Multivariate Analysis

- Multivariate analysis techniques are popular because they enable organizations to create knowledge and thereby improve their decision making.
- **Definition :**
  - In statistical terms,

    Multivariate analysis 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 consider as multivariate analysis

Multivariate Analysis

- Many multivariate techniques are extensions of univariate analysis and bivariate analysis.
- Confusion arises in term multivaraite analysis,
  - Sometimes it is simply mean that the examining relationships between or among more than two variables.
  - Some others mean that it is used for problems in which all the multiple variables are assumed to have multivariate normal distributions.
- Thus the multivariate character lies in the multiple combinations of variables and not only in the number of variables.
- Multivariate analysis will include both multivariable techniques and truly multivariate techniques.

Basic Concepts of Multivariate Analysis

- **The Variate**
  - The building block of multivariate analysis is the variate, a linear combination of variables with empirically determined weights.
  - Variables specified by the researcher.
  - Weights determined by the multivariate techniques.
- **Mathematical Representation of variate**
  - A variate of ‘n’ weighted variables (X1 to Xn) can be stated mathematically as :

$$\text{Variate value} = w_1X_1 + w_2X_2 + w_3X_3 + \cdots + w_nX_n$$

where Xn is the observed variable and Wn is the weight determined by the multivariate technique.

Introduction to Multivariate Analysis

- Multivariate data analysis refers to any statistical techniques used to analyze data that arises from more than one variable.
- This essentially models reality where each situation, product or decision involves more than a single variable. The information age has resulted in masses of data in every field.
- Despite the quantum of data available, the ability to obtain a clear picture of what is going on and make intelligent decision is a challenge.
- When available information is stored in database tables containing rows and columns, multivariate analysis can be used to process the information in a meaningful fashion.

Introduction to Multivariate Analysis

- Multivariate analysis methods typically used for :
  - Consumer and market research
  - Quality control and quality assurance across a range of industries such as food and beverage, pharmaceutics, chemicals, energy, telecommunications etc.
  - Process optimization and process control.
  - Research & development.

Measurement Scales

- The amount of information that can be provided by a variable is its type of measurement scale.
- Specifically variables are classified under two categories :-
  - Nonmetric (qualitative) scale
  - Metric (quantitative) scale
- **Qualitative (categorical) data :-**
  - Qualitative also known as categorical data cannot be measured on a numerical scale (quantified).
  - Example:
    - Gender (Male or Female)
    - Size of T-shirt (S, M, L, XL & XXL)
  - Yet these two variables differ in a sense :
    - Nominal
    - Ordinal

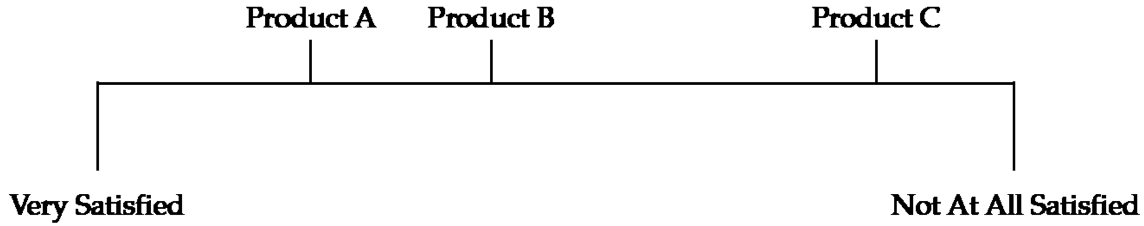
Measurement Scales

- **Nominal (purely categorical) data :**
  - Nominal variables allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories.
  - Example:-
    - Gender, race, colour, city, marital status etc.
    - Marital Status
      - 1) Never Married
      - 2) Divorced
      - 3) Widowed
      - 4) Married
      - 5) Separated

Measurement Scales

- **Ordinal data :**
  - Ordinal scales are the next “higher” level of measurement precision.
  - Ordinal variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say how much more.
  - Or else we can define that an ordinal scale rank-order observations. Class rank and horse race results are the examples.
  - There are two salient attributes of an ordinal scale
    - There is an underlying quantitative measure on which the observation differ.
    - The individual ignorance.
  - A typical example of an ordinal variable i.e., the socioeconomic status of families.

Measurement Scales

- For example, different levels of an individual consumer's satisfaction with several new products can be illustrated, first using an ordinal scale. The following scale shows a respondent's view of three products.
- 

- When we measure this variable with an ordinal scale, we "rank order" the products based on satisfaction level.
- We want a measure that reflects that the respondent is more satisfied with Product A and Product B and more satisfied with Product B and Product C.

Measurement Scales

- For example, Employer’s performance
  - 1) Excellent 2) Good 3) Average 4) Poor 5) Very Poor
- **Quantitative (numerical) Data:**
  - Quantitative data can be easily measured on a numerical scale : variables which can be quantified in terms of units are all quantitative.
  - Example :
    - No. Of students per class and height.
    - Yet, these two variables differ in their nature.
      - Discrete
      - Continuous
- **Discrete Data :**
  - Discrete data occur as definite and separate values. It assumes values which are countable so that these are gaps between its successive values.
  - Example :
    - When counting the number of children in a class , we use natural numbers (0,1,2, ... N)

Measurement Scales

- **Continuous Data:**
  - Continuous data occur as the whole set of real numbers or a subset of it. In other words, there are no gaps between successive so that a continuous variable assumes all the values (including all the decimals) between gives boundaries.
  - Example :
    - Temperature , Height, weight and speed
    - Continuous data can be measured on :
      - Interval Scale and
      - Ratio Scale

