Basic Concepts of Multivariate Analysis

MCF112M

NDD W1

Overview

- What Is Multivariate Analysis?
- Three Converging Trends
- Multivariate Analysis in Statistical Terms
- Some Basic Concepts of Multivariate Analysis
- Managing the Multivariate Model
- A Classification of Multivariate Techniques
- Types of Multivariate Techniques
- Guidelines for Multivariate Analyses and Interpretation
- A Structured Approach to Multivariate Model Building

Overview

- The Challenge of Big Data Research Efforts
- Preliminary Examination of the Data
- Missing Data
- Outliers
- Testing the Assumptions of Multivariate Analysis
- Data Transformations
- An Illustration of Testing the Assumptions Underlying Multivariate Analysis
- Incorporating Nonmetric Data with Dummy Variables

What Is Multivariate Analysis?

What is Multivariate Analysis?

What is multivariate?

- Univariate, bivariate, and multivariate data
- Multiple measurements on each individual or object under investigation.

What is multivariate analysis?

- All statistical methods that simultaneously analyze multivariate measurements.

Why use it?

- Methodological Benefits
  - Measurement
  - Explanation & Prediction
  - Hypothesis Testing
- Improved decision making

THREE CONVERGING TRENDS

- ❑ Topic 1: Rise of Big Data
- ❑ Topic 2: Statistical Versus Data Mining Models
- ❑ Topic 3: Causal Inference

Topic 1: The Rise of Big Data

Unique elements of Big Data focused in 6V areas:

- Volume
- Variety
- Velocity
- Veracity
- Variability
- Value

Impacts on:

- Organizational Decisions and Academic Research – improved decision-making capabilities as well as the explosion of data available to characterize situations on dimensions never before available
- Analytics and the Analyst – expansion of the domains of study embracing analytics as well as methodological challenges due to seemingly “unlimited” data

Problems – emerging on technological, ethical and sociological fronts

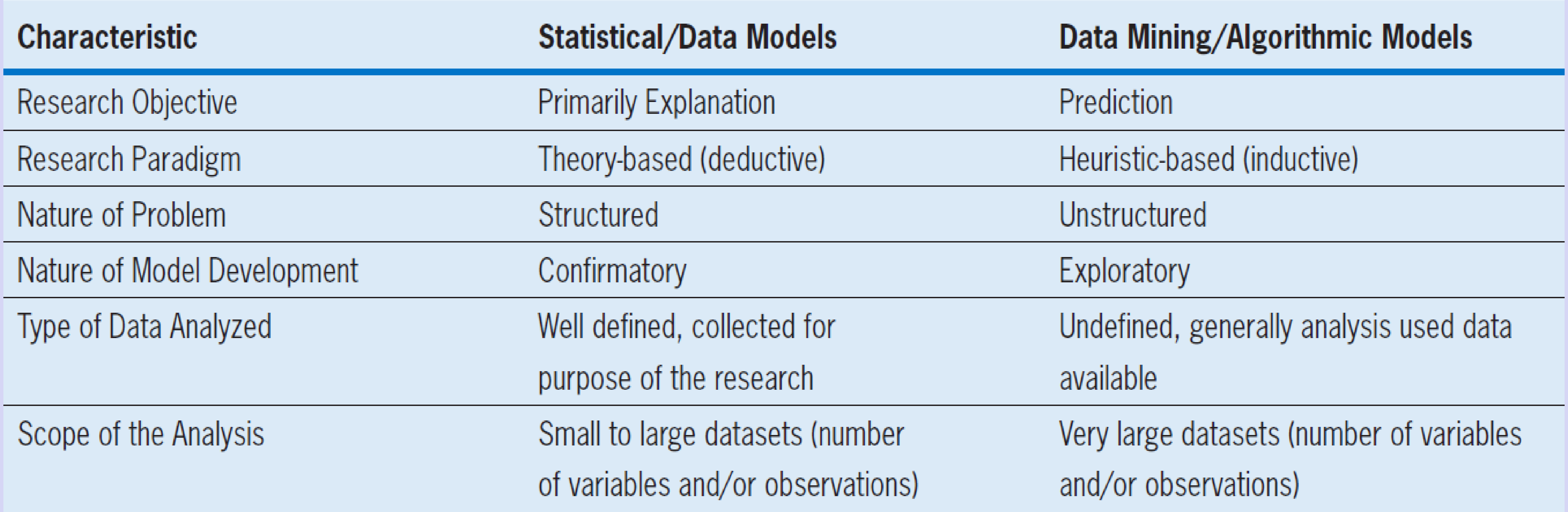
Topic 2: Statistical Versus Data Mining Models

Two fundamentally **different approaches to data analysis** (Breiman 2001):

- Statistical/data models – analysis where a specific model is proposed (e.g., dependent and independent variables to be analyzed by the general linear model), the model is then estimated and a statistical inference is made as to its generalizability to the population through statistical tests.
- Data mining/algorithmic models – models based on algorithms (e.g., neural networks, decision trees, support vector machine) that are widely used in many Big Data applications. Their emphasis is on predictive accuracy rather than statistical inference and explanation.

These two models truly **represent different "cultures" of model building**, coming from different research disciplines, operating on fundamentally different assumptions about how the models should operate and what are the most important objectives.

Comparing Between Statistical/Data Models and Data Mining/Algorithmic Models



Characteristic	Statistical/Data Models	Data Mining/Algorithmic Models
Research Objective	Primarily Explanation	Prediction
Research Paradigm	Theory-based (deductive)	Heuristic-based (inductive)
Nature of Problem	Structured	Unstructured
Nature of Model Development	Confirmatory	Exploratory
Type of Data Analyzed	Well defined, collected for purpose of the research	Undefined, generally analysis used data available
Scope of the Analysis	Small to large datasets (number of variables and/or observations)	Very large datasets (number of variables and/or observations)

- No “best” approach, each has strengths and weaknesses.
- Analysts today must assess each research situation and identify the best modeling approach for that specific situation (i.e., objective, data, etc.).

Topic 3: Causal Inference

Attempts to move beyond statistical inference to the **stronger statement of "cause and effect" in non-experimental situations**.

While causal statements have been primarily conceived as the domain of randomized controlled experiments, **recent developments** have provided researchers with . . .

- the theoretical frameworks for understanding the requirements for causal inferences in non-experimental settings.
- some techniques applicable to data not gathered in an experimental setting that still allow some causal inferences to be drawn.

Chapter 6 introduces **confounds and other "threats" to causal inference** and a popular methods for causal inference—**propensity score models**.

Multivariate Analysis in Statistical Terms

Defining Multivariate Analysis

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.

Many multivariate techniques are **extensions of univariate procedures**

- Simple regression → Multiple regression
- ANOVA → MANOVA

Many other techniques are **uniquely multivariate**

- Factor analysis, cluster analysis, discriminant analysis

Some Basic Concepts of Multivariate Analysis

- ❑ The Variate
- ❑ Measurement Scales
- ❑ Measurement Error and Multivariate Measurement

The Variate

The variate is a linear combination of variables with empirically determined weights.

- Weights are determined to best achieve the objective of the specific multivariate technique.
- Each respondent has a variate value (Y’).
- The Y’ value is a linear combination of the entire set of variables. It is the dependent variable.

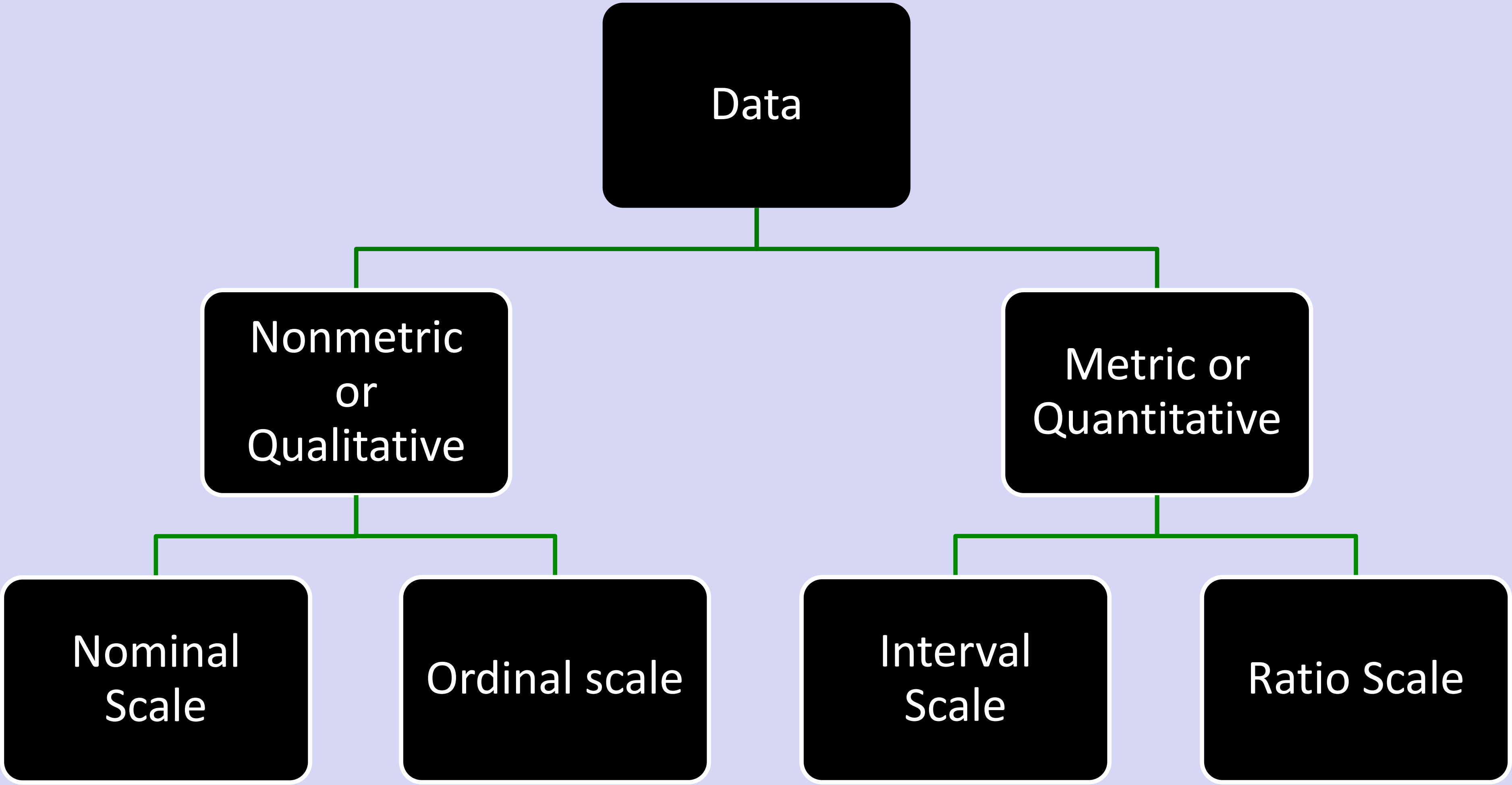
Variate equation: $(Y') = W_1 X_1 + W_2 X_2 + . . . + W_n X_n$

- W_n = empirically determined weights
- X_n = independent variables

Potential Independent Variables:

- X_1 = income
- X_2 = education
- X_3 = family size
- X_4 = ??

