

- **SUPERVISED LEARNING**
- Much of practical machine learning uses supervised learning.
- In this type, the system tries to learn from the previous examples its given. (On the other hand, in unsupervised learning the system attempts to find the patterns directly from the example given.)
- Speaking mathematically, Supervised Learning is when you have both input variables (*x*) and output variables (*y*) and can use an algorithm to derive the mapping function from the input to the output.
- Supervised learning problems can be further divided into two parts, namely classification and regression:
- **Classification**: A classification problem is when the output variable is a category or a group, such as “black” or “white,” or “spam” and “no spam."
- **Regression**: A regression problem is when the output variable is a real value, such as “Rupees” or “height.”

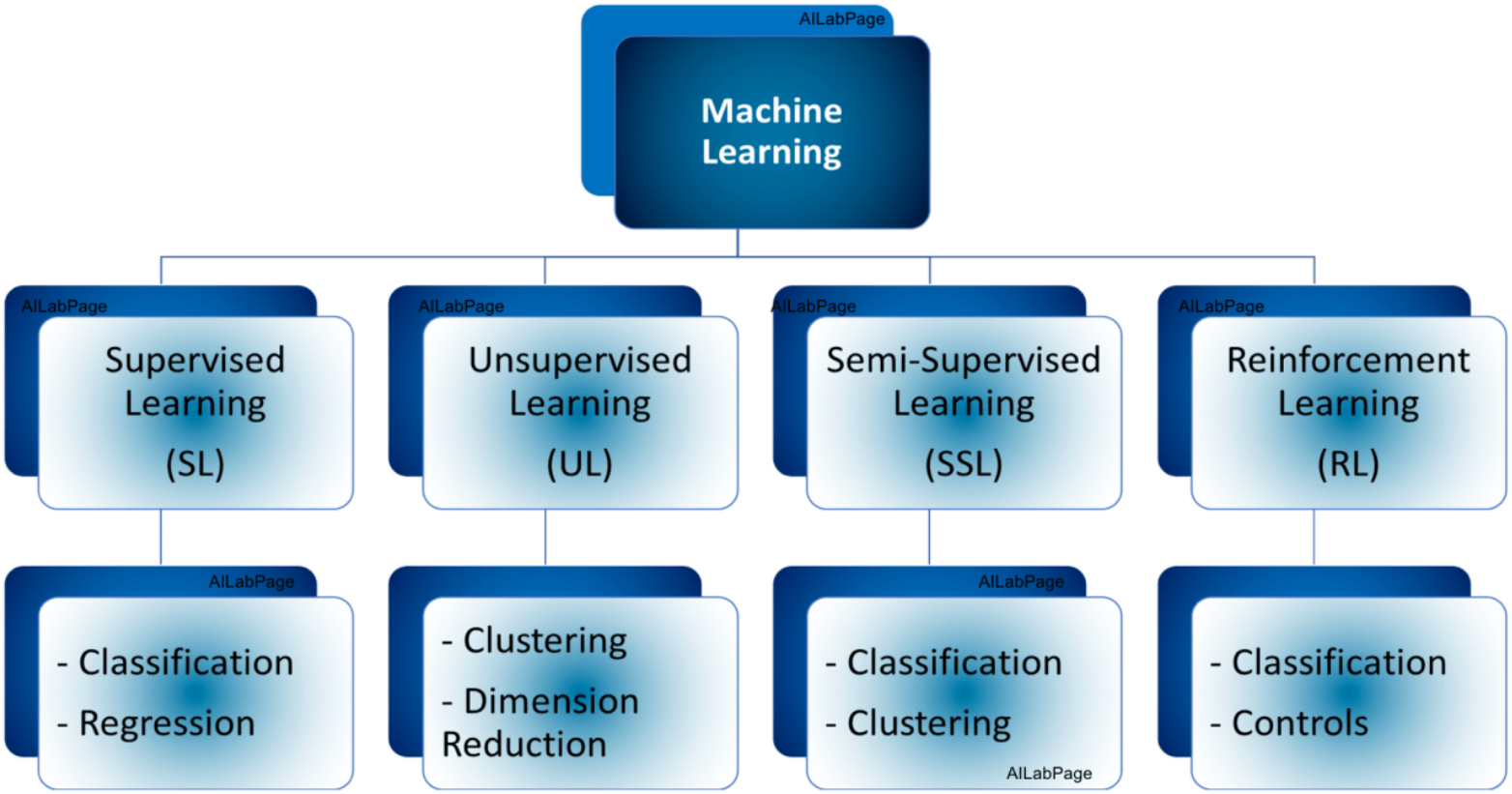
Machine Learning

- **UNSUPERVISED LEARNING**
- In unsupervised learning, the algorithms are left to themselves to discover interesting structures in the data.
- Mathematically, Unsupervised Learning is when you have only input data (*x*) and no corresponding output variables.
