

Multivariate Analysis

Dr. Hemal Pandya

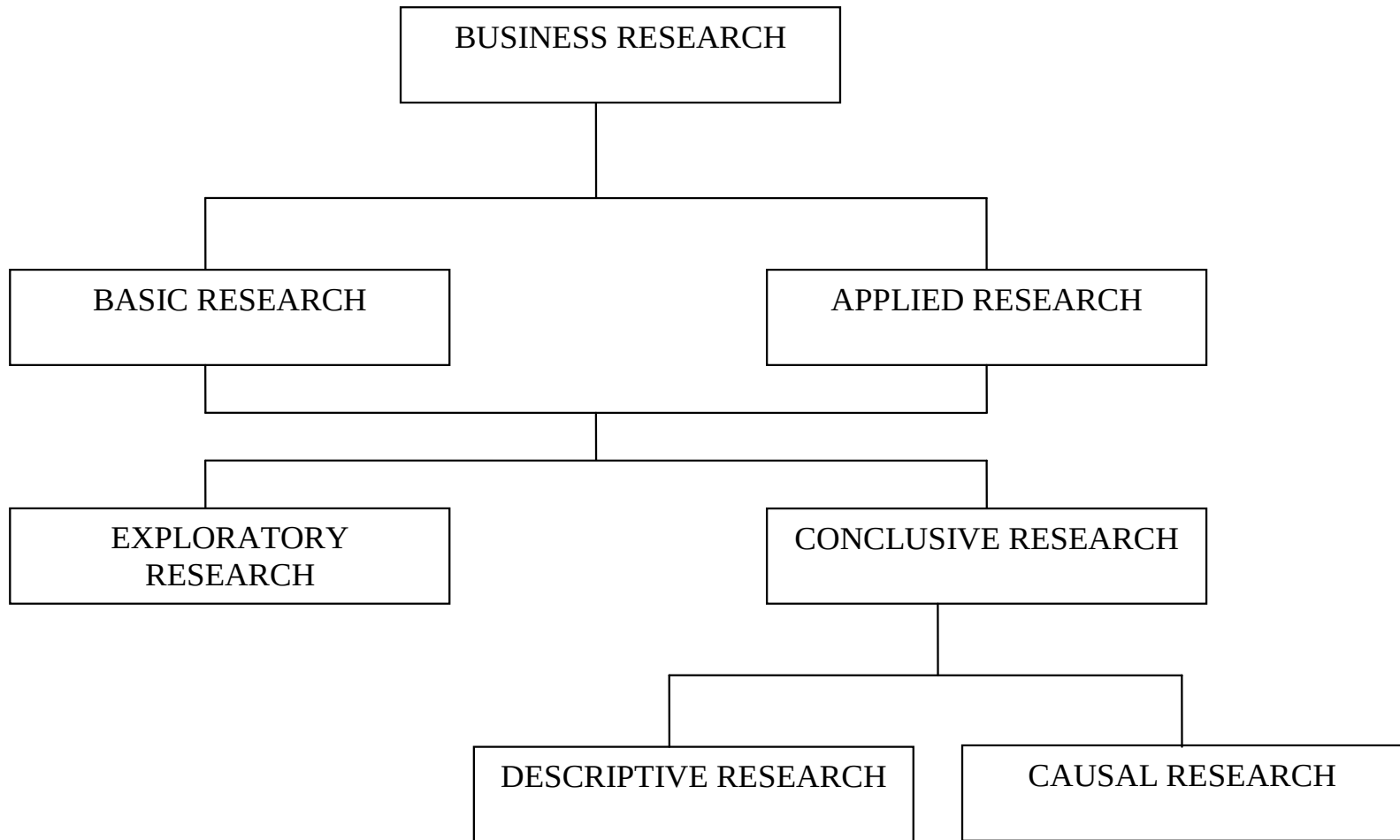
What is Research?

- Research in Social Sciences and 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

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

Types of Statistical Data Analysis

Five types of statistical analysis

Descriptive

What are the characteristics of the respondents?

Inferential

What are the characteristics of the population?

Differences

Are two or more groups the same or different?

Associative

Are two or more variables related in a systematic way?