Types of Data and Measurement Scales



Data

Nonmetric or Qualitative

Metric or Quantitative

Nominal Scale

Ordinal scale

Interval Scale

Ratio Scale

Measurement Scales

Nonmetric

- Nominal – size of number is not related to the amount of the characteristic being measured.
- Ordinal – larger numbers indicate more (or less) of the characteristic measured, but not how much more (or less).

Metric

- Interval – contains ordinal properties, and in addition, there are equal differences between scale points.
- Ratio – contains interval scale properties, and in addition, there is a natural zero point.

NOTE: The level of measurement is critical in determining the appropriate multivariate technique to use!

Measurement Error

- All variables have **some error**. What are the sources of error?

- **Measurement error**
  - distorts observed relationships and makes multivariate techniques less powerful.

- Researchers use summated scales, for which several variables are summed or averaged together to form a composite representation of a concept.

Assessing Measurement Error

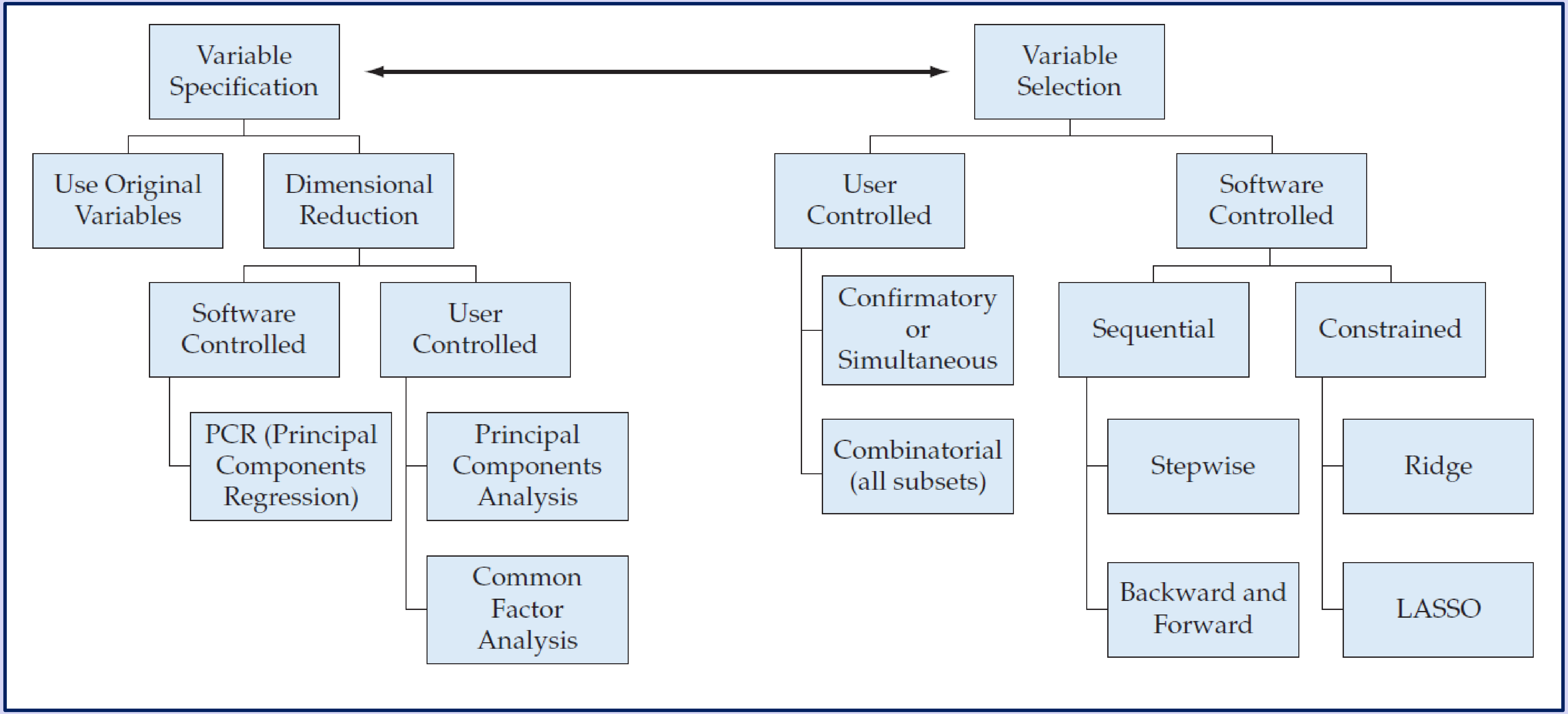
In addressing measurement error, researchers evaluate two important characteristics of measurement:

- **Validity**
  - the degree to which a measure accurately represents what it is supposed to.
- **Reliability**
  - the degree to which the observed variable measures the “true” value and is thus error free.

Managing the Multivariate Model

- ❑ Managing the Variate
- ❑ Managing the Dependence Model
- ❑ Statistical Significance Versus Statistical Power

Managing the Variate



Preference is always for user control, but a large number of variables may necessitate software control.

Variable Specification and Variable Selection

Variable Specification – preparing the variables for analysis

- Use original variables – retain most detailed attributes, but may suffer due to multicollinearity.
- Dimensional reduction – develop some form of composites of original variables
    - User controlled – Exploratory factor analysis (principal components or common factor analysis).
    - Software controlled – Principal components regression (PCR).

Variable Selection – identifying the variables included in the analysis

- User controlled – researcher explicitly defines variables in analysis.
    - Confirmatory or combinatorial
- Software controlled.
    - Sequential – subset of variables included based on algorithm (e.g., stepwise)
    - Constrained – regression weight constrained to force some to zero or very low, effectively eliminating those variables from the analysis (e.g., ridge or LASSO)

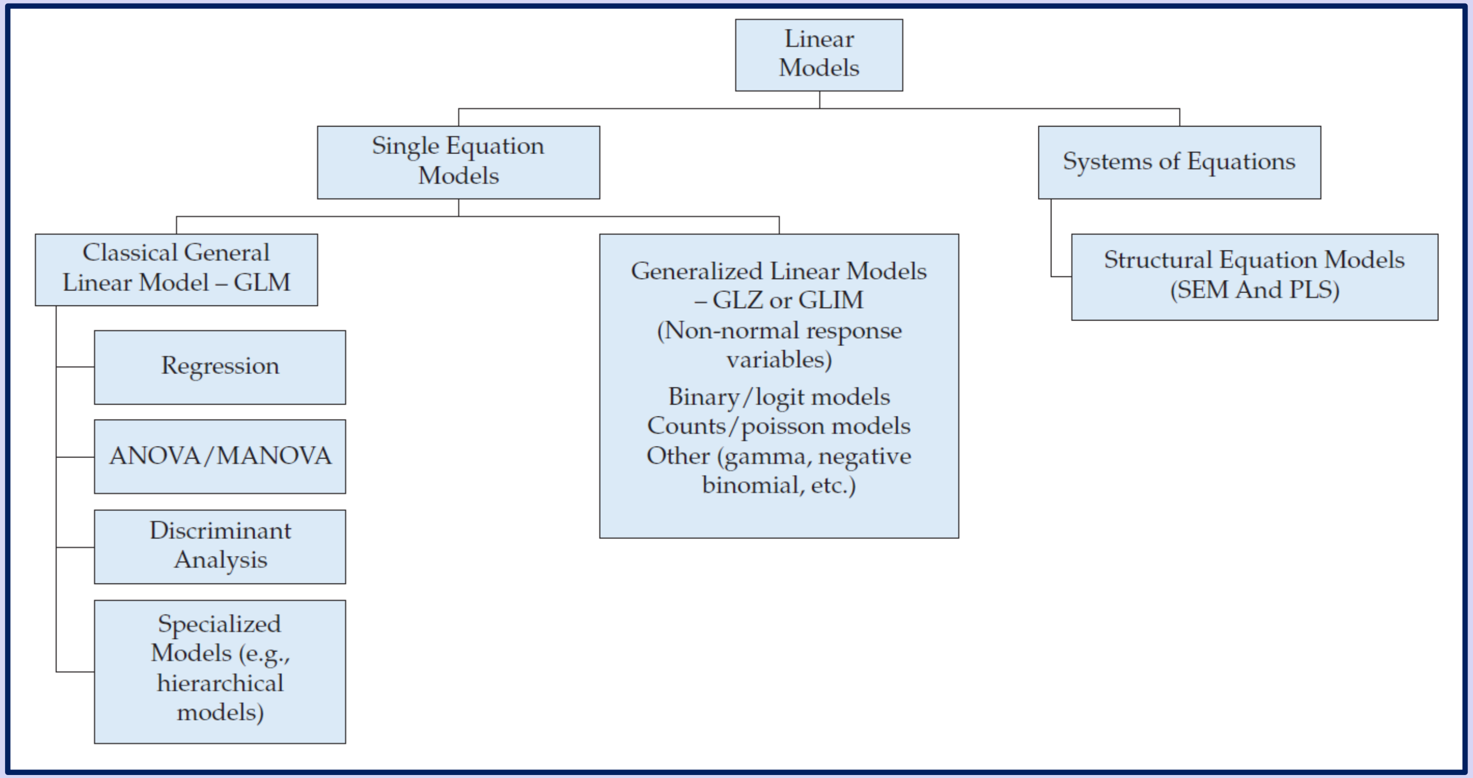
Managing the Dependence Model

Single Equation versus Multiple Equation Models

- <u>Single equation</u> – most widely used methods in the past
- <u>Multiple equation</u> – represents interrelated relationships

GLM versus GLZ/GLIM

- <u>GLM</u> – classical OLS-based system of techniques
- <u>GLZ/GLIM</u> – developed for non-normal response variables