- This is called unsupervised learning because unlike supervised learning there are no given correct answers and the machine itself finds the solutions.
- Unsupervised learning problems can be further divided into association and clustering:
- **Association**: An association rule learning problem is the need to discover rules that describe large portions of data, like “people who buy X also tend to buy Y."
- **Clustering**: A clustering problem is the need to discover inherent groupings in the data, such as grouping customers by purchasing behavior.

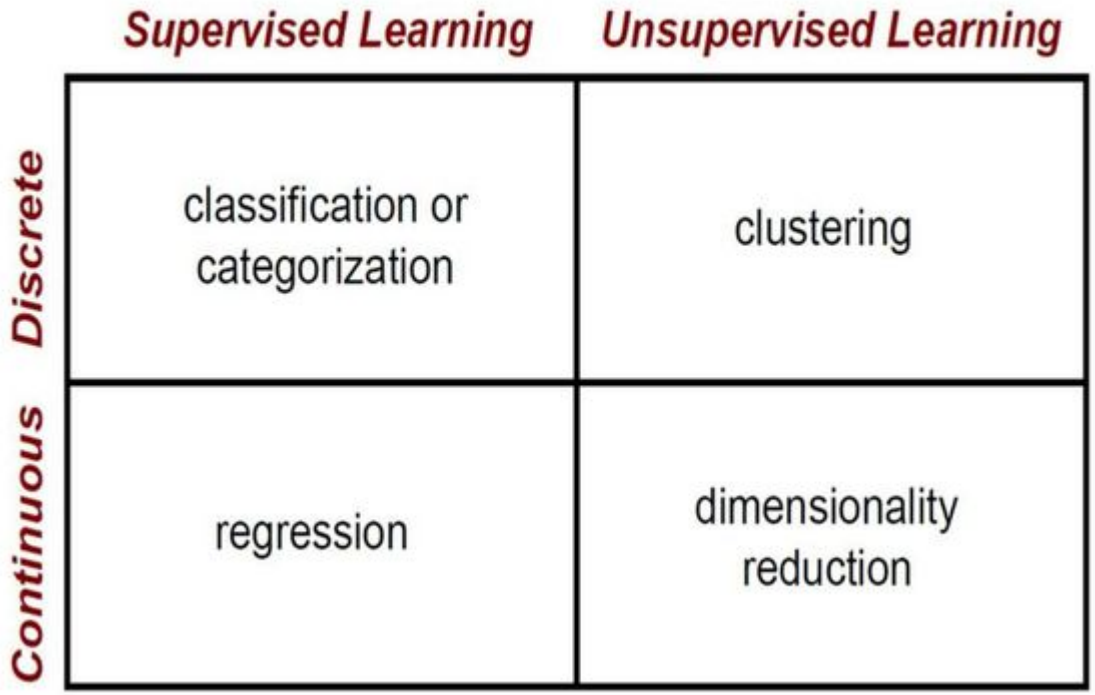
Machine Learning

- **REINFORCEMENT LEARNING**
- This is a particular type of machine learning where the computer program will interact with a dynamic environment in which it must perform a particular goal, such as playing a game with an opponent or driving a car. The program is provided feedback in the form of rewards and punishments as it navigates its problem space.
- Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it continuously trains itself using trial and error method.