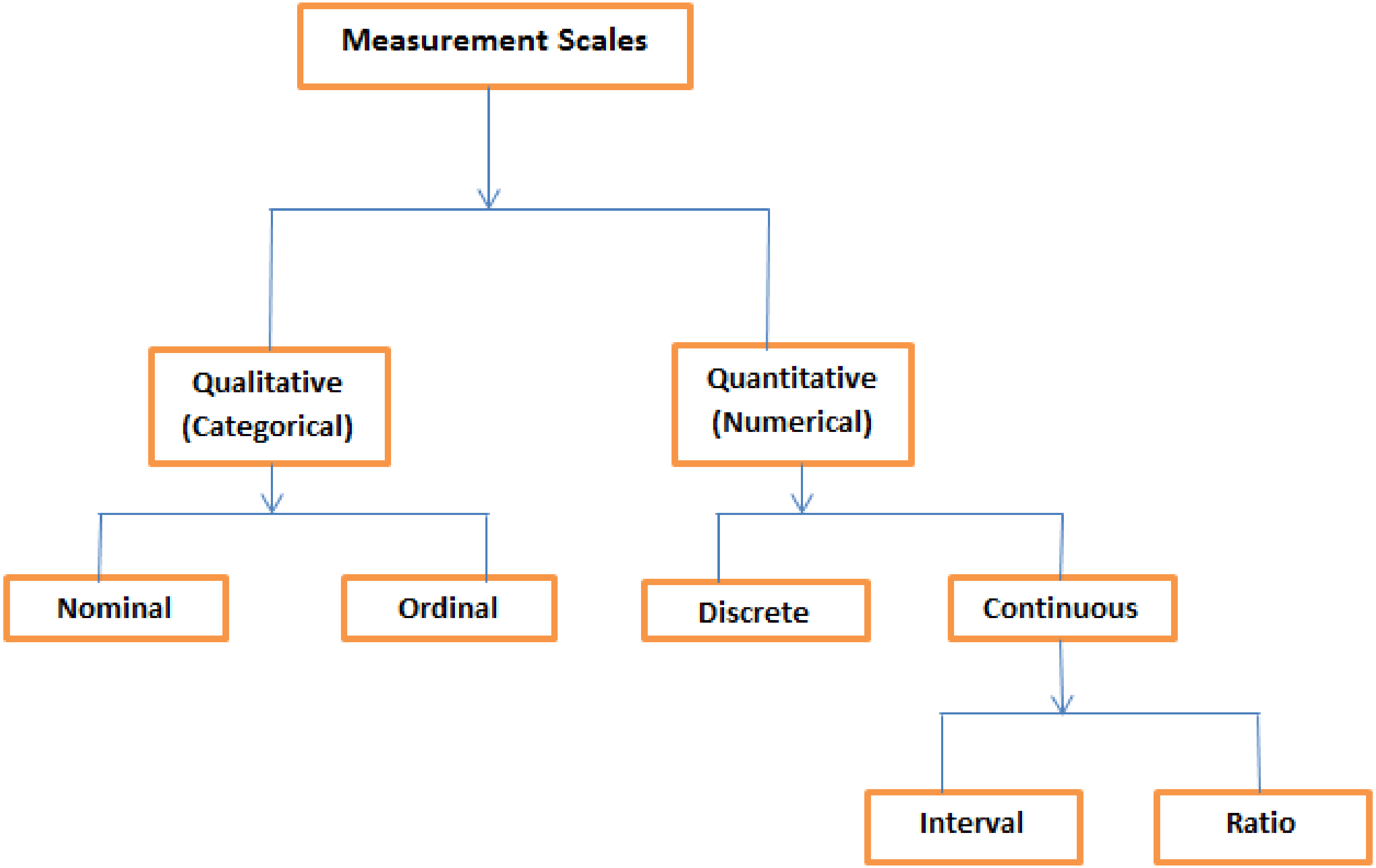
Measurement Scales

- **Interval Scale :**
  - Interval variables allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of difference between them. It don't have absolute zero.
  - These two scales have constant units of measurement, so differences between any two adjacent points on part of the scale are equal.
  - The only real difference between interval and ratio scales is that interval scales use an arbitrary zero point, whereas ratio scales include an absolute zero point.
  - The most familiar interval scales are the Fahrenheit and Celsius temperature scales.

Measurement Scales

- **Ratio Scale :**
  - Ratio variables are very similar to interval variables, in addition to all the properties of interval variables, they feature on identifiable absolute zero point, thus they allow for statements such as x is two times more than y.
  - Typical examples of ratio scales are measures of time and space.
  - Example :
    - Height, since if a person is twice as tall as another, he/she will remain so, irrespective of the units used (centimeter, inches etc.,).

Measurement Scales



Measurement Scales

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Measurement Error

- Measurement Error is the degree to which the observed values are not representative of the true values.
- It arises from many sources, ranging from data entry errors to imprecision of measurement to the inability of respondents to accurately provide information.
- Thus, all variables used in multivariate techniques must be assumed to have some degree of measurement error.
- It adds noise to the observed or measured variable.
- Thus the observed value obtained represents both the “true level” and the “noise”.

Measurement Error

- Our goal is to reduce the degree of measurement error present in any measure.
- By addressing two important characteristics of a measure :
  - Validity
  - Reliability
- **Validity :**
  - Degree to which a measure accurately represents what is its supposed to be.
  - Ensuring validity starts with a thorough understanding of what is to be measured and making the measurement as “correct” and accurate as possible.
  - Accuracy does not measure validity.

Measurement Error

- **Reliability:**
  - Degree to which the observed variable measures the “true” and is “error free”; thus its the opposite of measurement error.
  - More reliable measures will show greater consistency than less reliable measures.
  - Choose the variable with the higher reliability.
- **Employing Multivariate Measurement:**
  - Also called as summated scales, for which several variables are joined in a composite measure to represent a concept.
  - The objective is to avoid the use of only a single variable to represent a concept and instead to use several variables as indicators.
  - The use of multiple indicator enables to precisely specify the desired responses.