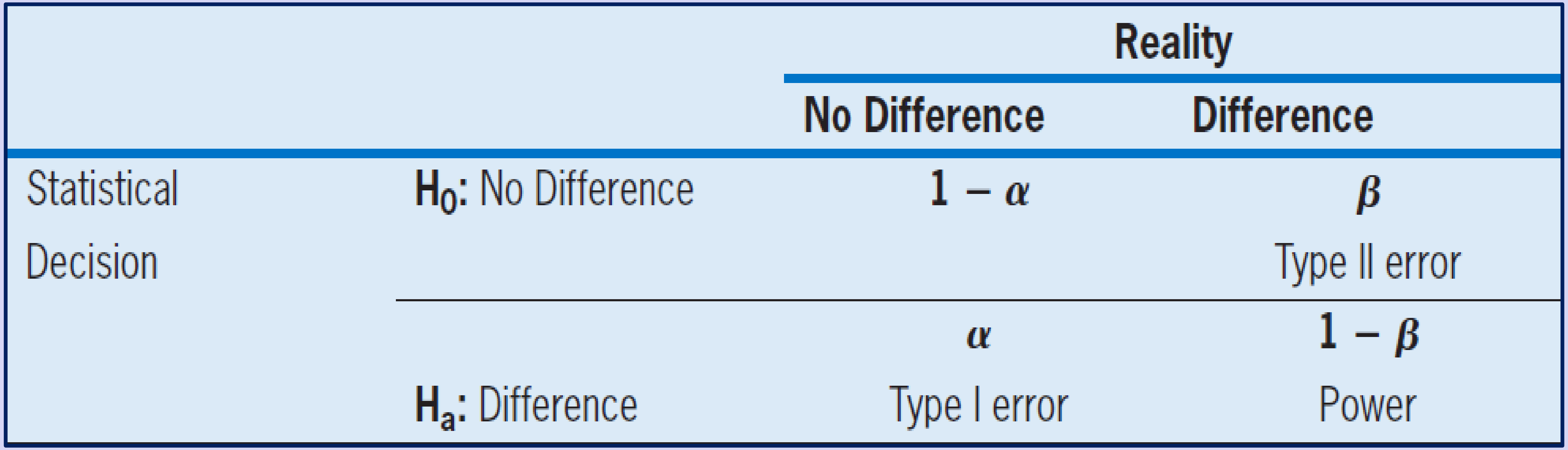

Linear Models
Single Equation Models
Systems of Equations
Classical General Linear Model – GLM
Regression
ANOVA/MANOVA
Discriminant Analysis
Specialized Models (e.g., hierarchical models)
Generalized Linear Models – GLZ or GLIM (Non-normal response variables)
Binary/logit models
Counts/poisson models
Other (gamma, negative binomial, etc.)
Structural Equation Models (SEM And PLS)

Statistical Significance and Power

Type I error (α) – probability of rejecting the null hypothesis when it is true.

Type II error (β) – probability of failing to reject the null hypothesis when it is false.

Power (1 – β) – probability of rejecting the null hypothesis when it is false.



		Reality	
		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

Power is Determined by Three Factors

Effect size:

- the actual magnitude of the effect of interest (e.g., the difference between means or the correlation between variables).

#The overall effect sizes $f^2 \geq .02$, .15, or .35 are regarded as small, moderate, and large effects, respectively.

$$Cohen's\ f^2 = \frac{R^2_{full} - R^2_{reduced}}{1 - R^2_{full}}$$

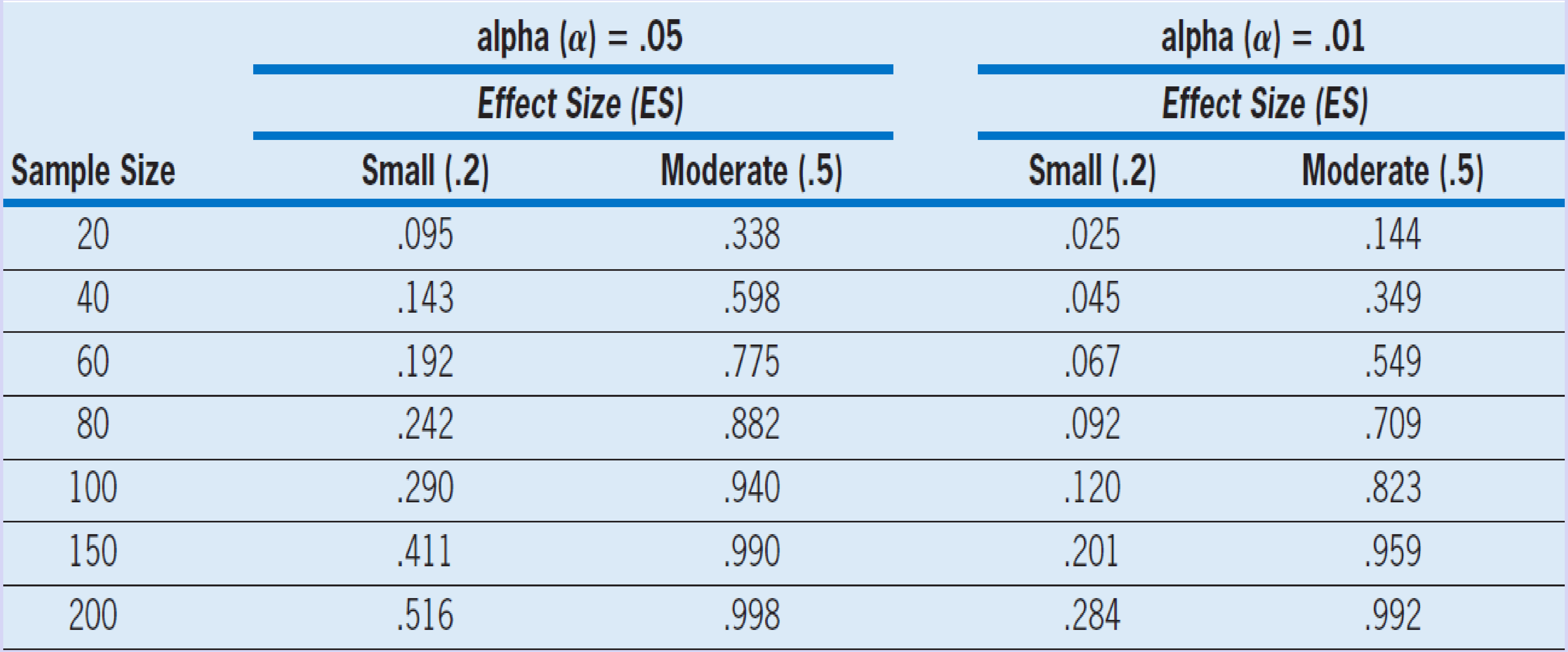
Alpha (α):

- as α is set at smaller levels, power decreases. Typically, α = .05.

Sample size:

- as sample size increases, power increases. With very large sample sizes (1,000+), even very small effects can be statistically significant, raising the issue of practical significance vs. statistical significance.

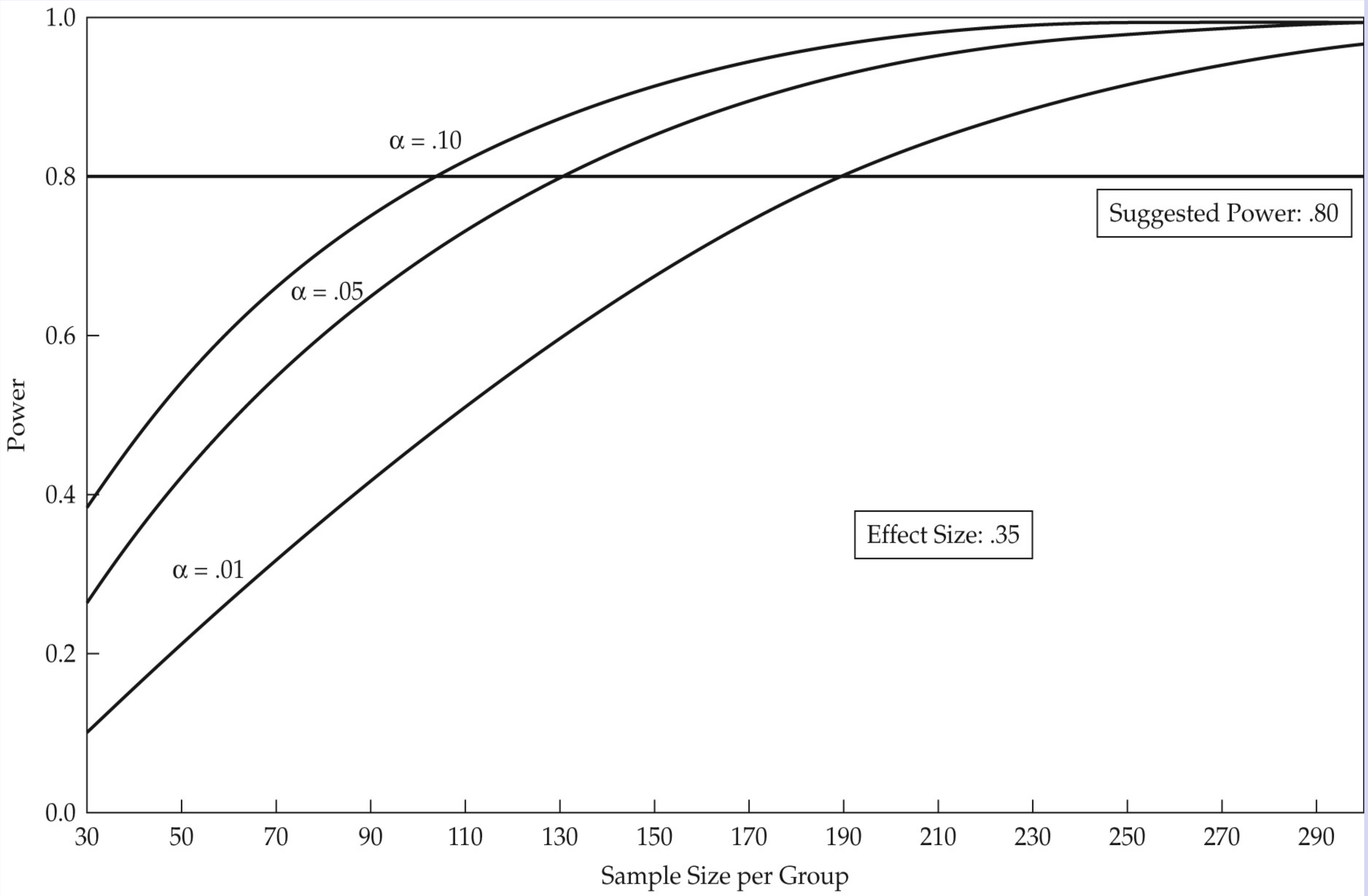
Power Levels for the Comparison of Two Means: Variations by Sample Size, Significance Level, and Effect Size



	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

Source: SOLO Power Analysis, BMDP Statistical Software, Inc.

Impact of Sample Size on Power



Rules of Thumb: Statistical Power Analysis

- Researchers should always design the study to achieve a power level of .80 at the desired significance level.
- More stringent significance levels (e.g., .01 instead of .05) require larger samples to achieve the desired power level.
- Conversely, power can be increased by choosing a less stringent alpha level (e.g., .10 instead of .05).
- Smaller effect sizes always require larger sample sizes to achieve the desired power.
- Any increase in power is most likely achieved by increased sample size.