Machine Learning



Machine Learning



Data Mining

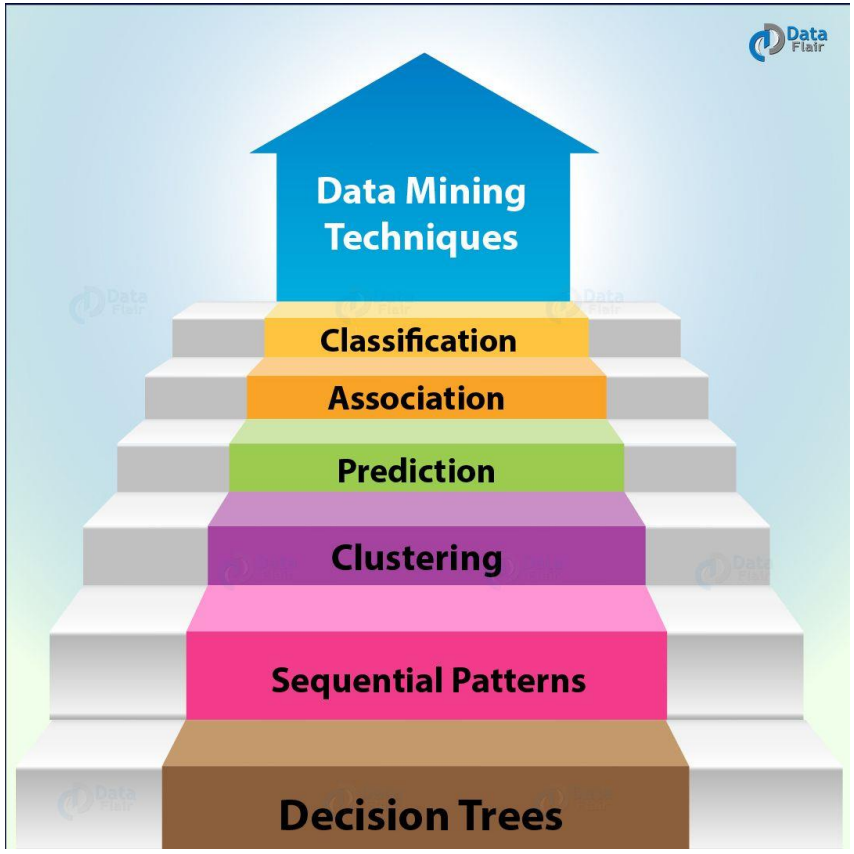
What is Data Mining ?

- process of analyzing data from different perspectives
- summarizing it into useful information
- information that can be used to increase revenue, cuts costs, or both.

Data Mining

- Data mining is a collective term used to describe different analysis techniques such as statistics, artificial intelligence and machine learning that are employed to scan huge amounts of data found in the organization's databases or online databases.
- In data mining, by identifying possible patterns, it could then help to predict, associate and group events, products, or customers in a more effective manner so that the organizations could provide better products or services to the customers or improve their operations.

Data Mining Techniques



Top 5 Predictive Analytics Models

- **Classification Model**
- The classification model is, in some ways, the simplest of the several types of predictive analytics models we're going to cover. It puts data in categories based on what it learns from historical data.
- Classification models are best to answer yes or no questions, providing broad analysis that's helpful for guiding decisive action. These models can answer questions such as:
- For a retailer, "Is this customer about to churn?"
- For a loan provider, "Will this loan be approved?" or "Is this applicant likely to default?"
- For an online banking provider, "Is this a fraudulent transaction?"
- The breadth of possibilities with the classification model—and the ease by which it can be retrained with new data—means it can be applied to many different industries.

Top 5 Predictive Analytics Models

- **Clustering Model**
- The clustering model sorts data into separate, nested smart groups based on similar attributes. If an ecommerce shoe company is looking to implement targeted marketing campaigns for their customers, they could go through the hundreds of thousands of records to create a tailored strategy for each individual. But is this the most efficient use of time? Probably not. Using the clustering model, they can quickly separate customers into similar groups based on common characteristics and devise strategies for each group at a larger scale.
- Other use cases of this predictive modeling technique might include grouping loan applicants into “smart buckets” based on loan attributes, identifying areas in a city with a high volume of crime, and benchmarking SaaS customer data into groups to identify global patterns of use.