Measurement Error

- **Employing Multivariate Measurement:**
  - It does not place a total reliance on a single response, but instead on the average response to a set of related responses.
- **The Impact of measurement error**
  - The impact of measurement error and poor reliability cannot be directly seen because they are embedded in the observed variables.
  - Increase in reliability and validity, which in turn will result in a more accuracy of the variables.
  - Poor results are not always due to measurement error, but the presence of error will make multivariate techniques less powerful.
  - Reducing measurement error, although it takes effort, time and additional resources, it may improve marginal results and strengthen the proven results as well.

Measurement Error

- **Types of statistical Error and Statistical Power :**
  - Interpreting statistical inferences requires to specify the acceptable levels of statistical error that result from using a sample (**known as sampling error**).
  - The most common approach is to specify the level of **Type I error**, also known as **alpha (α)**.
  - Type I error is the probability of rejecting the null hypothesis when it is actually true – generally referred to as a **false positive**.
  - When specifying the level of Type I error, it also determines an associated error, termed **Type II error, or beta (β)**.
  - The Type II error is the probability of not rejecting the null hypothesis when it is actually false.

Measurement Error

- **Types of statistical Error and Statistical Power :**
  - An extension of Type II error is **1 – β**, referred to as the **power** of the statistical inference test.
  - Power is the probability of correctly rejecting the null hypothesis when it should be rejected.
  - Thus, power is the probability that statistical significance will be indicated if it is present.
  - The relationship of the different error probabilities in testing for the difference in two means is shown here

		Real y	
		No Difference	*Difference*
Statistical Decision	H_0: *No Difference*	**$1 - \alpha$**	**β** **Type II error**
	H_a: *Difference*	**α** **Type I error**	**$1 - \beta$** **Power**

Measurement Error

- Specifying alpha establishes the level of acceptable statistical significances, it is the level of power that gives the probability of success in finding the differences if they actually exist.
- The Type I and Type II errors are inversely related.
- Thus, Type I error becomes more restrictive as the probability of a Type II error increases.
- That is, reducing Type I errors reduces the power of the statistical test.
- Needs a balance between the level of alpha and the resulting power.

Impacts on Statistical Power

- Power is not solely a function of alpha. Power is determined by three factors:
- **Effect size :**
  - The probability of achieving statistical significance is based not only on statistical considerations, but also on the actual size of the effect.
  - Thus, the effect size helps researchers determine whether the observed relationship (difference or correlation) is meaningful.
  - For example, the effect size could be a difference in the means between two groups or the correlation between variables.
  - When examining effect sizes, a larger effect is more likely to be found than a smaller effect and is thus more likely to impact the power of the statistical test.
  - To assess the power of any statistical test, the researcher must first understand the effect being examined.
  - Effect sizes are defined in standardized terms for ease of comparison.
  - Mean differences are stated in terms of standard deviations, thus an effect size of .5 indicates that the mean difference is one-half of a standard deviation.
  - For correlations, the effect size is based on the actual correlation between the variables.

Impacts on Statistical Power

- **Alpha (α) :**
  - As alpha becomes more restrictive, power decreases.
  - Therefore, as the researcher reduces the chance of incorrectly saying an effect is significant when it is not, the probability of correctly finding an effect decreases.
  - Conventional guidelines suggest alpha levels of .05 or .01.
  - Researchers should consider the impact of a particular alpha level on the power before selecting the alpha level.
- **Sample size :**
  - At any given alpha level, increased sample sizes always produce greater power for the statistical test.
  - As sample sizes increase, researchers must decide if the power is too high.
  - By "too high" we mean that by increasing sample size, smaller and smaller effects (e.g., correlations) will be found to be statistically significant, until at very large sample sizes almost any effect is significant.
  - The researcher must always be aware that sample size can affect the statistical test either by making it insensitive (at small sample sizes) or overly sensitive (at very large sample sizes).
  - To achieve such power levels, all three factors—alpha, sample size, and effect size—must be considered simultaneously.