A Classification of Multivariate Techniques

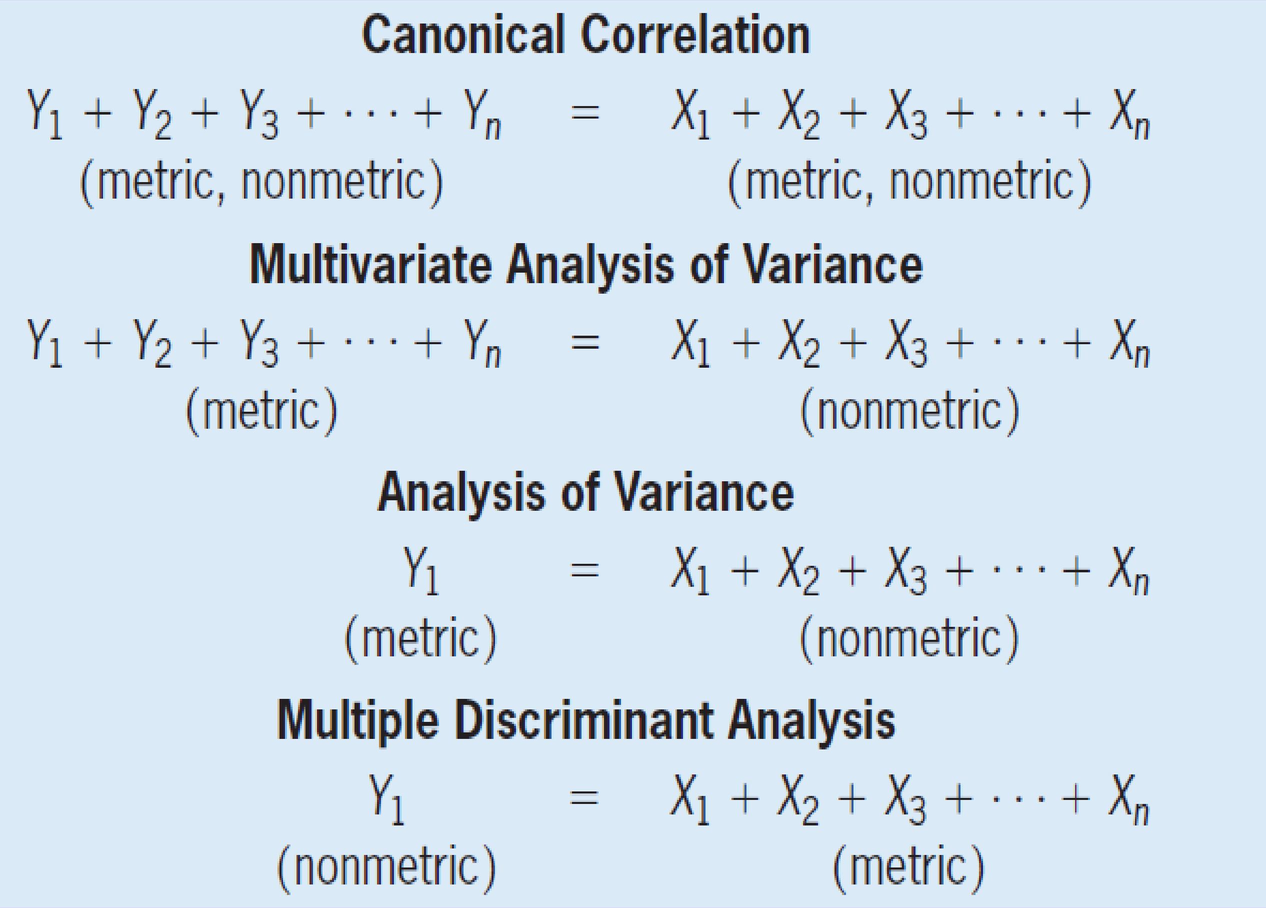
- ❑ Dependence Techniques
- ❑ Interdependence Techniques

Dependence Techniques

Dependence techniques: a variable or set of variables is identified as the dependent variable to be predicted or explained by other variables known as independent variables.

- Multiple Regression
- Multiple Discriminant Analysis
- Logit/Logistic Regression
- Multivariate Analysis of Variance (MANOVA) and Covariance
- Conjoint Analysis
- Canonical Correlation
- Structural Equations Modeling (SEM)
- Partial Least Squares (PLS) Modeling

The Relationship Between Multivariate Dependence Methods



Canonical Correlation

$$\begin{array}{ccc} Y_1 + Y_2 + Y_3 + \cdots + Y_n & = & X_1 + X_2 + X_3 + \cdots + X_n \\ \text{(metric, nonmetric)} & & \text{(metric, nonmetric)} \end{array}$$

Multivariate Analysis of Variance

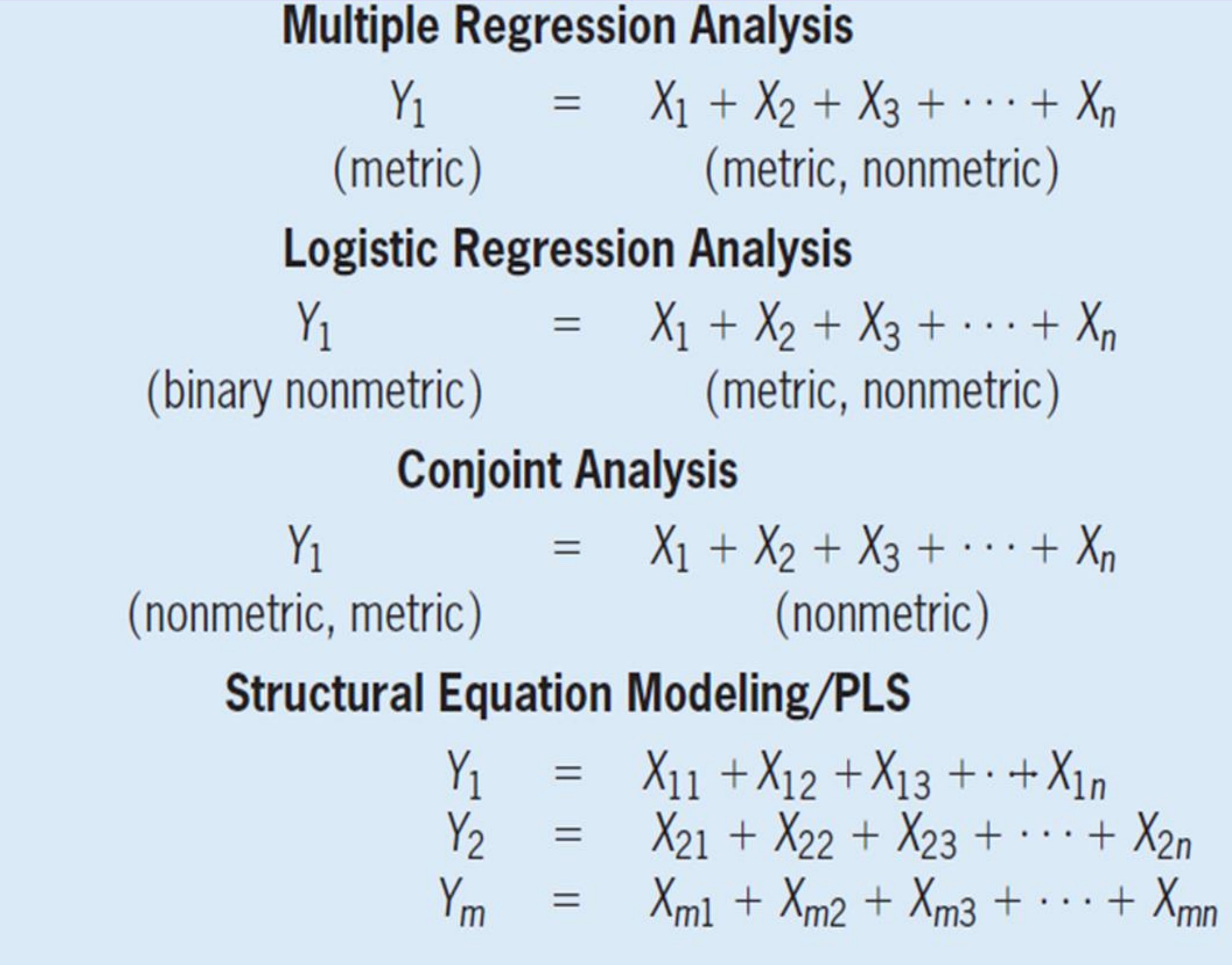
$$\begin{array}{ccc} Y_1 + Y_2 + Y_3 + \cdots + Y_n & = & X_1 + X_2 + X_3 + \cdots + X_n \\ \text{(metric)} & & \text{(nonmetric)} \end{array}$$

Analysis of Variance

$$\begin{array}{ccc} Y_1 & = & X_1 + X_2 + X_3 + \cdots + X_n \\ \text{(metric)} & & \text{(nonmetric)} \end{array}$$

Multiple Discriminant Analysis

$$\begin{array}{ccc} Y_1 & = & X_1 + X_2 + X_3 + \cdots + X_n \\ \text{(nonmetric)} & & \text{(metric)} \end{array}$$



Multiple Regression Analysis

$$\begin{array}{ccc} Y_1 & = & X_1 + X_2 + X_3 + \cdots + X_n \\ \text{(metric)} & & \text{(metric, nonmetric)} \end{array}$$

Logistic Regression Analysis

$$\begin{array}{ccc} Y_1 & = & X_1 + X_2 + X_3 + \cdots + X_n \\ \text{(binary nonmetric)} & & \text{(metric, nonmetric)} \end{array}$$

Conjoint Analysis

$$\begin{array}{ccc} Y_1 & = & X_1 + X_2 + X_3 + \cdots + X_n \\ \text{(nonmetric, metric)} & & \text{(nonmetric)} \end{array}$$

Structural Equation Modeling/PLS

$$\begin{array}{ccc} Y_1 & = & X_{11} + X_{12} + X_{13} + \cdot + X_{1n} \\ Y_2 & = & X_{21} + X_{22} + X_{23} + \cdots + X_{2n} \\ Y_m & = & X_{m1} + X_{m2} + X_{m3} + \cdots + X_{mn} \end{array}$$

Interdependence Techniques

Interdependence techniques: involve the simultaneous analysis of all variables in the set, without distinction between dependent variables and independent variables.