Predictive

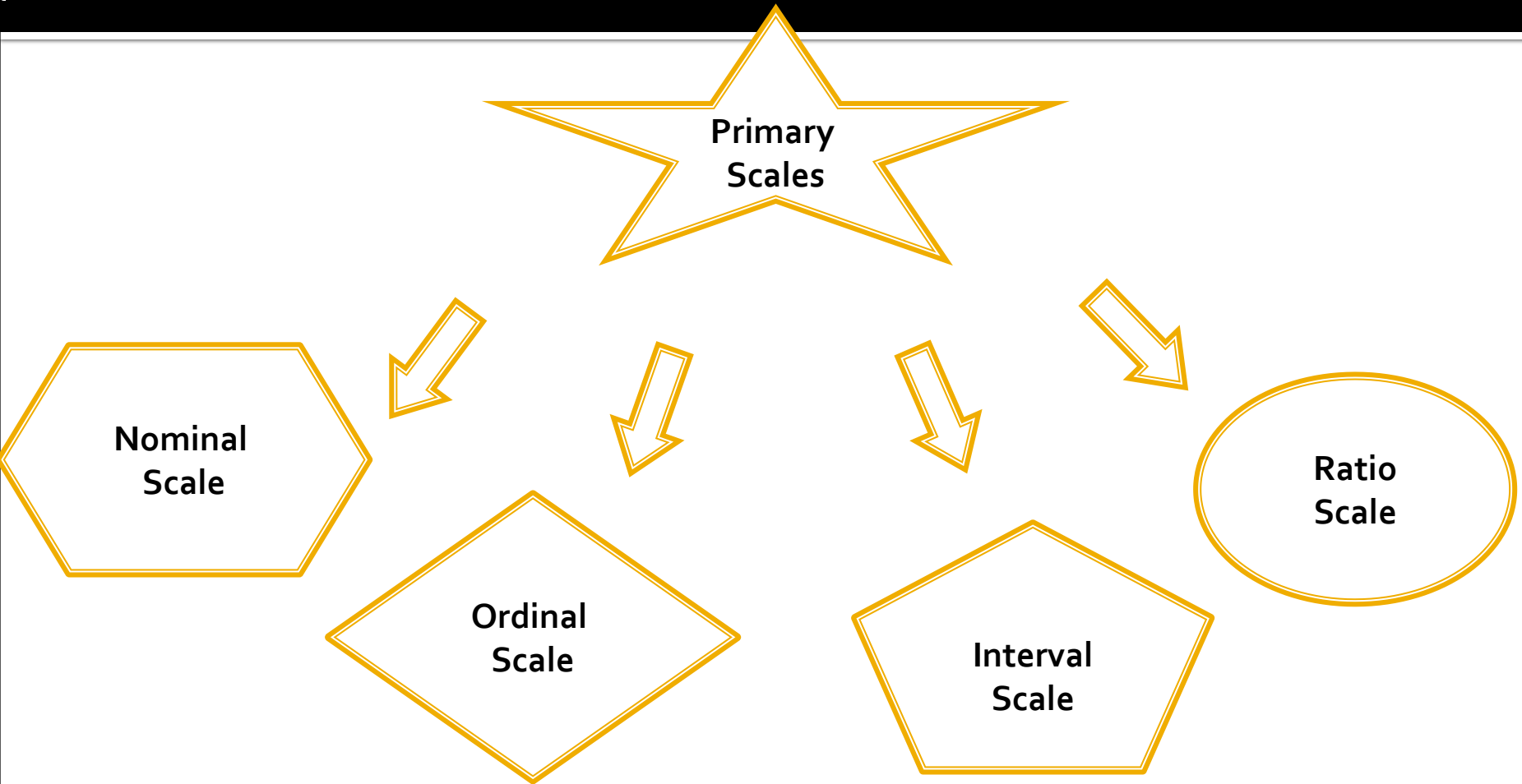
Can we predict one variable if we know one or more other variables?

SOME BASIC CONCEPTS OF MVA

- **MEASUREMENT SCALE**

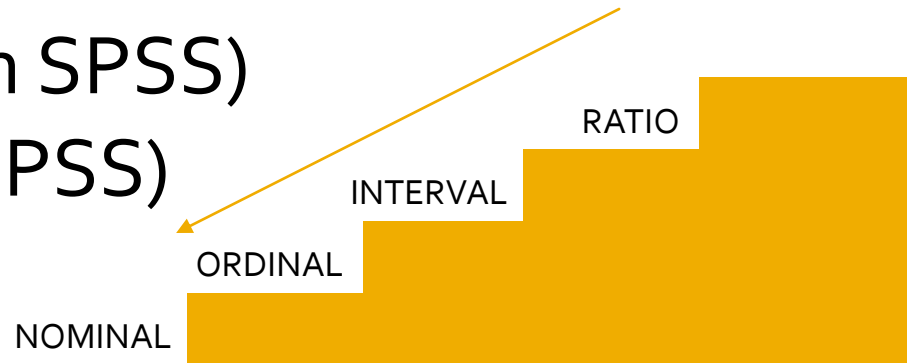
- Data can be classified in to two categories:
 - a) Non-metric (qualitative)
 - b) Metric (quantitative)