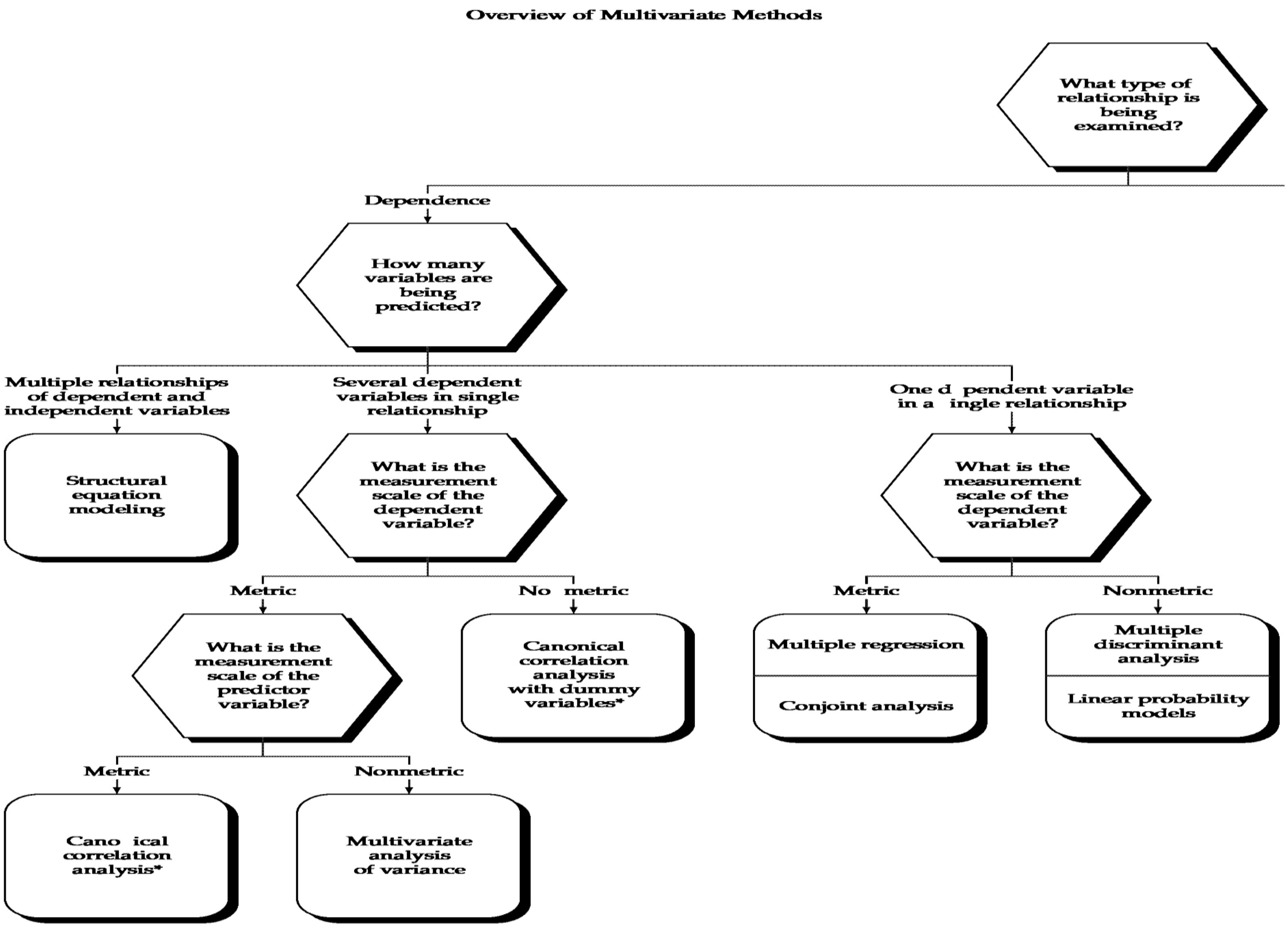
Impacts on Statistical Power

- These interrelationships can be illustrated by a simple example. The example involves testing for the difference between the mean scores of two groups.
- **Using Power with Multivariate Techniques**
- Researchers can use power analysis either in the study design or after data is collected.
- In designing research studies, the sample size and alpha level are selected to achieve the desired power.
- Power also is examined after analysis is completed to determine the actual power achieved so the results can be interpreted.

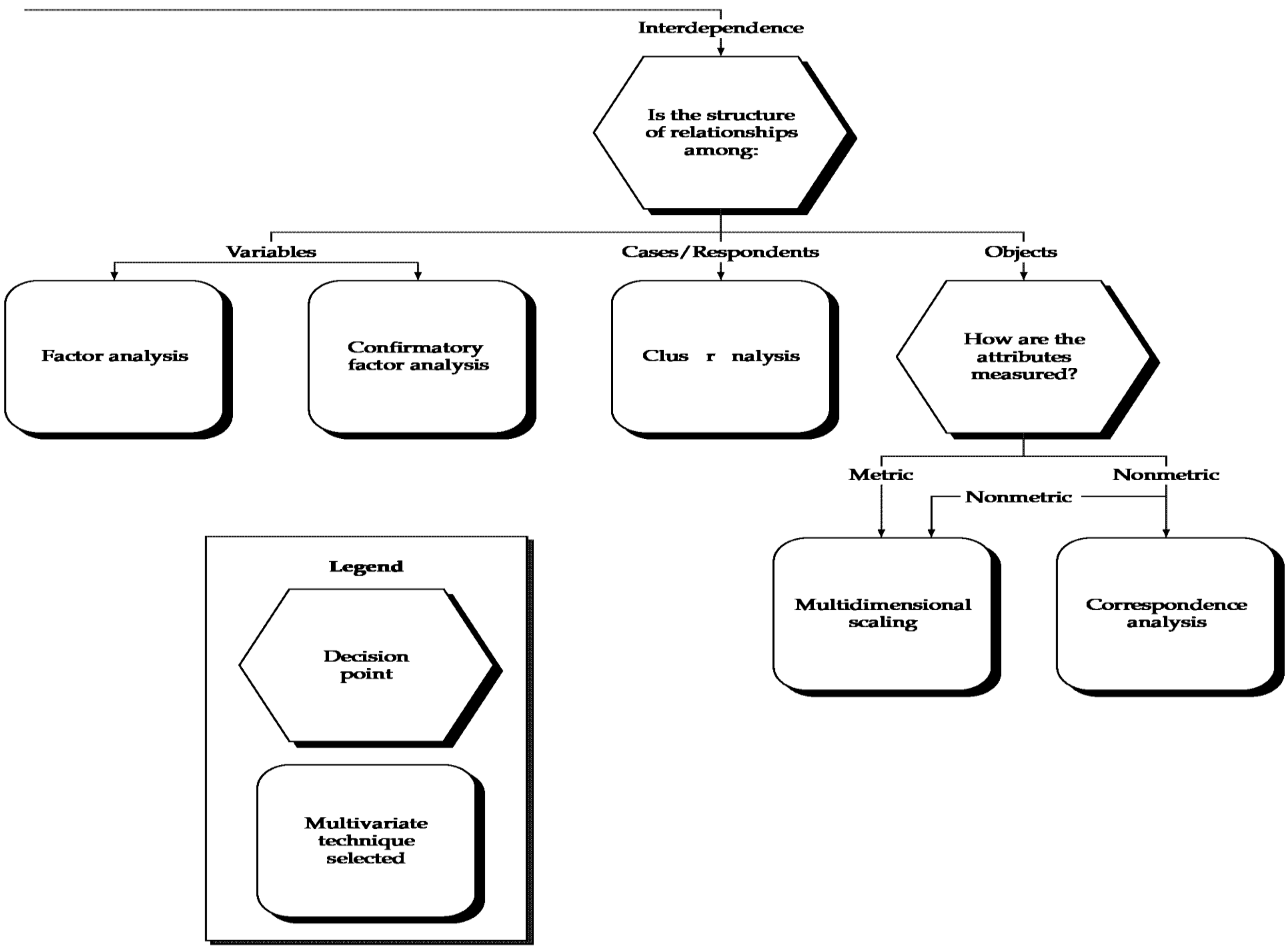
TABLE 1 Power Levels for the Comparison of Two Means: Variations by Sample Size, Significance Level, and Effect Size