- Principal Components and Common Factor Analysis
- Cluster Analysis
- Multidimensional Scaling (perceptual mapping)
- Correspondence Analysis

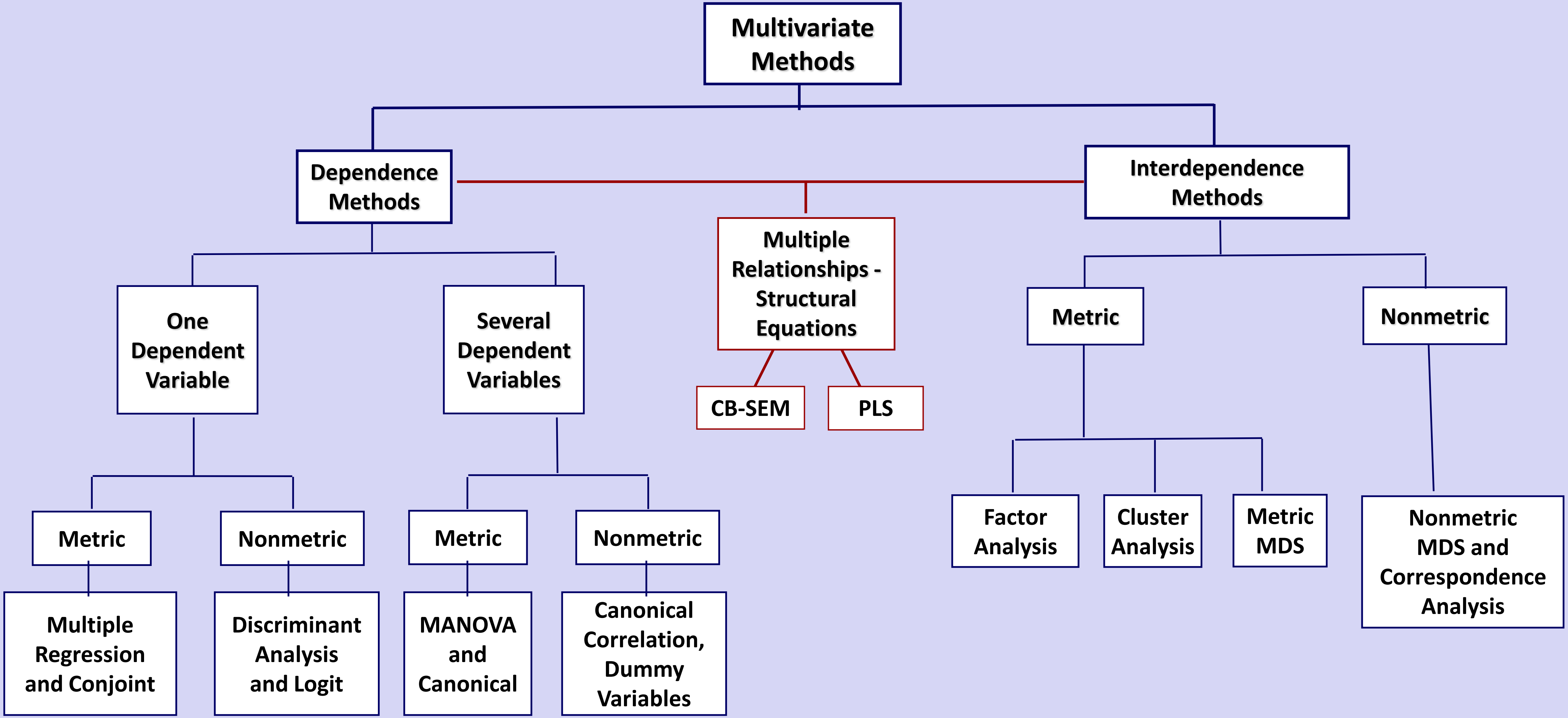
Selecting a Multivariate Technique

- What **type of relationship** is being examined – dependence or interdependence?

- **Dependence relationship**:
  - How many variables are being predicted?
  - What is the measurement scale of the dependent variable?
  - What is the measurement scale of the predictor variable (s)?

- **Interdependence relationship**:
  - Are you examining relationships between variables, respondents, or objects?

Selecting the Correct Multivariate Method



Multivariate Methods

Dependence Methods

Interdependence Methods

Multiple Relationships - Structural Equations

CB-SEM

PLS

One Dependent Variable

Several Dependent Variables

Metric

Nonmetric

Metric

Nonmetric

Metric

Nonmetric

Multiple Regression and Conjoint

Discriminant Analysis and Logit

MANOVA and Canonical

Canonical Correlation, Dummy Variables

Factor Analysis

Cluster Analysis

Metric MDS

Nonmetric MDS and Correspondence Analysis

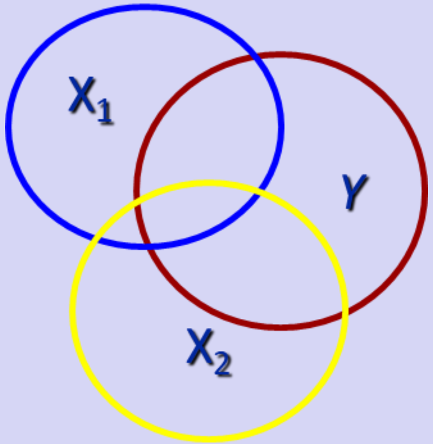
Types of Multivariate Techniques

Dependence Techniques

Multiple Regression

What is it?

. . . a **single** metric dependent variable is predicted by several metric independent variables.



MANOVA

What is it?

. . . **several** metric dependent variables are predicted by a set of nonmetric (categorical) independent variables.

Dependence Techniques

Discriminant Analysis

What is it?

. . . single, non-metric (categorical) dependent variable is predicted by several metric independent variables.

Why use it?

Examples of Dependent Variables:

- Gender – Male vs. Female
- Culture – USA vs. Outside USA
- Purchasers vs. Non-purchasers
- Member vs. Non-Member
- Good, Average and Poor Credit Risk

Logistic Regression

What is it?

. . . single nonmetric dependent variable is predicted by several metric independent variables.

This technique is similar to discriminant analysis, but relies on calculations more like regression.

Dependence Techniques

Canonical Analysis

What is it?

. . . several metric dependent variables are predicted by several metric independent variables.

Conjoint Analysis

What is it?

. . . quasi-experimental design based on attributes and levels of attributes which develops combinations of attributes/levels which are then evaluated by respondents.

Dependence Techniques

Structural Equation Modeling (SEM)

What is it?

. . . estimates multiple, interrelated dependence relationships based on two components:

a. Measurement Model
b. Structural Model

Two basic methodologies

CB-SEM – covariance-based SEM

PLS-SEM – Partial Least Squares Modeling (variance-based)

Interdependence Techniques

Exploratory Factor Analysis (EFA)

What is it?

. . . analyzes the structure of the interrelationships among a large number of variables to determine a set of common underlying dimensions (factors).

Cluster Analysis

What is it?

. . . . groups objects (respondents, products, firms, variables, etc.) so that each object is similar to the other objects in the cluster and different from objects in all the other clusters.

Interdependence Techniques

Multidimensional Scaling (MDS)

What is it?

. . . identifies "unrecognized" dimensions that affect purchase behavior based on customer judgments of

similarities or preferences

and transforms these into distances represented as perceptual maps.

Correspondence Analysis

What is it?

. . . uses non-metric data and evaluates either linear or non-linear relationships in an effort to develop a perceptual map representing the association between objects (firms, products, etc.) and a set of descriptive characteristics of the objects.

Guidelines for Multivariate Analyses and Interpretation

- ❑ Establish Practical Significance as well as Statistical Significance.
- ❑ Sample Size Affects All Results.
- ❑ Know Your Data.
- ❑ Strive for Model Parsimony.
- ❑ Look at Your Errors.
- ❑ Simplify Your Models By Separation
- ❑ Validate Your Results.

Guidelines for Multivariate Analysis

Establish Practical Significance as well as Statistical Significance.

- Practical significance asks the question, “So what?”
- Applicable to both managerial and academic contexts.

Sample Size Affects All Results.

- Small samples – too little statistical power or too easily “overfitting” the data.
- Large Samples – make the statistical tests overly sensitive.

Know Your Data.

- Multivariate analyses require an even more rigorous examination of the data because the influence of outliers, violations of assumptions, and missing data can be compounded across several variables to create substantial effects.

Guidelines for Multivariate Analysis

Strive for Model Parsimony.

- Irrelevant variables usually increase a technique's ability to fit the sample data, but at the expense of overfitting the sample data and making the results less generalizable to the population.
- Even though irrelevant variables typically do not bias the estimates of the relevant variables, they can mask the true effects due to an increase in multicollinearity.

Look at Your Errors.

- starting point for diagnosing the validity of the obtained results.
- indication of the remaining unexplained relationships.

Simplify Your Models By Separation

- Estimate separate models when possible (e.g., in presence of moderators).

Validate Your Results.

- Use split-sample or cross-validation to assess generalizability of any model.

A Structured Approach to Multivariate Model Building

Stage 1: Define the Research Problem, Objectives, and Multivariate Technique(s) to be Used

Stage 2: Develop the Analysis Plan

Stage 3: Evaluate the Assumptions Underlying the Multivariate Technique(s)

Stage 4: Estimate the Multivariate Model and Assess Overall Model Fit

Stage 5: Interpret the Variate(s)

Stage 6: Validate the Multivariate Model

The Challenge of Big Data Research Efforts

- ❑ Data Management
- ❑ Data Quality

Era of Big Data and Data Examination

Data Management

- Many consider most daunting challenge
- Many times majority of research effort expended in this task
- Complexity arises from . . .
  - Merging disparate sources of data
  - Use of unstructured data

Data Quality

- True “value” of analysis may rest in data quality
- Conceptualized in eight dimensions
- Many times “hidden” in basic nature of the data (e.g., binary measures)

Preliminary Examination of the Data

- ❑ Univariate Profiling: Examining the Shape of the Distribution
- ❑ Bivariate Profiling: Examining the Relationship Between Variables
- ❑ Bivariate Profiling: Examining Group Differences
- ❑ Multivariate Profiles
- ❑ New Measures of Association