Measurement Scales



LEVELS OF DATA

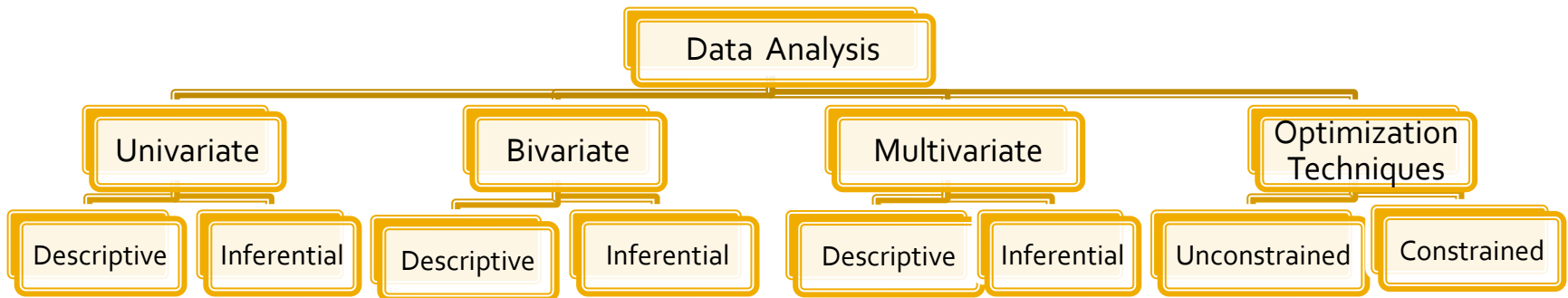
- Nominal
- Ordinal
- Interval (Scale in SPSS)
- Ratio (Scale in SPSS)



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.