Top 5 Predictive Analytics Models

- **Forecast Model**
- One of the most widely used predictive analytics models, the forecast model deals in metric value prediction, estimating numeric value for new data based on learnings from historical data.
- This model can be applied wherever historical numerical data is available. Scenarios include:
- A SaaS company can estimate how many customers they are likely to convert within a given week.
- A call center can predict how many support calls they will receive per hour.
- A shoe store can calculate how much inventory they should keep on hand in order to meet demand during a particular sales period.
- The forecast model also considers multiple input parameters. If a restaurant owner wants to predict the number of customers she is likely to receive in the following week, the model will take into account factors that could impact this, such as: Is there an event close by? What is the weather forecast? Is there an illness going around?

Top 5 Predictive Analytics Models

- **Outliers Model**
- The outliers model is oriented around anomalous data entries within a dataset. It can identify anomalous figures either by themselves or in conjunction with other numbers and categories.
- Recording a spike in support calls, which could indicate a product failure that might lead to a recall
- Finding anomalous data within transactions, or in insurance claims, to identify fraud
- Finding unusual information in your NetOps logs and noticing the signs of impending unplanned downtime
- The outlier model is particularly useful for predictive analytics in retail and finance. For example, when identifying fraudulent transactions, the model can assess not only amount, but also location, time, purchase history and the nature of a purchase (i.e., a $1000 purchase on electronics is not as likely to be fraudulent as a purchase of the same amount on books or common utilities).

Top 5 Predictive Analytics Models

- **Time Series Model**
- The time series model comprises a sequence of data points captured, using time as the input parameter. It uses the last year of data to develop a numerical metric and predicts the next three to six weeks of data using that metric. Use cases for this model includes the number of daily calls received in the past three months, sales for the past 20 quarters, or the number of patients who showed up at a given hospital in the past six weeks. It is a potent means of understanding the way a singular metric is developing over time with a level of accuracy beyond simple averages. It also takes into account seasons of the year or events that could impact the metric.
- If the owner of a salon wishes to predict how many people are likely to visit his business, he might turn to the crude method of averaging the total number of visitors over the past 90 days. However, growth is not always static or linear, and the time series model can better model exponential growth and better align the model to a company's trend. It can also forecast for multiple projects or multiple regions at the same time instead of just one at a time.

Interdependence Techniques

Interdependence Techniques

- Factor Analysis (Data Reduction Technique) (Metric Variables)
- Cluster Analysis (Respondents (Cases) Grouping Technique) (Metric variables)
- Multidimensional Scaling (Perceptual Mapping) (Metric /Non-Metric)
- Correspondence Analysis (Perceptual Mapping) (Non-Metric Attributes)

Factor Analysis

- When there are many variables in a research design, it is often helpful to reduce the variables to a smaller set of factors.
- This is an independence technique, in which there is no dependent variable. Rather, the researcher is looking for the underlying structure of the data matrix.
- Ideally, the independent variables are normal and continuous, with at least three to five variables loading onto a factor.
- The sample size should be over 50 observations, with over five observations per variable.
- Multicollinearity is generally preferred between the variables, as the correlations are key to data reduction.
- Kaiser's Measure of Statistical Adequacy (MSA) is a measure of the degree to which every variable can be predicted by all other variables. An overall MSA of .80 or higher is very good, with a measure of under .50 deemed poor.

Factor Analysis

- There are two main factor analysis methods: common factor analysis, which extracts factors based on the variance shared by the factors, and principal component analysis, which extracts factors based on the total variance of the factors.
- Common factor analysis is used to look for the latent (underlying) factors, whereas principal component analysis is used to find the fewest number of variables that explain the most variance.
- The first factor extracted explains the most variance.
- Typically, factors are extracted as long as the eigenvalues are greater than 1.0 or the Scree test visually indicates how many factors to extract.
- The factor loadings are the correlations between the factor and the variables.
- Typically a factor loading of .4 or higher is required to attribute a specific variable to a factor.
- An orthogonal rotation assumes no correlation between the factors, whereas an oblique rotation is used when some relationship is believed to exist.