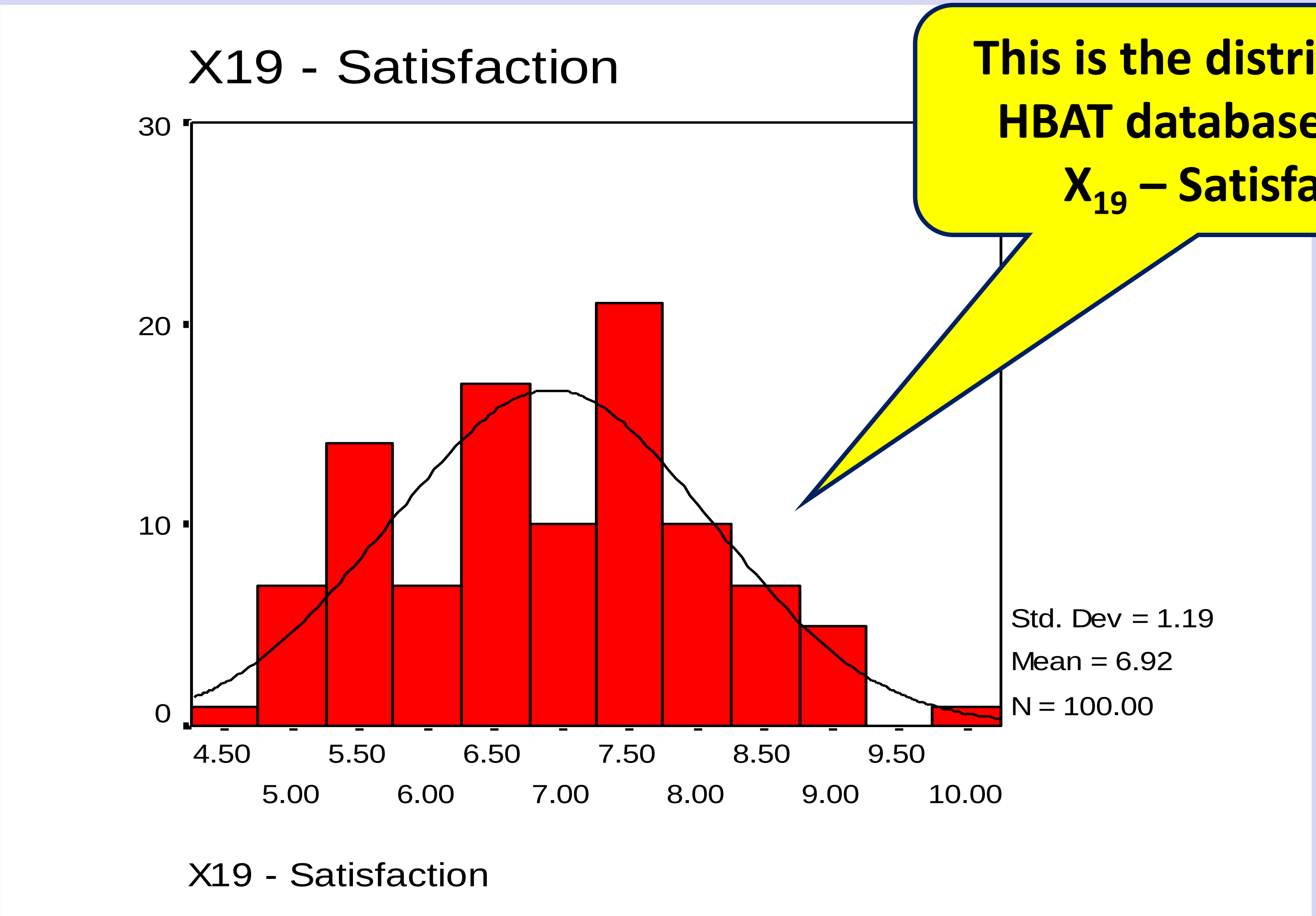
Graphical Examination

Fundamental tool in data examination is graphical examination

Some basic **types** of graphical tools:

- Shape:
  - Histogram
  - Bar Chart
  - Box & Whisker plot
  - Stem and Leaf plot
- Relationships:
  - Scatterplot
  - Outliers

Univariate Profiling: Histograms and The Normal Curve



This is the distribution for HBAT database variable X_{19} – Satisfaction.

Univariate Profile: Stem & Leaf Diagram – HBAT Variable X_6

```
X6 - Product Quality
          Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

    3.00       5 .  012
   10.00       5 .  5567777899
   10.00       6 .  0112344444
   10.00       6 .  5567777999
    5.00       7 .  01144
   11.00       7 .  55666777899
    9.00       8 .  000122234
   14.00       8 .  55556667777778
   18.00       9 .  001111222333333444
    8.00       9 .  56699999
    2.00      10 .  00

 Stem width:    1.0
 Each leaf:     1 case(s)
```

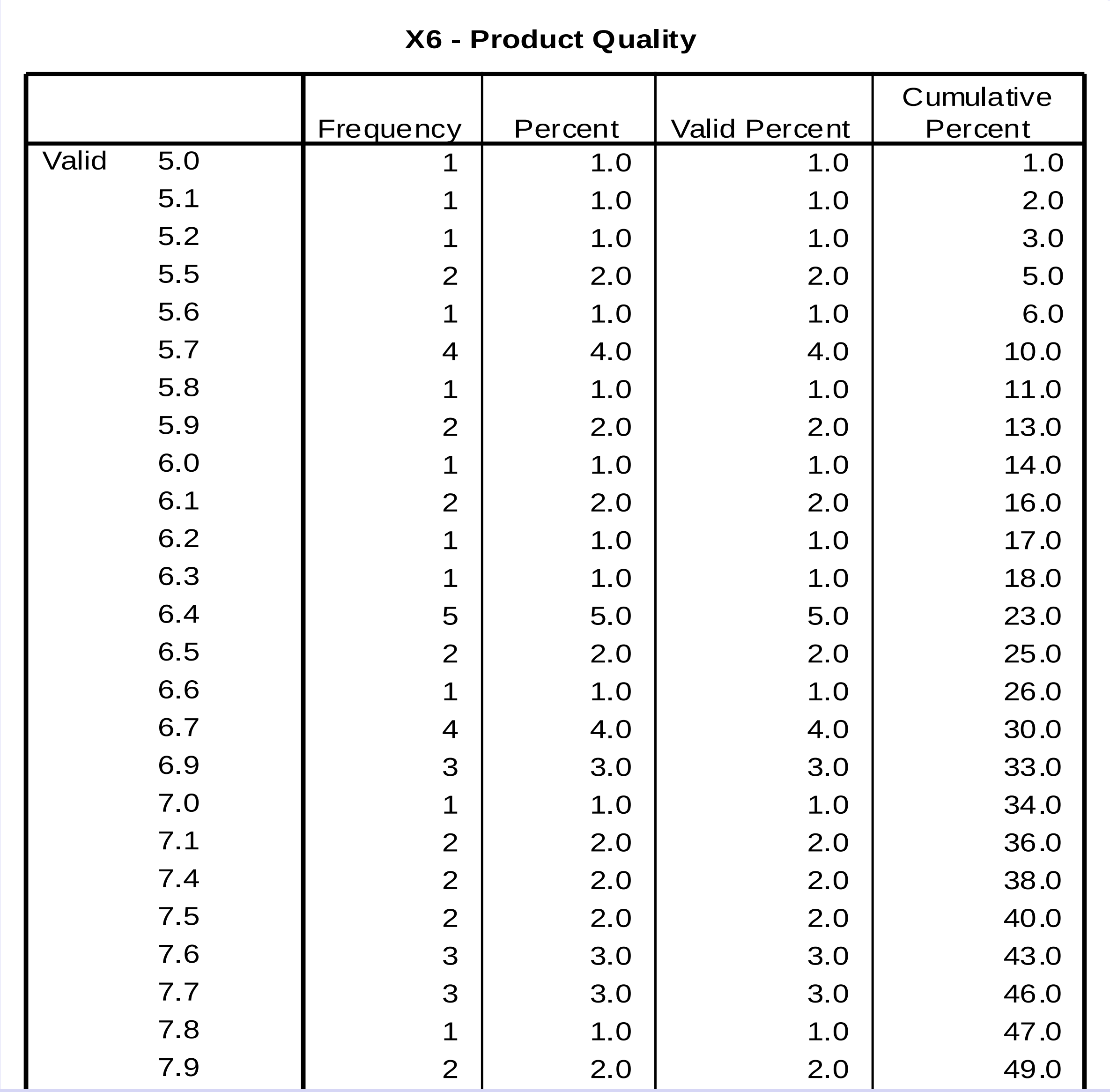
Each stem is shown by the numbers (in this case from 5 to 10), and each number is a leaf. This stem has 10 leaves.

The length of the stem, indicated by the number of leaves, shows the frequency distribution. For this stem, the frequency is 14, representing values ranging from 8.5 to 8.8.

This table shows the distribution of X_6 with a stem and leaf diagram.

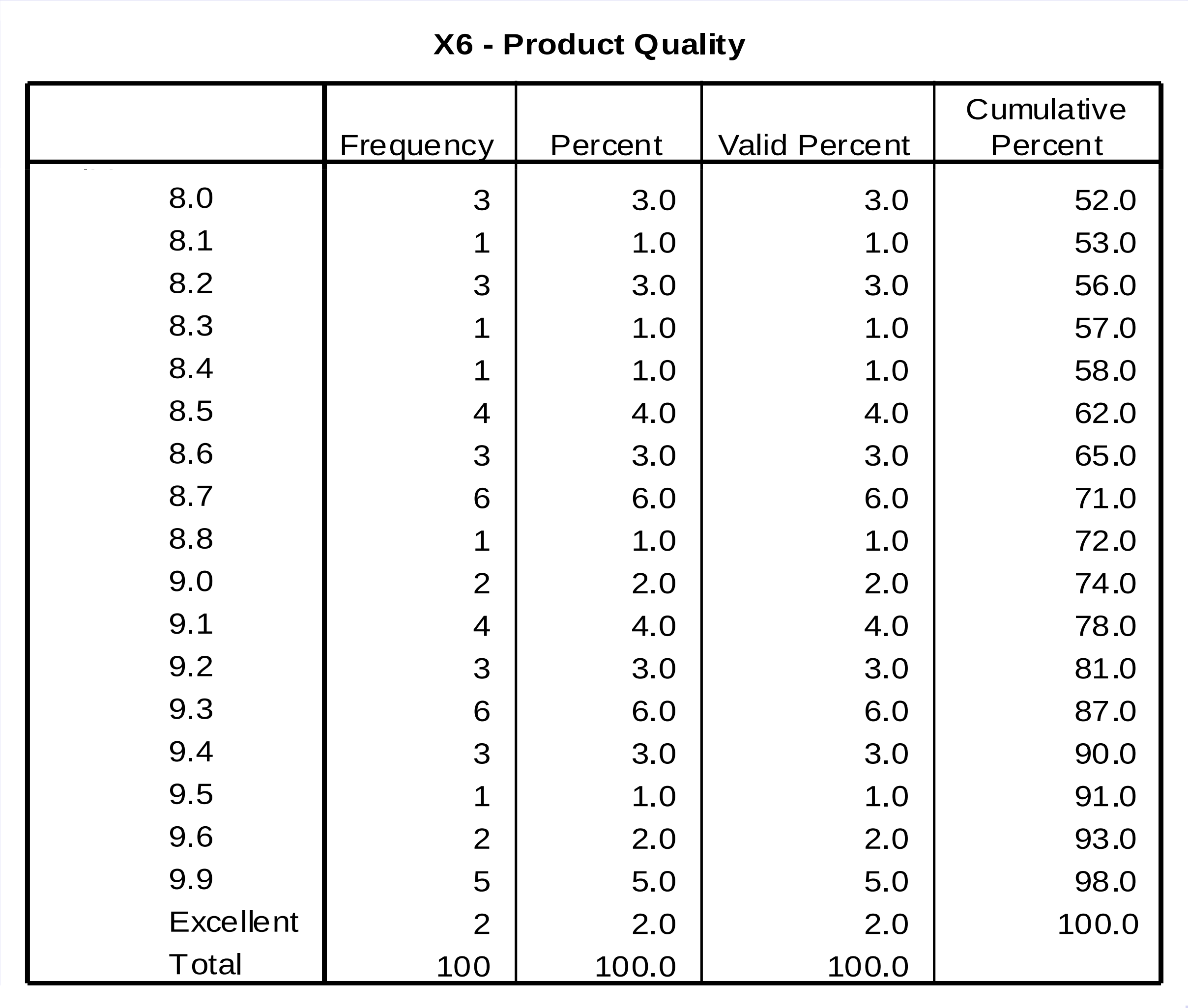
- The first category is from 5.0 to 5.4, thus the stem is 5.0. There are three observations with values in this range (5.0, 5.1 and 5.2). This is shown as three leaves of 0, 1 and 2. These are also the three lowest values for X6.
- In the next stem, the stem value is again 5.0 and there are ten observations, ranging from 5.5 to 5.9. These correspond to the leaves of 5.5 to 5. 9.
- At the other end of the figure, the stem is 10.0. It is associated with two leaves (0 and 0), representing two values of10.0, the two highest values for X_6.