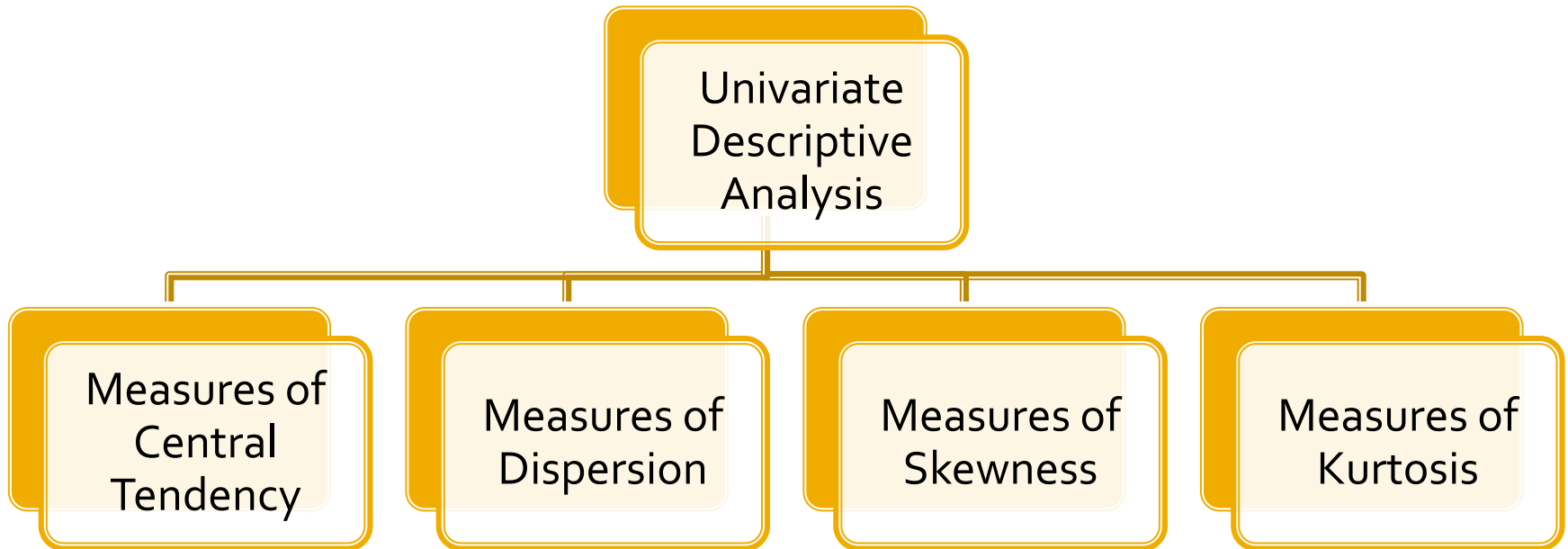
Techniques of Data Analysis



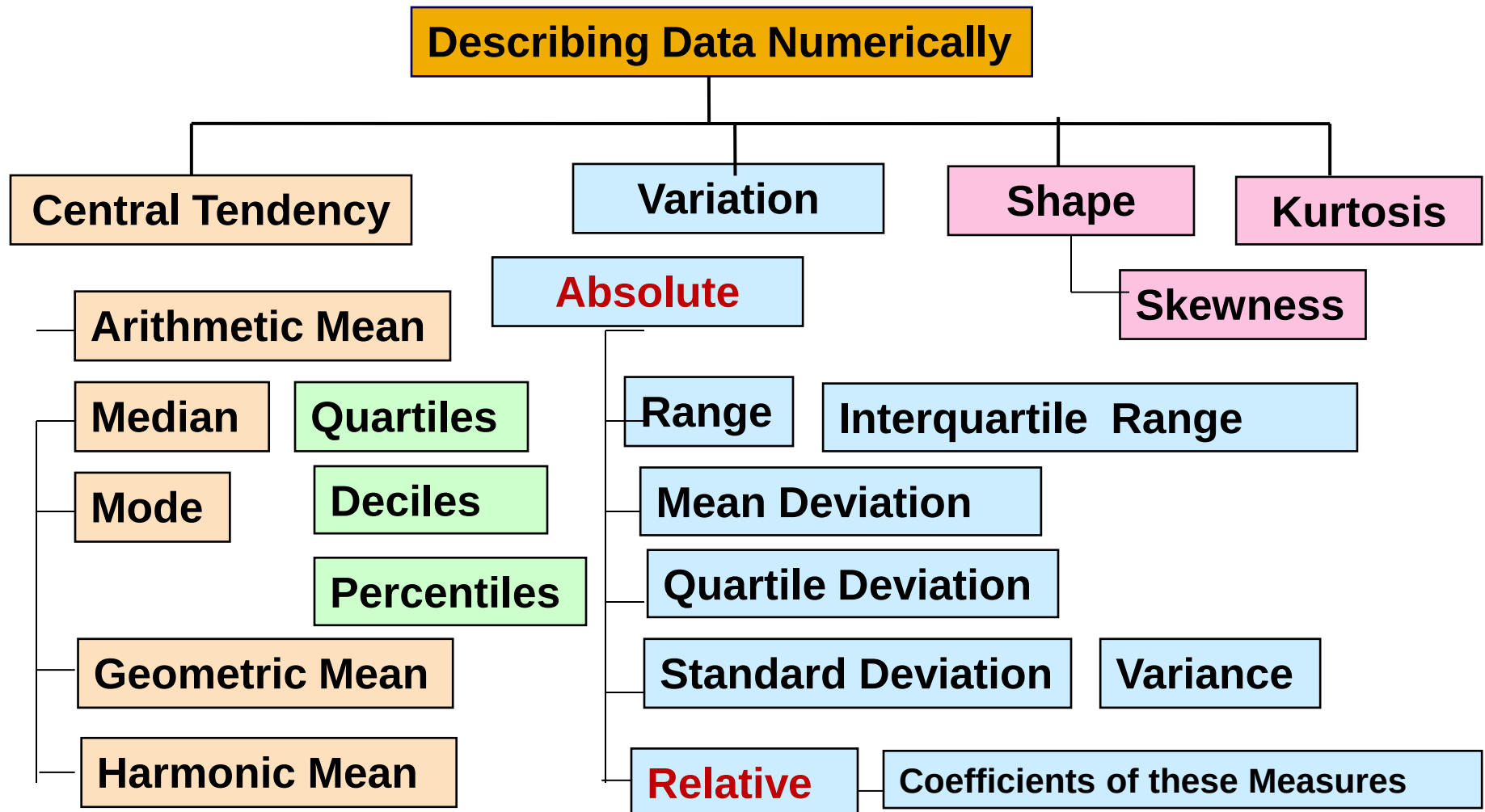
UNIVARIATE ANALYSIS

- 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



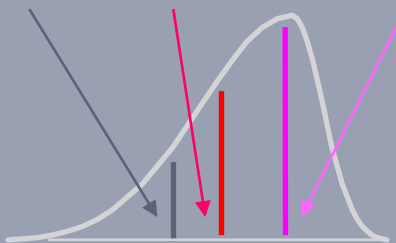
Descriptive Analysis of Univariate Data: Skewness