	alpha (α) = .05		alpha (α) = .01	
	Effect Size (ES)		*Effect Size (ES)*	
Sample Size	**Small (.2)**	**Moderate (.5)**	**Small (.2)**	**Moderate (.5)**
20	.095	.338	.025	.144
40	.143	.598	.045	.349
60	.192	.775	.067	.549
80	.242	.882	.092	.709
100	.290	.940	.120	.823
150	.411	.990	.201	.959
200	.516	.998	.284	.992

Classification of Multivariate Techniques



Classification of Multivariate Techniques



Classification of Multivariate Techniques

- Dependence Technique :
  - Defined as one in which a variable or a set of variables is identified as the dependent variable to be predicted or explained by other variables known as independent variables.
  - Example : Multiple Regression analysis
- Inter-Dependence Technique :
  - Defined as one in which no single variable or a group of variables is defined as being dependent or independent.
  - Example : Factor Analysis

Classification of Multivariate Techniques

- Dependence Techniques :
  - Can be categorized by two characteristics
  - (i) the number of dependent variables
  - (ii) the type of measurement scale employed by the variables
  - Based on dependent variables it can be classified as
    - Single dependent variable
    - Several dependent variables
    - Several dependent/ independent variables
  - Based on type of measurement it can be classified as
    - Metric (quantitative/numerical)
    - Nonmetric (qualitative/categorical)

Classification of Multivariate Techniques

- Dependence Techniques :
  - Single dependent variable and metric
    - Use either **multiple regression analysis or conjoint analysis**
  - Single dependent variable and nonmetric
    - Use either **multiple discriminant analysis and linear probability models**
  - Several dependent variables are metric
    - If independent variables are nonmetric
      - **Use multivariate analysis of variance (MANOVA)**
    - If independent variables are metric
      - **Use canonical correlation**
  - Several dependent variables are nonmetric
    - They can be transformed through dummy variable coding (0-1) and **canonical analysis** can be used.
  - Thus multivariate techniques range from the general method of canonical analysis to specialized technique of structural equation modelling.

TABLE 2 The Relationship Between Multivariate Dependence Methods

Canonical Correlation

$Y_1 + Y_2 + Y_3 + \cdots + Y_n$	$=$	$X_1 + X_2 + X_3 + \cdots + X_n$
(metric, nonmetric)		(metric, nonmetric)

Multivariate Analysis of Variance

$Y_1 + Y_2 + Y_3 + \cdots + Y_n$	$=$	$X_1 + X_2 + X_3 + \cdots + X_n$
(metric)		(nonmetric)

Analysis of Variance

Y_1	$=$	$X_1 + X_2 + X_3 + \cdots + X$
(metric)		(nonmetric)

Multiple Discriminant Analysis

Y_1	$=$	$X_1 + X_2 + X \quad \cdots + X_n$
(nonmetric)		(m tric)

Multiple Regression alysis

Y_1	$=$	$X_1 + X_2 + X_3 + \cdots + X_n$
(metric)		(metric, nonmetric)

Conjoint Analysis

Y_1	$=$	$X_1 + X_2 + X_3 + \cdots + X_n$
(nonmetric, metric)		(nonmetric)

St uctural Equation Modeling

Y_1	$=$	$X_{11} + X_{12} + X_{13} + \cdots + X_{1n}$
Y_2	$=$	$X_{21} + X_{22} + X_{23} + \cdots + X_{2n}$
Y_m	$=$	$X_{m1} + X_{m2} + X_{m3} + \cdots + X_{mn}$
(metric)		(metric, nonmetric)

Classification of Multivariate Techniques

- Interdependence Techniques :
  - In this variables cannot be classified as either dependent or independent.
  - All the variables are analysed simultaneously to find an underlying structure to the entire set of variables or subject.
  - If the structure of the variables is to be analysed, then **factor analysis or confirmatory factor analysis** is the appropriate technique.
  - If cases or respondents are to be grouped to represent a structure, the **cluster analysis** is selected.
  - Generally factor analysis and cluster analysis are considered to be metric interdependence techniques.
  - **Non-metric data**
    - Dummy variable coding to use with special forms of factor analysis and cluster analysis.
  - If the inter dependencies of objects measured by non metric data are to be analyzed, **correspondence analysis** is appropriate technique.