Cluster Analysis

- The purpose of cluster analysis is to reduce a large data set to meaningful subgroups of individuals or objects.
- The division is accomplished on the basis of similarity of the objects across a set of specified characteristics.
- Outliers are a problem with this technique, often caused by too many irrelevant variables.
- The sample should be representative of the population, and it is desirable to have uncorrelated factors.
- There are three main clustering methods: hierarchical, which is a treelike process appropriate for smaller data sets; non-hierarchical, which requires specification of the number of clusters a priori; and a combination of both.
- There are four main rules for developing clusters: the clusters should be different, they should be reachable, they should be measurable, and the clusters should be profitable (big enough to matter).
- This is a great tool for market segmentation.

Multidimensional Scaling (MDS)

- The purpose of MDS is to transform consumer judgments of similarity into distances represented in multidimensional space.
- This is a decomposition approach that uses perceptual mapping to present the dimensions.
- As an exploratory technique, it is useful in examining unrecognized dimensions about products and in uncovering comparative evaluations of products when the basis for comparison is unknown.
- Typically there must be at least four times as many objects being evaluated as dimensions. It is possible to evaluate the objects with nonmetric preference rankings or metric similarities (paired comparison) ratings.
- Kruskal's Stress measure is a "badness of fit" measure; a stress percentage of 0 indicates a perfect fit, and over 20% is a poor fit.
- The dimensions can be interpreted either subjectively by letting the respondents identify the dimensions or objectively by the researcher.

Correspondence Analysis

- This technique provides for dimensional reduction of object ratings on a set of attributes, resulting in a perceptual map of the ratings.
- However, unlike MDS, both independent variables and dependent variables are examined at the same time.
- This technique is more similar in nature to factor analysis.
- It is a compositional technique, and is useful when there are many attributes and many companies.
- It is most often used in assessing the effectiveness of advertising campaigns.
- It is also used when the attributes are too similar for factor analysis to be meaningful.
- The main structural approach is the development of a contingency (crosstab) table. This means that the form of the variables should be nonmetric.
- The model can be assessed by examining the Chi-square value for the model.
- Correspondence analysis is difficult to interpret, as the dimensions are a combination of independent and dependent variables.

Multivariate Analysis

Dependence Techniques

Dependence Techniques

- Multiple Regression Analysis (Prediction of Scaled Dependent Variable) (DV: Metric vs. IV: Metric/Non-Metric)
- Discriminant Analysis (Classification Technique)
- (DV: Non metric (Categorical) vs.
- Logistic Regression Analysis (Classification Technique)
- Conjoint Analysis
- Linear Probability Models
- Canonical Correlation
- MANOVA and MANCOVA
- Structural Equations Modeling

Multiple Regression Analysis

- Multiple regression is the most commonly utilized multivariate technique.
- It examines the relationship between a single metric dependent variable and two or more metric independent variables.
- The technique relies upon determining the linear relationship with the lowest sum of squared variances; therefore, assumptions of normality, linearity, and equal variance are carefully observed.
- The beta coefficients (weights) are the marginal impacts of each variable, and the size of the weight can be interpreted directly. Multiple regression is often used as a forecasting tool.

Discriminant Analysis

- The purpose of discriminant analysis is to correctly classify observations or people into homogeneous groups.
- The independent variables must be metric and must have a high degree of normality.
- Discriminant analysis builds a linear discriminant function, which can then be used to classify the observations.
- The overall fit is assessed by looking at the degree to which the group means differ (Wilk's Lambda or D2) and how well the model classifies.
- To determine which variables have the most impact on the discriminant function, it is possible to look at partial F values.
- The higher the partial F, the more impact that variable has on the discriminant function.
- This tool helps categorize people, like buyers and nonbuyers.