Shape of a Distribution: Symmetric or skewed

- ▶ Describes how data is distributed
- ▶ Measures of shape

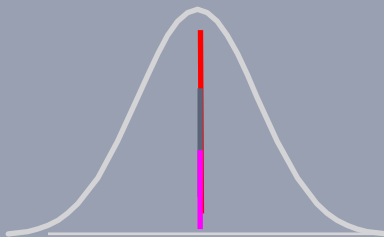
Left-Skewed

Mean < Median < Mode



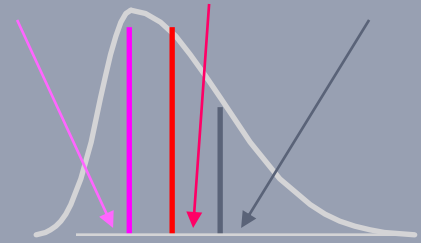
Symmetric

Mean = Median = Mode



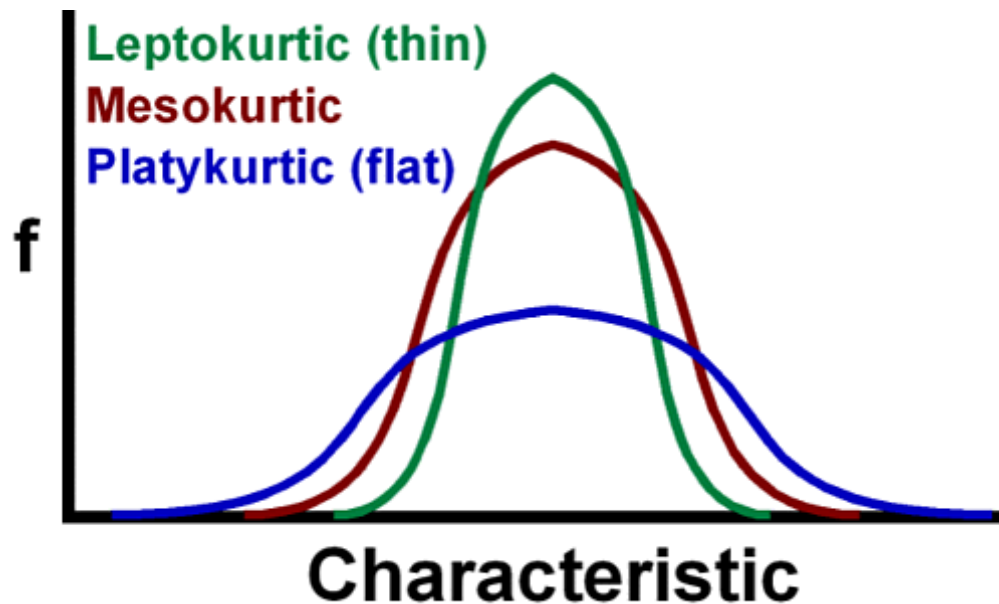
Right-Skewed

Mode < Median < Mean



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 MacGillivray 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