Types of Multivariate Techniques

- Principal components and common factor analysis
- Multiple regression and multiple correlation
- Multiple discriminant analysis and logistic regression
- Canonical correlation analysis
- Multivariate analysis of variance and covariance
- Conjoint analysis
- Cluster analysis
- Perceptual mapping (multidimensional scaling)
- Correspondence analysis
- Structural equation modeling and confirmatory factor analysis

Principal Components and Common Factor Analysis

- Factor analysis, including both Principal components and common factor analysis is a statistical approach that can be used to analyze interrelationships among a large number of variables.
- To explain these variables in terms of their common underlying dimensions called **factors**.
- The objective is to find a way of condensing the information contained in a number of original variables into a small set of variates (**factors**) with a minimum loss of information.
- Example :
  - Understand the relationships between customers ratings of fast-food restaurant.
  - Six variables : food taste, food temperature, freshness, waiting time, cleanliness and friendliness of employees
  - By analyzing customer responses,
    - food taste, food temperature and freshness – **food quality (one factor)**
    - waiting time, cleanliness and friendliness of employees – **service quality (one factor)**

Multiple Regression

$$\underset{\text{(metric)}}{Y_1} = \underset{\text{(metric, nonmetric)}}{X_1 + X_2 + X_3 + \cdots + X_n}$$

- It is the appropriate technique, if the problem involves a single metric dependent variable to be related to two or more metric independent variables.
- The objective is to predict the changes in the dependent variable in response to changes in the independent variables.
- Its is achieved through the statistical rule of least squares.
- Example :
  - Monthly expenditures on dining out **(dependent variable)**
  - Family’s income, its size and the age of head of household **(independent variables)**
  - Company’s sales from its expenditure for advertising, the number of salespeople and the number of product stores.

Multiple Discriminant Analysis (MDA) and Logistic Regression

$$\underset{\text{(nonmetric)}}{Y_1} = \underset{\text{(m tric)}}{X_1 + X_2 + X \cdots + X_n}$$

- MDA is the appropriate multivariate technique if the single dependent variable is dichotomous or multi-chotomous and also nonmetric.
- Multiple regression the independent variables are assumed to be metric.
- It is mainly applicable for the total sample can be divide into groups based on a nonmetric dependent variables characterizing into several known classes.
- The primary objectives are to understand group differences and to predict the likelihood that an entity will belong to a particular class based on several metric independent variables.
- Example :
  - Its is used to distinguishing applications include
  - Heavy product users from light users
  - Males from females
  - National-brand buyers from private-label buyers
  - Good credit risks from poor credit risks.

Multiple Discriminant Analysis and Logistic Regression

- Logistic regression models, are often referred to as logit analysis, are a combination of multiple regression and multiple discriminant analysis,
- It is similar to multiple regression analysis, but a main difference is it uses dependent variable as nonmetric as in discriminant analysis.
- The nonmetric scale of dependent variable requires differences in the estimation method and assumptions about the type of underlying distribution, otherwise its quite similar to multiple regression.
- Logistic models are distinguished from discriminant analysis that they include all types of independent variables (metric and nonmetric) and don not require the assumption of multivariate normality.
- But in many example, with more than two levels of dependent variable, discriminant analysis is the more appropriate technique.
- Example :
  - Financial advisors were trying to selecting emerging firms for start-up investment.
  - Two classes :
    - Successful over a period of five years
    - Unsuccessful after five years
  - For each firm they had financial and managerial data.

Canonical Correlation

$$\underset{\text{(metric, nonmetric)}}{Y_1 + Y_2 + Y_3 + \cdots + Y_n} \quad = \quad \underset{\text{(metric, nonmetric)}}{X_1 + X_2 + X_3 + \cdots + X_n}$$

- Canonical correlation can be viewed as a logical extension of multiple regression analysis.
- Canonical analysis the objective is to correlate simultaneously several metric dependent variables and several metric independent variables.
- Multiple regression involves single dependent variable, whereas canonical correlation involves multiple dependent variables.
- The underlying principle is to develop a linear combination of each set of variables in a manner that maximizes the correlation between the two sets.
- The procedure involves obtaining a set of weights for the dependent and independent variables that provides the maximum simple correlation between the set of independent variables and dependent variables.
- Example :
  - A company conducts a study that collects information on its service quality based on answers to 50 metrically measured questions.
  - Canonical correlation could be used to compare the perceptions of the world-class companies on the 50 questions with the perceptions of the company.

Multivariate Analysis of Variance and Covariance

$$\underset{\text{(metric)}}{Y_1 + Y_2 + Y_3 + \cdots + Y_n} = \underset{\text{(nonmetric)}}{X_1 + X_2 + X_3 + \cdots + X_n}$$

- Multivariate analysis of variance (MANOVA) is a statistical technique that can be used to simultaneously explore the relationship between several categorical independent variables and two or more metric dependent variables.
- It is the extension of univariate analysis of variance (ANOVA).
- Multivariate analysis of covariance (MANCOVA) can be used in conjunction with MANOVA to remove the effect of any uncontrolled metric independent variables (Covariates) on the dependent variables.
- MANOVA is useful for an experimental situation to test hypotheses concerning he variance in group responses on two or more metric dependent variables.
- Example :
  - A company wants to know if a humorous AD will be more effective with its customers than a nonhumorous AD.
  - It could ask its ad agency to develop two ads – one humorous and one nonhumorous and then show group of customers the two ads.
  - MANOVA would be the technique to determine the extent of any statistical differences between the perception of customers who saw the humorous AD versus those who saw the nonhumorous.

Conjoint Analysis

$$\underset{\text{(nonmetric, metric)}}{Y_1} = \underset{\text{(nonmetric)}}{X_1 + X_2 + X_3 + \cdots + X_n}$$

- Conjoint analysis is an emerging technique that bring new sophistication to the evaluation of objects, such as new products, services or ideas.
- It is mainly used in the direct application of new product or service development.
- It is able to assess the importance of attributes as well as the levels of each attribute while consumers evaluate only a few product profiles, which are combinations of product levels.
- Example :
  - A product company has three attributes (price, quality, and color), each at three possible levels (red, yellow and blue).
  - Instead of having to evaluate all 27 (3X3X3) possible combinations, a subset of (9 or more) can be evaluated for their attractiveness of customers.
  - It provides not only the importance of each attribute but also the importance of each level.