Linear Probability Models: Logistic Regression Analysis

- Sometimes referred to as "choice models," this technique is a variation of multiple regression that allows for the prediction of an event.
- It is allowable to utilize nonmetric (typically binary) dependent variables, as the objective is to arrive at a probabilistic assessment of a binary choice.
- The independent variables can be either discrete or continuous.
- A contingency table is produced, which shows the classification of observations as to whether the observed and predicted events match.
- The sum of events that were predicted to occur which actually did occur and the events that were predicted not to occur which actually did not occur, divided by the total number of events, is a measure of the effectiveness of the model.
- This tool helps predict the choices consumers might make when presented with alternatives.

Linear Probability Models: Logistic Regression

- Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy).
- The prediction is based on the use of one or several predictors (numerical and categorical).
- A logistic regression produces a logistic curve, which is limited to values between 0 and 1.
- Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability.
- Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

Linear Probability Models: Probit Model

- A **probit model** is a type of regression where the dependent variable can take only two values, for example married or not married. The word is a portmanteau, coming from ***prob****ability* + *un****it***.[1] The purpose of the model is to estimate the probability that an observation with particular characteristics will fall into a specific one of the categories; moreover, classifying observations based on their predicted probabilities is a type of binary classification model.
- A probit model is a popular specification for an ordinal[2] or a binary response model. As such it treats the same set of problems as does logistic regression using similar techniques. The probit model, which employs a probit link function, is most often estimated using the standard maximum likelihood procedure, such an estimation being called a **probit regression**.

Linear Probability Models: Probit Model

- Suppose a response variable *Y* is *binary*, that is it can have only two possible outcomes which we will denote as 1 and 0. For example, *Y* may represent presence/absence of a certain condition, success/failure of some device, answer yes/no on a survey, etc. We also have a vector of regressors *X*, which are assumed to influence the outcome *Y*. Specifically, we assume that the model takes the form
- $P(Y=1|X)=\Phi(X'\beta)$
- where Pr denotes probability, and Φ is the Cumulative Distribution Function (CDF) of the standard normal distribution. The parameters *β* are typically estimated by maximum likelihood.

Linear Probability Models: Tobit Model

- The **Tobit model** is a statistical model proposed by James Tobin (1958)[1] to describe the relationship between a non-negative dependent variable y_i and an independent variable (or vector)x_i. The term *Tobit* was derived from Tobin's name by truncating and adding *-it* by analogy with the probit model. The Tobit model is distinct from the truncated regression model, which is in general different and requires a different estimator.

Linear Probability Models: Tobit Model

The Tobit Model

- Can also have latent variable models that don't involve binary dependent variables
- Say $y^* = \boldsymbol{x}\boldsymbol{\beta} + u$, $u|\boldsymbol{x} \sim \text{Normal}(0,\sigma^2)$
- But we only observe $y = \max(0, y^*)$
- The Tobit model uses MLE to estimate both $\boldsymbol{\beta}$ and σ for this model
- Important to realize that $\boldsymbol{\beta}$ estimates the effect of $\boldsymbol{x}$ on y*, the latent variable, not y

Multivariate Analysis of Variance (MANOVA)

- This technique examines the relationship between several categorical independent variables and two or more metric dependent variables.
- Whereas analysis of variance (ANOVA) assesses the differences between groups (by using T tests for two means and F tests between three or more means), MANOVA examines the dependence relationship between a set of dependent measures across a set of groups.
- Typically this analysis is used in experimental design, and usually a hypothesized relationship between dependent measures is used.
- This technique is slightly different in that the independent variables are categorical and the dependent variable is metric.
- Sample size is an issue, with 15-20 observations needed per cell. However, too many observations per cell (over 30) and the technique loses its practical significance.
- Cell sizes should be roughly equal, with the largest cell having less than 1.5 times the observations of the smallest cell. That is because, in this technique, normality of the dependent variables is important.
- The model fit is determined by examining mean vector equivalents across groups. If there is a significant difference in the means, the null hypothesis can be rejected and treatment differences can be determined.

Conjoint Analysis

- Conjoint analysis is often referred to as "trade-off analysis," since it allows for the evaluation of objects and the various levels of the attributes to be examined.
- It is both a compositional technique and a dependence technique, in that a level of preference for a combination of attributes and levels is developed.
- A part-worth, or utility, is calculated for each level of each attribute, and combinations of attributes at specific levels are summed to develop the overall preference for the attribute at each level.
- Models can be built that identify the ideal levels and combinations of attributes for products and services.