Univariate Profile: Frequency Distribution: Variable X_6



X6 - Product Quality

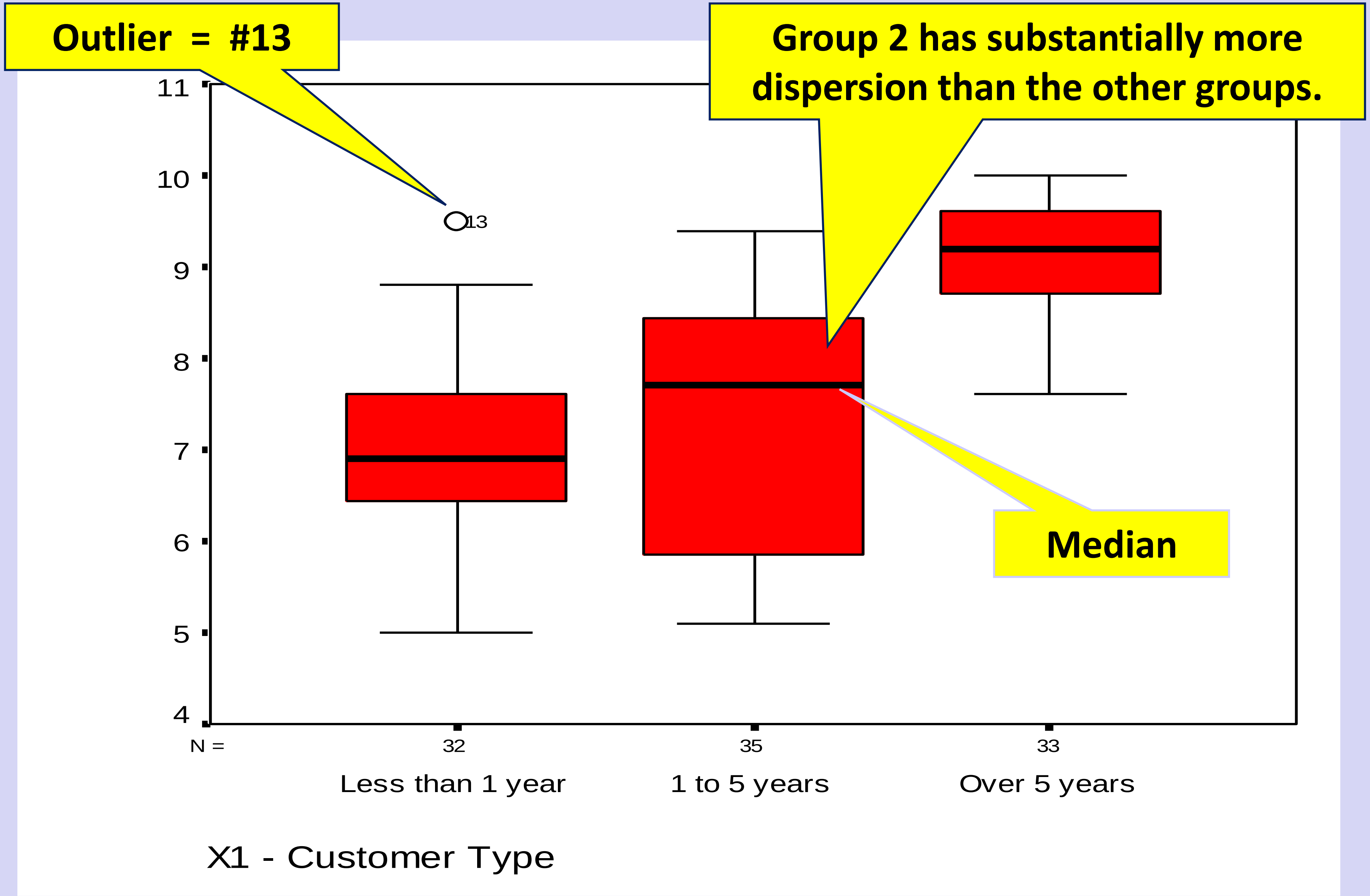
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5.0	1	1.0	1.0	1.0
	5.1	1	1.0	1.0	2.0
	5.2	1	1.0	1.0	3.0
	5.5	2	2.0	2.0	5.0
	5.6	1	1.0	1.0	6.0
	5.7	4	4.0	4.0	10.0
	5.8	1	1.0	1.0	11.0
	5.9	2	2.0	2.0	13.0
	6.0	1	1.0	1.0	14.0
	6.1	2	2.0	2.0	16.0
	6.2	1	1.0	1.0	17.0
	6.3	1	1.0	1.0	18.0
	6.4	5	5.0	5.0	23.0
	6.5	2	2.0	2.0	25.0
	6.6	1	1.0	1.0	26.0
	6.7	4	4.0	4.0	30.0
	6.9	3	3.0	3.0	33.0
	7.0	1	1.0	1.0	34.0
	7.1	2	2.0	2.0	36.0
	7.4	2	2.0	2.0	38.0
	7.5	2	2.0	2.0	40.0
	7.6	3	3.0	3.0	43.0
	7.7	3	3.0	3.0	46.0
	7.8	1	1.0	1.0	47.0
	7.9	2	2.0	2.0	49.0



X6 - Product Quality

	Frequency	Percent	Valid Percent	Cumulative Percent
8.0	3	3.0	3.0	52.0
8.1	1	1.0	1.0	53.0
8.2	3	3.0	3.0	56.0
8.3	1	1.0	1.0	57.0
8.4	1	1.0	1.0	58.0
8.5	4	4.0	4.0	62.0
8.6	3	3.0	3.0	65.0
8.7	6	6.0	6.0	71.0
8.8	1	1.0	1.0	72.0
9.0	2	2.0	2.0	74.0
9.1	4	4.0	4.0	78.0
9.2	3	3.0	3.0	81.0
9.3	6	6.0	6.0	87.0
9.4	3	3.0	3.0	90.0
9.5	1	1.0	1.0	91.0
9.6	2	2.0	2.0	93.0
9.9	5	5.0	5.0	98.0
Excellent	2	2.0	2.0	100.0
Total	100	100.0	100.0	

Bivariate Profile: Box & Whiskers Plots

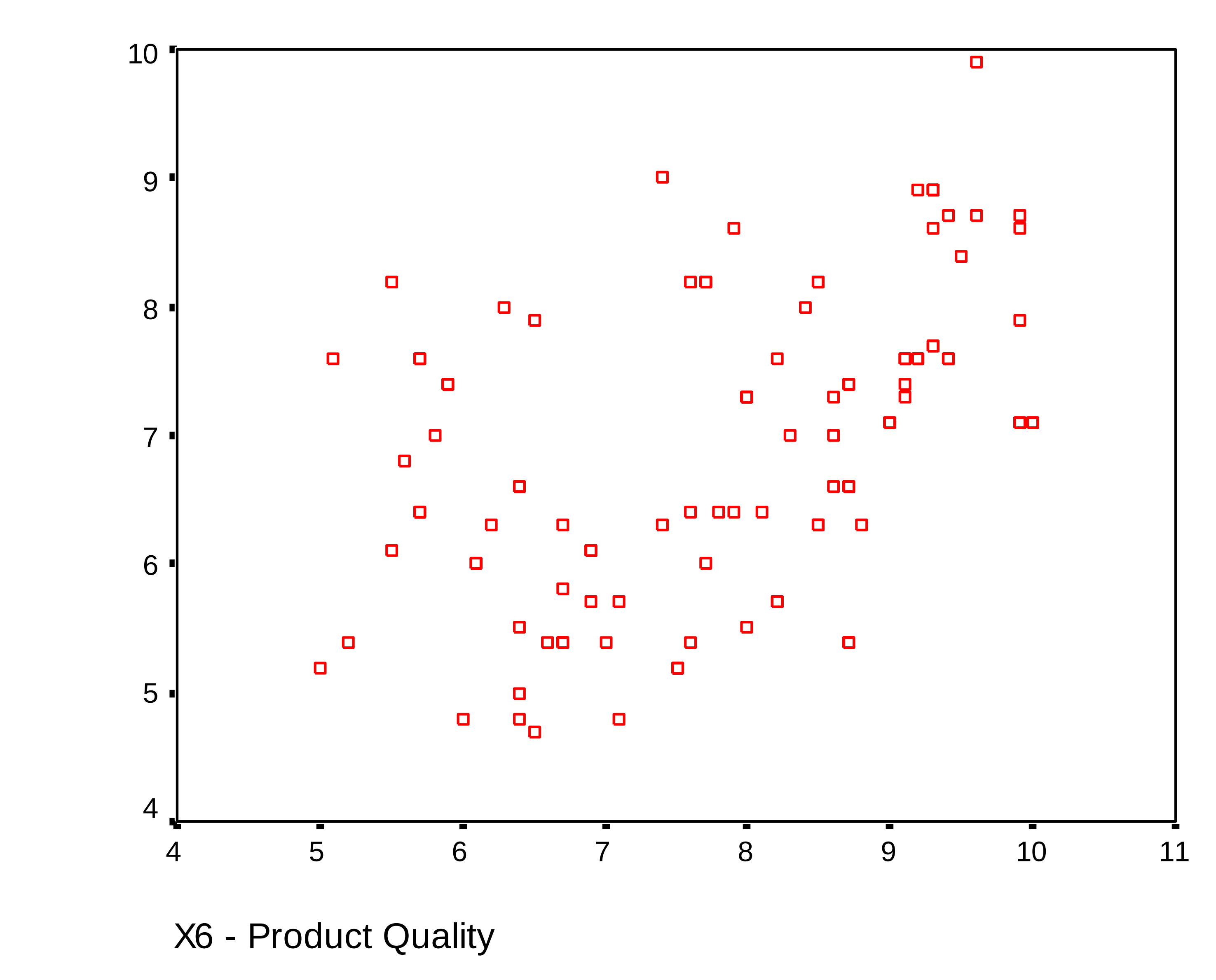


Outlier = #13

Group 2 has substantially more dispersion than the other groups.

Median

Bivariate profile: HBAT Scatterplot -- Variables X_{19} and X_6



New Measures of Association

Wide range of existing measures beyond Pearson correlation (parametric and non-parametric).

New measures from data mining:

- Hoeffding’s D
  - nonparametric measure of association based on departures from independence.
- dCor (the distance correlation)
  - distance-based measure of association which also is more sensitive to nonlinear patterns in the data.
- MIC (mutual information correlation)
  - Pattern-matching approach amenable to identifying both non-linear relationships as well as a range of distinct patterns.