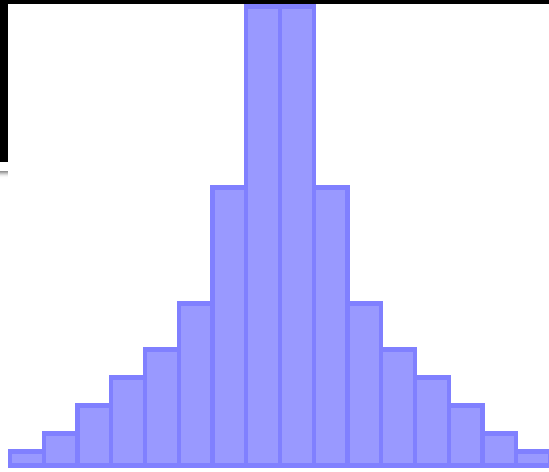


Descriptive Analysis of Univariate Data: Kurtosis

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.

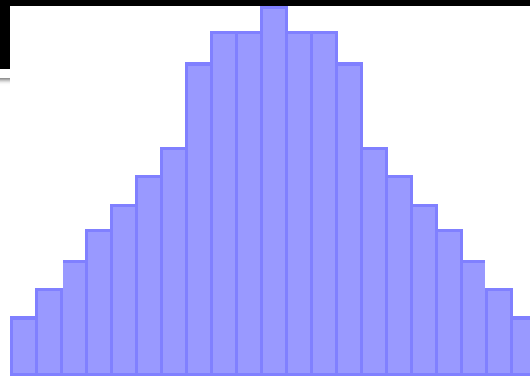
Kurtosis



Leptokurtic

(high peak)

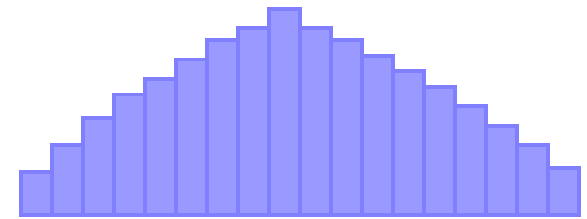
(+ve kurtosis)



Mesokurtic

(normal)

(zero kurtosis)



Platykurtic

(low peak)

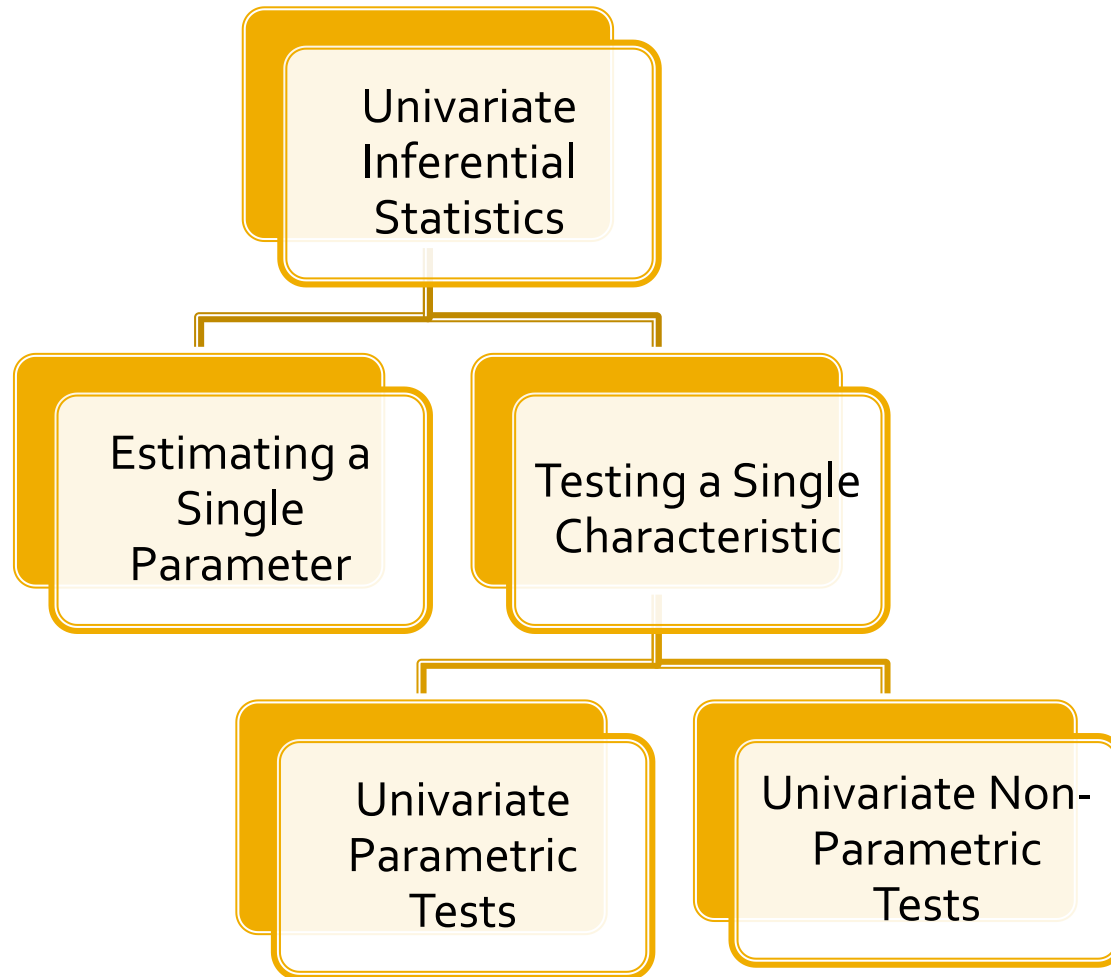
(-ve kurtosis)