Canonical Correlation

- Canonical correlation analysis is used to identify and measure the associations among two sets of variables. Canonical correlation is appropriate in the same situations where multiple regression would be, but where are there are multiple intercorrelated outcome variables. Canonical correlation analysis determines a set of canonical variates, orthogonal linear combinations of the variables within each set that best explain the variability both within and between sets.

- The most flexible of the multivariate techniques, canonical correlation simultaneously correlates several independent variables and several dependent variables.
- This powerful technique utilizes metric independent variables, unlike MANOVA, such as sales, satisfaction levels, and usage levels.
- It can also utilize nonmetric categorical variables.
- This technique has the fewest restrictions of any of the multivariate techniques, so the results should be interpreted with caution due to the relaxed assumptions.
- Often, the dependent variables are related, and the independent variables are related, so finding a relationship is difficult without a technique like canonical correlation.

Structural Equation Modelling

- Unlike the other multivariate techniques discussed, structural equation modelling (SEM) examines multiple relationships between sets of variables simultaneously.
- This represents a family of techniques, including LISREL, latent variable analysis, and confirmatory factor analysis. SEM can incorporate latent variables, which either are not or cannot be measured directly into the analysis.
- For example, intelligence levels can only be inferred, with direct measurement of variables like test scores, level of education, grade point average, and other related measures.
- These tools are often used to evaluate many scaled attributes or to build summated scales.

Dependence Techniques

MVT	Dependent Variables	Independent Variables	Number of Variables
Multiple Regression	Metric	Metric/ Non-Metric	D.V.-One I.V. Many
Discriminant Analysis	Non- Metric (Categorical)	Metric	D.V.-One I.V. Many
Logistic Regression	Non-Metric: Binary	Metric/ Non metric	D.V.-One I.V. Many
Conjoint Analysis	Metric/ Non-Metric	Non-Metric	D.V.-One I.V. Many
Linear Probability Models	Non-Metric: Binary	Metric/ Non-Metric	D.V.-One I.V. Many
Canonical Correlation	Metric	Metric	D.V. –Many I.V. - Many
MANOVA/ MANCOVA	Metric	Non-Metric Categorical	D.V. –Many I.V. - Many
Structural Equations Modelling	Metric	Metric/ Categorical	D.V. –Many I.V. - Many

Recent Trends in Data Analytics
&
Multivariate Techniques

Trends in Data Analytics

- **Big Data Analysis**
  With an increasing emphasis on digitization in every aspect of life, datasets continue to expand at an unprecedented rate. This expansion is both an advantage and a disadvantage—more data means more potential insights, but the sheer volume can be overwhelming.

- **Artificial Intelligence**
  As AIs become smarter and able to teach themselves, AIs created by AIs are being developed and launched in industry verticals such as banking, financial services, insurance, retail, hospitality, engineering, manufacturing, and more.

- **Deep Learning**
  The next step up from machine learning, deep learning leverages the advantages of vast computing power to manage enormous data sets, identifying patterns and delivering predictive results that were formerly impossible.

Methods based on Machine Learning and Artificial Intelligence

- **1. Decision Trees:** Decision tree analysis is a graphical representation, similar to a tree-like structure in which the problems in decision making can be seen in the form of a flow chart, each with branches for alternative answers. Decision trees are a top-down approach type, with the first decision node at the top, based on the answer at first decision node it will be divided into branches, and it will continue until the tree arrives at a final decision. The branches which do not divide any more are known as leaves.
- **2. Neural Networks:** Neural networks are a set of algorithms, which are designed to mimic the human brain. It is also known as the "Network of Artificial neurons". The applications of neural network in data mining are very broad. They have a high acceptance ability for noisy data and high accuracy results. Based on the necessity many types of neural networks are currently being used, few of them are recurrent neural networks and convolutional neural networks. Convolutional neural networks are mostly used in Image processing, natural language processing, and recommender systems. Recurrent neural networks are mainly used for handwriting and speech recognition.