Cluster Analysis

- Cluster analysis is an analytical technique for developing meaningful subgroups of individuals or objects.
- Specifically, the objective is to classify a sample of entities (individuals or objects) into a small number of mutually exclusive groups based on the similarities among the entities.
- In cluster analysis, unlike discriminant analysis, the groups are not predefined. Instead, the technique is used to identify the groups.
- Cluster analysis usually involves at least three steps.
  - The first is the measurement of some form of similarity or association among the entities to determine how many groups really exist in the sample.
  - The second step is the actual clustering process, whereby entities are partitioned into groups (clusters).
  - The final step is to profile the persons or variables to determine their composition.
- Many times this profiling may be accomplished by applying discriminant analysis to the groups identified by the cluster technique.
- Example :
  - assume a restaurant owner wants to know whether customers are patronizing the restaurant for different reasons.
  - Data could be collected on perceptions of pricing, food quality, and so forth.
  - Two clusters can be formed based on
    - highly motivated by low prices versus
    - those who are much less motivated to come to the restaurant based on price considerations.

Perceptual Mapping

- In perceptual mapping (also known as multidimensional scaling), the objective is to transform consumer judgments of similarity or preference (e.g., preference for stores or brands) into distances represented in multidimensional space.
- If objects A and B are judged by respondents as being the most similar compared with all other possible pairs of objects, perceptual mapping techniques will position objects A and B in such a way that the distance between them in multidimensional space is smaller than the distance between any other pairs of objects.
- The resulting perceptual maps show the relative positioning of all objects, but additional analyses are needed to describe or assess which attributes predict the position of each object.
- Example :
  - assume an owner of a Burger King franchise wants to know whether the strongest competitor is McDonald's or Wendy's.
  - A sample of customers is given a survey and asked to rate the pairs of restaurants from most similar to least similar.
  - The results show that the Burger King is most similar to Wendy's, so the owners know that the strongest competitor is the Wendy's restaurant because it is thought to be the most similar.
  - Follow-up analysis can identify what attributes influence perceptions of similarity or dissimilarity.

Correspondence Analysis

- Correspondence analysis is a recently developed interdependence technique that facilitates the perceptual mapping of objects (e.g., products, persons) on a set of nonmetric attributes.
- Researchers are constantly faced with the need to “quantify the qualitative data” found in nominal variables.
- Correspondence analysis differs from the interdependence techniques discussed earlier in its ability to accommodate both nonmetric data and nonlinear relationships.
- In its most basic form, correspondence analysis employs a contingency table, which is the cross-tabulation of two categorical variables.
- It then transforms the nonmetric data to a metric level and performs dimensional reduction (similar to factor analysis) and perceptual mapping.
- Correspondence analysis provides a multivariate representation of interdependence for nonmetric data that is not possible with other methods.
- Example :
  - respondents’ brand preferences can be cross-tabulated on demographic variables (e.g., gender, income categories, occupation)
  - by indicating how many people preferring each brand fall into each category of the demographic variables.
  - Brands perceived as similar are located close to one another.

Structural Equation Modeling

$$\begin{aligned} Y_1 &= X_{11} + X_{12} + X_{13} + \cdots + X_{1n} \\ Y_2 &= X_{21} + X_{22} + X_{23} + \cdots + X_{2n} \\ Y_m &= X_{m1} + X_{m2} + X_{m3} + \cdots + X_{mn} \\ \text{(metric)} & \quad \text{(metric, nonmetric)} \end{aligned}$$

- Structural equation modeling (SEM) is a technique that allows separate relationships for each of a set of dependent variables.
- In its simplest sense, structural equation modeling provides the appropriate and most efficient estimation technique for a series of separate multiple regression equations estimated simultaneously.
- It is characterized by two basic components:
  - (1) the structural model and
  - (2) the measurement model.
- The structural model is the path model, which relates independent to dependent variables.
- In such situations, theory, prior experience, or other guidelines enable the researcher to distinguish which independent variables predict each dependent variable.
- Models discussed previously that accommodate multiple dependent variables—multivariate analysis of variance and canonical correlation—are not applicable in this situation because they allow only a single relationship between dependent and independent variables.
- The measurement model enables the researcher to use several variables (indicators) for a single independent or dependent variable.

Structural Equation Modeling

- Example :
  - A study by management consultants identified several factors that affect worker satisfaction: supervisor support, work environment, and job performance.
  - In addition to this relationship, they noted a separate relationship wherein supervisor support and work environment were unique predictors of job performance.
  - Hence, they had two separate, but interrelated relationships.
- SEM provides a means of not only assessing each of the relationships simultaneously rather than in separate analyses, but also incorporating the multi-item scales in the analysis to account for measurement error associated with each of the scales.

A Structured Approach to Multivariate Model Building

- Numerous multivariate techniques are available and the large set of issues involved in their application.
- Successful completion of a multivariate analysis involves the selection of a correct method.
- Issues ranging from problem definition to a critical diagnosis of the results must be addressed.
- To aid the researcher or user in applying multivariate methods, a six-step approach to multivariate analysis is presented.
- The intent is not to provide a set of procedures to follow but to provide a series of guidelines that emphasize a model-building approach.
- This six-step model-building process provides a framework for developing, interpreting, and validating any multivariate analysis.

A Structured Approach to Multivariate Model Building

- **Stage 1 : Define the Research Problem, Objectives, and Multivariate technique to be used**
  - The starting point for any multivariate analysis is to define the research problem and analysis objectives in conceptual terms before specifying ant variables or measures.
  - The role of conceptual model development cannot be overstated.
  - First view the problem in conceptual terms by defining the concepts and identifying the fundamental relationships to be investigated.
  - It should not be complex and detailed. It can be a just simple representation of the relationships to be studied.
  - For a dependence problem specify both the dependent and independent concepts. For an interdependent application, the dimensions of structure or similarity should be specified.
  - This minimizes the chance that relevant concepts will be omitted in the effort to develop measures and to define the specifics of the research design.