Mesokurtic distribution...kurtosis = 3

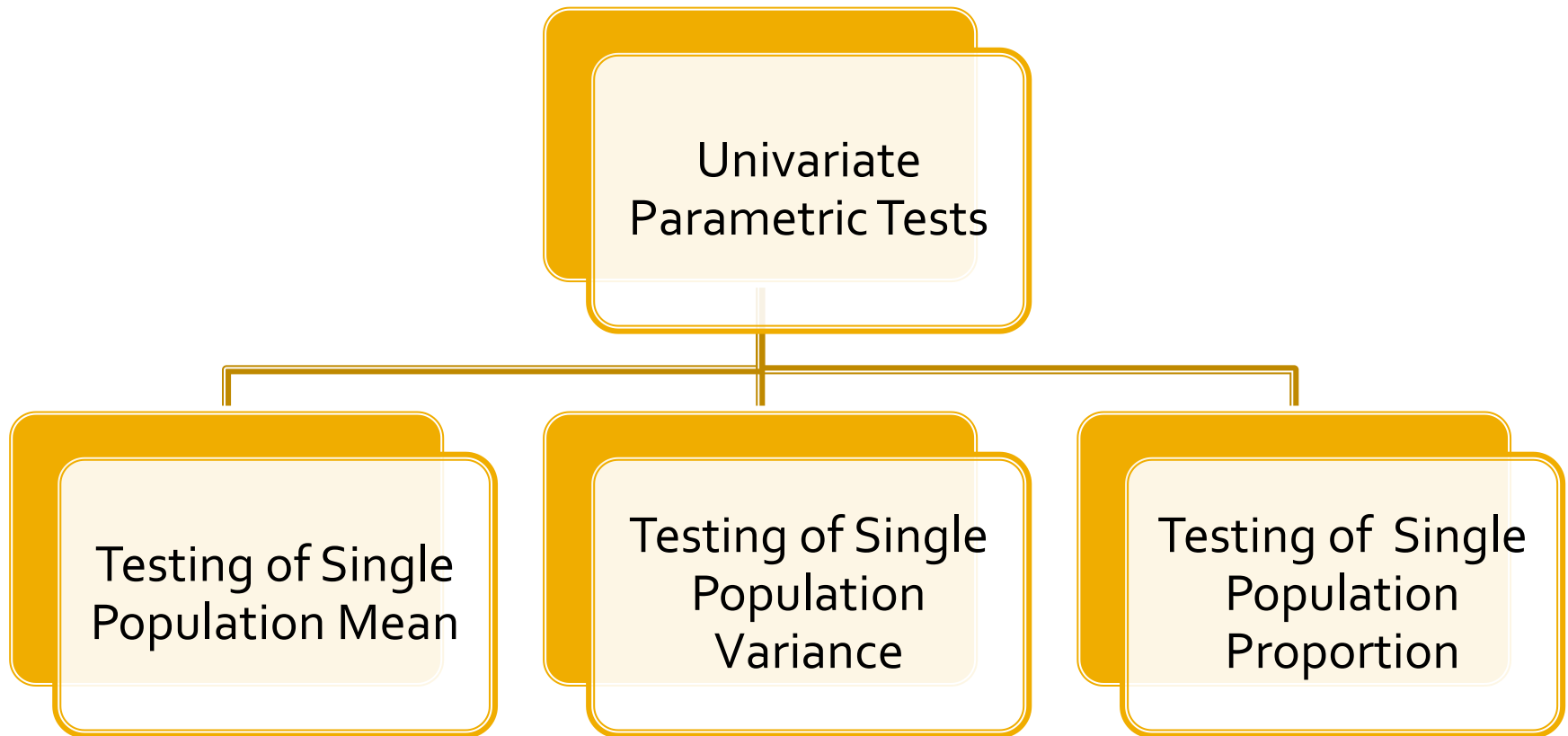
Platykurtic distribution...kurtosis < 3

Leptokurtic distribution...kurtosis > 3

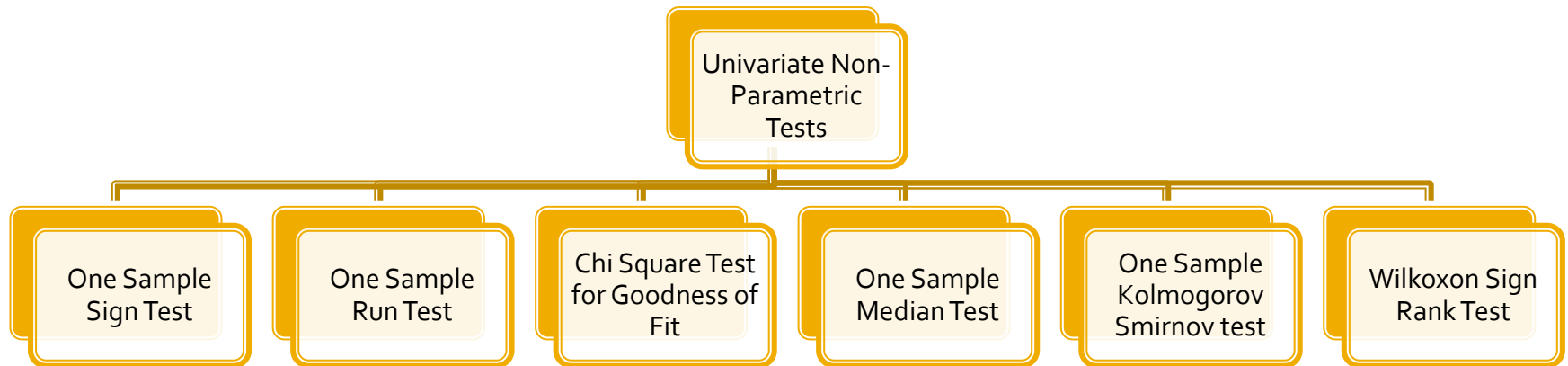
Univariate Inferential Statistics



Univariate Inferential Statistics



Univariate Inferential Statistics



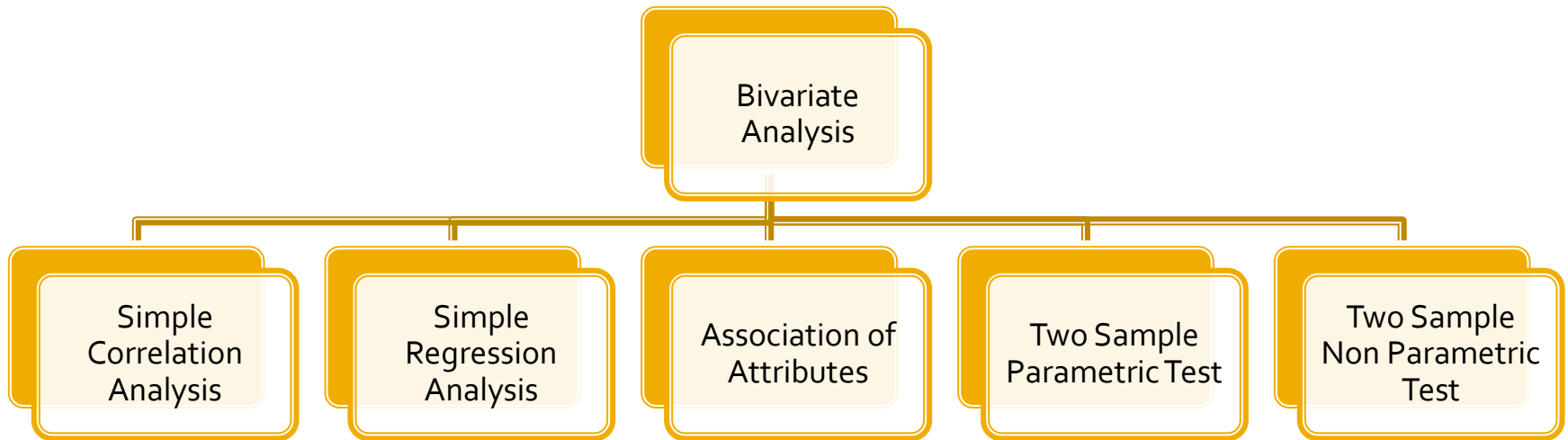
BIVARIATE ANALYSIS

- 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	<i>t</i> -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	<i>t</i> -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	<i>t</i> -test	Paired <i>t</i> -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 BUILDING

□ 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 BUILDING

□ 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 BUILDING

- 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 BUILDING

- 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 BUILDING

□ Stage- 5

□ Interpret the variates

- Interpreting the variates reveals 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 BUILDING