Common Modelling Methods

- Modeling Methods
- The most widely used predictive modeling methods are as below,

- 1. Simple linear regression: A statistical method to mention the relationship between two variables which are continuous.

- 2. Multiple linear regression: A statistical method to mention the relationship between more than two variables which are continuous.

- 3. Polynomial regression: A non-linear relationship between residuals versus a predictor will lead to a nonlinear relationship. This can be archived through a polynomial regression model.

Common Modelling Methods

- 4. Support vector regression: Support Vector Machine is another regression method, which characterizes the algorithm based on all key features. The Support Vector Regression (SVR) apply similar principles as the SVM for classification, with some minor differences.
- **5. Decision tree regression:** A tree-like structure is used in these decision tree models to build a classification or regression related algorithms. Here the decision tree is incrementally developed by sub-setting the given dataset into smaller chunks.
- **6. Naive Bayes:** In machine learning, they are simple probabilistic classifiers that are predicted by applying Bayes theorem alongside independent assumptions.

Conclusions

- Each of the multivariate techniques described above has a specific type of research question for which it is best suited.
- Each technique also has certain strengths and weaknesses that should be clearly understood by the analyst before attempting to interpret the results of the technique.
- Current statistical packages (SAS, SPSS, S-Plus, and others) make it increasingly easy to run a procedure, but the results can be disastrously misinterpreted without adequate care.

REFERENCES

- Hair, Anderson, Tatham, Black: "Multivariate Data Analysis" Pearson Education
- Johnson and Wichern: "Applied Multivariate Statistical Analysis" Pearson Education
- Morrison Donald : "Multivariate Statistical Methods" Mc. Graw Hill
- Rao C. R. : "Linear Statistical Inference and Its Applications" Wiley Eastern
- Sharma K. R. : "Business Research Methods" National Publishing House
- Zikmund R. A. : "Business Research Methods" PHI
- Nargudkar Rajendra: "Marketing Research: Text and Cases" Tata Mc. Graw Hill
- Damodar Gujarati: "Basic Econometrics" Tata Mc. Graw Hill "
- Hardle & Simar: "Applied Multivariate Statistical Analysis" Springer Publication
- Cooper & Schindler: " Business Research Method" Mc Graw Hill Education
- Deepak Chawla & Neena sondhi: Research Methodology: Concepts and Cases" Vikas publication

References

References. See Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, B, 26, 211-234; see also Maddala, G. S. (1977). *Econometrics*. New York: McGraw-Hill. (page 315-317); or Mason, R., L., R. F. Gunst, and J. L. Hess (1989). *Statistical design and analysis of experiments with applications to engineering and science.* New York: Wiley.

References

Bibliography

Bollinger, Christopher R. & Chandra , Amitabh (2005). Iatrogenic specification error: A cautionary tale of cleaning data. *Journal of Labor Economics*, 23(2), 235-257.

Boneau, C. A. (1960). The effect of violation of assumptions underlying the t-test. *Psychological Bulletin*, 57: 49-64.

Cohen, Jacob (1969). *Statistical power analysis for the behavioral sciences*. NY: Academic Press.

Hutcheson, Graeme and Nick Sofroniou (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Thousand Oaks, CA: Sage Publications. Chapter two covers data screening for assumptions under GLM.

Lipsey, Mark W. & Wilson, David B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Schumacker, Randall E. and Richard G. Lomax (2004). *A beginner's guide to structural equation modeling, Second edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Shapiro, S. S. & Wilk. M. B. (1965). An analysis of variance test for normality. *Biometrika*, 52(3), 591-599.

Steenkamp, J.E.M., & van Trijp, H.C.M. (1991). The use of LISREL in validating marketing constructs. International Journal of Research in Marketing, 8: 283-299.

Vasu, E. S. (1979). Non-normality in regression analysis: Monte Carlo investigation under the condition of multicollinearity. Working papers in methodology. Number 9. Institute for Research in Social Science, University of North Carolina. Chapel Hill, North Carolina, USA.

Thank You

Have a Good Day