A Structured Approach to Multivariate Model Building

- **Stage 1 : Define the Research Problem, Objectives, and Multivariate technique to be used**
  - With the objective and conceptual model specified, the researcher has only to choose the appropriate multivariate technique based on the measurement characteristics of the dependent and independent variables.
  - Variables for each concept are specified prior to the study in its design, but may be respecified or transformed after the data have been collected.
- **Stage 2 : Develop the Analysis Plan**
  - With the concept model established and multivariate technique selected, attention turns to the implementation issues.
  - The issues include general considerations such as minimum or desired sample sizes and a required type of variables (metric versus nonmetric) and estimation methods.

A Structured Approach to Multivariate Model Building

- **Stage 3 : Evaluate the Assumptions underlying the Multivariate Technique**
  - Once the data collected, the first task is not to estimate the multivariate model but to evaluate its underlying assumptions, both statistical and conceptual, that substantially affect their ability to represent multivariate relationships.
  - For the techniques based on statistical inference, the assumptions of multivariate normality, linearity, independence of the error terms, and equality if variances must all be met.
  - Before any model estimation is attempted, must ensure that both statistical and conceptual assumptions are met.

A Structured Approach to Multivariate Model Building

- **Stage 4 : Estimate the Multivariate Model and Assess Overall Model Fit**
  - With the assumptions satisfied, the analysis proceeds to the actual estimation of the multivariate model and an assessment of overall model fit.
  - In the estimation process, choose options to meet specific characteristics of the data (e.g. Use of Covariances in MANOVA) or maximize the fit to the data (e.g. rotation of factors or discriminant functions).
  - After the model is estimated, the overall model fit is evaluated to achieve,
    - Acceptable levels of statistical criteria (level of significance)
    - Identifies the proposed relationships
    - Achieves practical significance
  - Model will be respecified to achieve better levels of overall fit.
  - Determine whether the results affected by single or small set observations will indicate that the results may be unstable or not generalizable.
  - Ill-fitting observations are identified as outliers, influential observations or other disparate results.

A Structured Approach to Multivariate Model Building

- **Stage 5 : Interpret the Variates**
  - With an acceptable level of model fit, interpreting the variates reveals the nature of the multivariate relationship.
  - The interpretation effects for individual variables is made by examining the estimated coefficeints (weights) for each variable in the variate.
  - For multiple variates, represent underlying dimensions of comparison or association.
  - The interpretation may lead to, respecification of the variables and model formulation. Model is re-estiamated and then interpreted again.
  - The objective is to identify the empirical evidence of multivariate relationships in the sample data that can be generalized to the total population.

A Structured Approach to Multivariate Model Building

- **Stage 6 : Validate the Multivariate Model**
  - Before accepting the results, must subject them to one final set of diagnostic analyses that assess the degree of generalizability of the results by the available validation methods.
  - The attempts to validate the model are directed toward demonstrating the generalizability of the results to the total population.
  - These diagnostic analyses add little to the interpretation of the results but can be viewed as “insurance” that the results are the most descriptive of the data, yet generalizable to the population.
- **A decision Flowchart**
  - For each multivariate technique, the six-step approach to multivariate model building will be given in a decision flowchart partitioned into two sections.
  - The first section **(Stages 1 through 3)** : deals with the issues addressed while preparing for actual model estimation (i.e., research objectives, research design considerations and testing for assumptions).
  - The second section of the decision flowchart **(stages 4 through 6)** deals with the issues pertaining to model estimation, interpretation and validation).

Applications of Multivariate Techniques

- The published applications of multivariate methods have increased tremendously in recent years.
- In order to give some indication of the usefulness of multivariate techniques, following are the short descriptions of the results of studies from several disciplines.
- These descriptions are organized according to the categories of objectives of multivariate techniques.

Applications of Multivariate Techniques

- A taxonomy of multivariate statistical analyses shows that most techniques fall into one of the following categories:

1. Data reduction or structural simplification.
2. Sorting and grouping.
3. Investigation of the dependence among variables.
4. Prediction.
5. Hypothesis construction and testing.

Applications of Multivariate Techniques

- **Data Reduction or Simplification :**
  - Using data on several variables related to cancer patient responses to radio therapy, a simple measure of patient response to radiotherapy was constructed.
  - Track records from many nations were used to develop an index of performance for both male and female athletes.
  - Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions.
  - Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants.

Applications of Multivariate Techniques

- **Sorting and Grouping :**
  - Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing (or planned) computer utilization.
  - Measurements of several psychological variables were used to develop a screening procedure that discriminates alcoholics from non-alcoholics.
  - Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease.
  - The U.S Internal Revenue Service uses data collected from tax returns to sort taxpayers into two groups: those that will be audited and those that will not.

Applications of Multivariate Techniques

- **Investigation of the dependence among variables :**
  - Data on several variables were used to identify factors that were responsible for client success in hiring external consultants.
  - Measurements of variables related to innovation, on the one hand, and variables related to the business environment and business organization , on the other hand, were used to discover why some firms are product innovators and some firms are not.
  - Data on variables representing the outcomes of the 10 decathlon events in the Olympics were used to determine the physical factors responsible for the success in the decathlon.
  - The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance.

Applications of Multivariate Techniques

- **Prediction :**
  - The associations between test scores and several high school performance variables and several college performance variables were used to develop predictors of success in college.
  - Data on several variables related to the size distributions of sediments were used to develop rules for predicting different depositional environments.
  - Measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers.
  - Data on several variables for chickweed plants were used to develop a method for predicting the species of a new plant.

Applications of Multivariate Techniques

- **Hypotheses Testing :**
  - Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends.
  - Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores.
  - Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories.
  - Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innovation.

  The preceding descriptions offer glimpses into the use of multivariate methods in widely diverse fields.