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

Data Preparation

MULTIVARIATE ANALYSIS

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

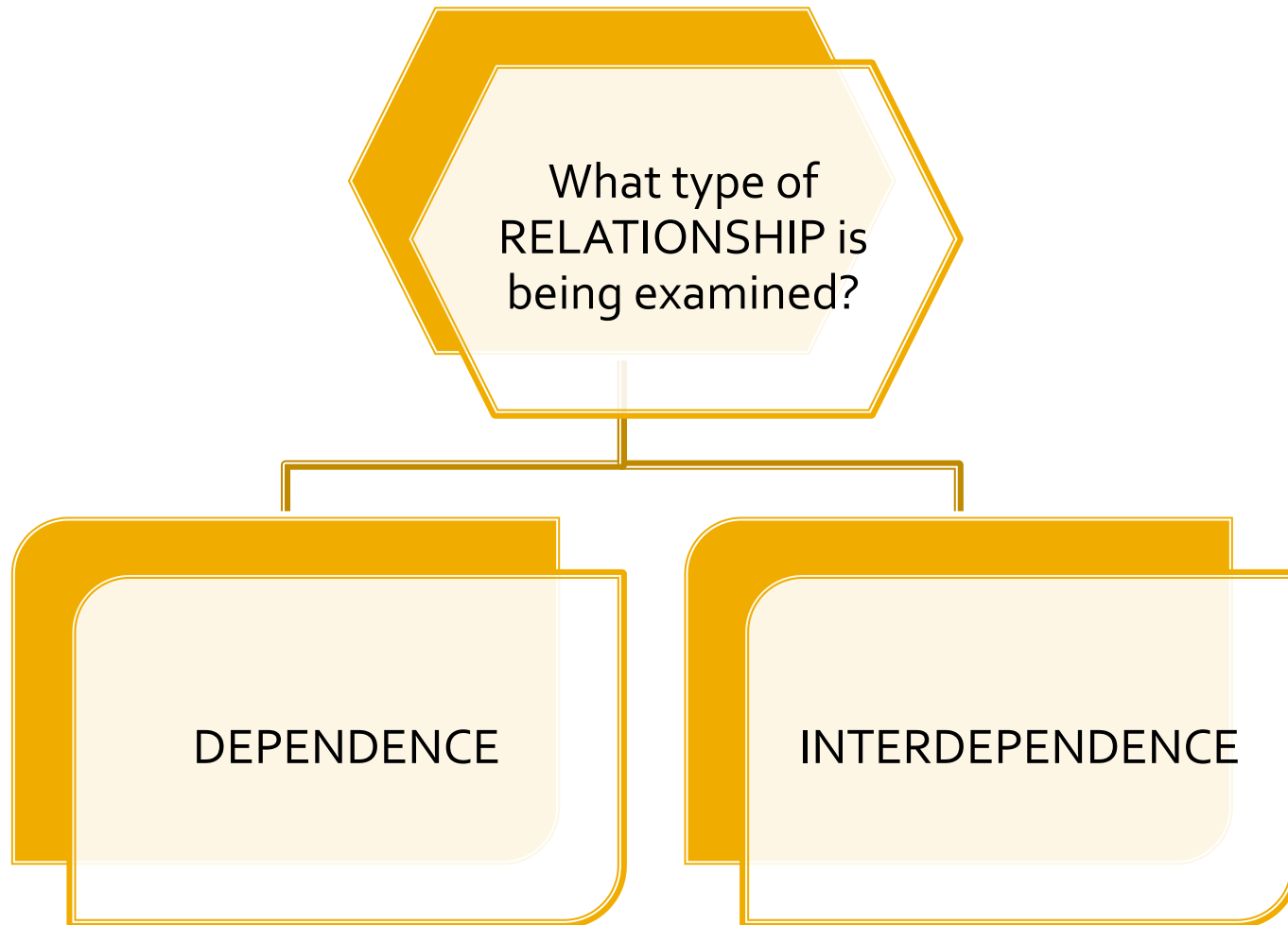
■ THE VARIATE:

- The building block of multivariate analysis is the variate.
- Variate is a linear combinations of Variables with empirically determined weights.
- Variate value = $w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n$
- 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.

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



Dependence Techniques

- Multiple Regression Analysis (Prediction of Scaled Dependent Variable)
- Discriminant Analysis (Classification Technique)
- Logistic Regression Analysis (Classification Technique)
- Structural Equations Modeling
- Conjoint Analysis
- Linear Probability Models
- Canonical Correlation
- MANOVA and MANCOVA

Interdependence Techniques

- Factor Analysis (Data Reduction Technique)
- Cluster Analysis (Respondents (Cases) Grouping Technique)
- Multidimensional Scaling (Perceptual Mapping)
- Correspondence Analysis (Perceptual Mapping)

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

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)

Multivariate Analysis

**TESTING STATISTICAL
ASSUMPTIONS**

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 (EIEWS)
 - 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 (EIEWS)
- Autocorrelation Function (ACF) & Partial Autocorrelation Function (PACF) (EIEWS)

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

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