Missing Data

- ❑ The Impact of Missing Data
- ❑ Recent Developments in Missing Data Analysis
- ❑ A Four-Step Process for Identifying Missing Data and Applying Remedies
- ❑ An Illustration of Missing Data Diagnosis with the Four-Step Process

Missing Data

- **Missing Data:** information not available for a subject (or case) about whom other information is available. Typically occurs when respondent fails to answer one or more questions in a survey.
  - ✓ Systematic?
  - ✓ Random?
- **Researcher's Concern:** to identify the patterns and relationships underlying the missing data in order to maintain as close as possible to the original distribution of values when any remedy is applied.

- **Impact** . . .
  - ✓ Reduces sample size available for analysis.
  - ✓ Can distort results.

Recent Developments in Missing Data Analysis

Two major issues experiencing a **resurgence of interest**.

- Wide range of data sources now being used in analysis.
- Expanded availability and improved usability of model-based methods of imputation.

Corresponding increase in:

- Study and applications across wide range of disciplines.
- Model-based methods availability in all major software.

Four-Step Process for Identifying Missing Data

Step 1: Determine the Type of Missing Data

Step 2: Determine the Extent of Missing Data

Step 3: Diagnose the Randomness of the Missing Data Processes

Step 4: Select the Imputation Method

Step 1: Determine the Type of Missing Data

Ignorable Missing Data – expected and part of research design.

- Sample – form of missing data, with excluded data the remaining population.
- Part of data collection – e.g., skip patterns.
- Censored data – some data not yet observed (e.g., survival data).

Not Ignorable Missing Data – data which must be addressed in the analysis.

- Known process – identified due to procedural factors (e.g., data entry or data management).
- Unknown process – primarily related to respondent, but important characteristic is level of randomness (e.g., straight lining; lack of attention).

Levels of Missingness

Three levels of missingness (Newman 2014):

- Item-level – missing data for individual variable.
- Construct-level – missing data for entire set of questions about a specific construct.
- Person-level – missing data related to individual’s willingness or ability to provide responses.

Step 2: Determine the Extent of Missing Data

Basic question: Is the extent or amount of missing data is low enough to not affect the results, even if it operates in a nonrandom manner.

Levels of analysis: percentage of data missing by . . .

- Variable – common form of assessment.
- Case – amount of missing data across all variables by case.

Guidelines for deleting variables and/or cases:

- <u>10 percent or less generally acceptable</u> – cases or observations with 10% or less are amenable to any imputation strategy.
- <u>Sufficient sample size</u> – be sure missing data remedy provides adequate sample size.
- Cases with missing data for dependent variable(s) typically are deleted.
- <u>When deleting a variable, ensure that alternative variables</u>, hopefully highly correlated, are available to represent the intent of the original variable.
- Perform the analysis both <u>with and without the deleted cases or variables</u>.

Step 3: Diagnose the Randomness of the Missing Data Processes

Levels of Randomness of the Missing Data Process

- Missing Data at Random (MAR)
  - missing values of Y depend on X, but not on Y.
  - Example – observed Y values represent a random sample of the actual Y values for each value of X, but the observed data for Y do not necessarily represent a truly random sample of all Y values.
- Missing Completely at Random (MCAR)
  - observed values of Y are truly a random sample of all Y values.
  - no underlying association to the other observed variables, characterized as “purely haphazard missingness”.
- Not Missing at Random (NMAR)
  - Distinct non-random pattern of missing data.
  - Non-random pattern not related to any other variables.
  - Example: all individuals with high income had missing data.

Diagnostic Tests for Levels of Randomness

t test of Missingness

- Test of differences between cases with missing data versus not missing data on other variables:
    1. For specific variable (e.g., X_1), create two groups of cases – cases with missing values on X_1 and those with valid values on X_1
    2. Compare these two groups with a *t* test for differences on other variables in the analysis (e.g., X_2, X_3 ….)
    3. Differences indicate MAR processes, no differences indicate MCAR processes

Little's MCAR Test

- analyzes the pattern of missing data on all variables and compares it with the pattern expected for a random missing data process.
- If no significant differences are found, the missing data can be classified as MCAR.

MAR or MCAR?

- Useful for selecting remedy, but less impactful when using model-based methods.

Imputation of MCAR Using Only Valid Data

IF MCAR, several approaches available:

1. Using only valid data
   - Complete case approach
     - Use only cases with no missing data.
   - Using all-available data
     - Calculate imputed values based on all valid pairwise information.
2. Using known replacement data
   - Hot or Cold Deck imputation
   - Case substitution
3. Calculating replacement values
   - Mean substitution
     - replaces the missing values with the mean value of that variable calculated from all valid responses.
   - Regression imputation
     - predict the missing values of a variable based on its relationship to other variables in the dataset.

Imputation of a MAR Missing Data Process

Best remedy is some form of **model-based approach** . . .

Two forms of model-based imputation which rely upon MAR relationships to estimate missing data:

- Maximum likelihood and EM:
  - Single step process of missing data estimation and model estimation.
  - No imputation for individual cases, rather direct estimation of means and covariance matrix.
- Multiple imputation:
  - Estimation of imputed values for missing data of individual cases by specified model.
  - Calculates multiple sets of imputed values, each set varying by adding a random element to imputed values and then forming a separate dataset for estimation.
  - Model estimates made for each imputed dataset and then combined for final model estimates.

Choosing between maximum likelihood and multiple imputation:

- Multiple imputation uses conventional techniques for model estimation while maximum likelihood limited in applicable methods.

Choosing Imputation Based on:

Extent of missing data:

- Under 10% – Any of the imputation methods can be applied, complete case method has been shown to be the least preferred.
- 10% to 20% – all-available, hot deck case substitution, and regression methods most preferred for MCAR data, whereas model-based methods are necessary with MAR missing data processes.
- Over 20% – if necessary, the preferred methods are:
  - The regression method for MCAR situations.
  - Model-based methods when MAR missing data occur.

Type of missing data process:

- MCAR – any imputation method can provide unbiased estimates if MCAR conditions met, but the model-based methods are preferred.
- MAR – only model-based methods.

Outliers

- ❑ Two Different Contexts for Defining Outliers
- ❑ Impacts of Outliers
- ❑ Classifying Outliers
- ❑ Detecting and Handling Outliers

Outlier

An observation/response with a unique combination of characteristics and identifiable as **distinctly different** from the other observations/responses.

Issue: "Is the observation/response representative of the population?"

Contexts for defining outliers:

- Pre-analysis Context: A Member of a Population.
  - focus is on each case as compared to the other observations under study.
- Post-analysis Context: Meeting Analysis Expectations.
  - defines "normal" as the expectations (e.g., predicted values, group membership predictions, etc.) generated by the analysis of interest.

Impacts of Outliers

Practical Impacts

- Can have substantial impact on the results of any analysis.

Substantive Impacts

- Non-representative outliers can distort results and lame them less generalizable to the population.

Outliers– Good or Bad?

- Good – identify perhaps, small, but unique, portions of the sample that should be included.
- Bad – distort results and impact generalizability.
- Which one – depends on context and objectives of the research.

Classifying Outliers

Types of impacts of outliers

- Error outliers – differ from expected values generated by the analysis.
- Interesting outliers – different enough to generate insight into the analysis.
- Influential outliers – different enough to substantively impact the results.

Reasons for Outlier Designation

- Procedural Error.
- Extraordinary Event.
- Extraordinary Observations.
- Observations unique in their combination of values.

Detecting Outliers

- Standardize data and then identify outliers in terms of number of standard deviations.
- Examine data using Box Plots, Stem & Leaf, and Scatterplots.
- Multivariate detection (Mahalanobis D^2).

Detecting Outliers

Univariate methods – examine all metric variables to identify unique or extreme observations.

- For small samples (80 or fewer observations), outliers typically are defined as cases with standard scores of 2.5 or greater.
- For larger sample sizes, increase the threshold value of standard scores up to 4.
- If standard scores are not used, identify cases falling outside the ranges of 2.5 versus 4 standard deviations, depending on the sample size.

Bivariate methods – focus their use on specific variable relationships, such as the independent versus dependent variables:

- use scatterplots with confidence intervals at a specified Alpha level.

Multivariate methods – best suited for examining a complete variate, such as the independent variables in regression or the variables in factor analysis:

- threshold levels for the D^2/df measure should be very conservative (.005 or .001), resulting in values of 2.5 (small samples) versus 3 or 4 in larger samples.

Impact of Dimensionality

Increased dimensionality (i.e., increased number of variables) dramatically impacts outlier detection and designation in three ways:

1. Distance measures become less useful
   - higher levels of dimensionality create a "natural" dispersion among observations that makes distance measures less useful for identifying observations.

2. Impact of irrelevant variables
   - as dimensionality increases, the presence of irrelevant variables has higher likelihood, thus confounding the ability to identify outliers.

3. Compatibility of dimensions
   - as dimensionality increases through the use of multiple sources of data, especially unstructured data, methods for assessing comparability among observations becomes more difficult.

Dealing with Outliers

Outlier Designation

- Researcher judgment should guide designation of outliers versus a strictly empirical designation.

Outlier Description and Profiling

- Outliers should be described on the variables used to compare between observations.
- Profiles on additional variables should be generated when possible to provide more insight into the character of outliers.

Retention versus Deletion

- Should be retained unless demonstrable proof indicates that they are truly aberrant and not representative of any observations in the population.
- If possible generate results with and without outliers to assess impact.
- Methods to minimize outlier influence (e.g., robust methods) are available.

Testing the Assumptions of Multivariate Analysis

❑ Assessing Individual Variables Versus the Variate.

❑ Four Important Statistical Assumptions.

Need For Testing of Assumptions

F**oundation** for making **statistical inferences and results**.

Need is increased in multivariate analysis **because the complexity** of the analysis:

1. Makes the potential distortions and biases more potent when the assumptions are violated.
2. May mask the indicators of assumption violations apparent in the simpler univariate analyses.

Important Note: **Must test for assumptions twice** –

- Individual variables – to understand basic sources of problems.
- Variate – to assess the combined effect across all variables.

Four Important Statistical Assumptions

Normality

- Comparison of distribution to normal distribution.
- Basis for statistical inference from sample to population.

Homoscedasticity

- Variance of the error terms appears constant over a range of predictor variables.
- Heteroscedasticity is when error terms have increasing or modulating variance.
- Analysis of residuals best illustrates this point.

Linearity

- Relationship represented by a straight line (i.e., constant unit change (slope) of the dependent variable for a constant unit change of the independent variable.

Non-correlated Errors

- Prediction errors are uncorrelated with each other.

Normality Assumptions . . .

Univariate versus Multivariate Normality

- Univariate normality – each individual variable.
- Multivariate normality – combinations of variables.

Impacts of Assumption Violations

- Shape of Distribution – skewness versus kurtosis.
- Impact of sample size – increased sample size reduces detrimental effects.

Testing for Normality Assumptions

- Visual check of histogram or normal probability plot.
- Statistical tests of skewness and kurtosis.

Remedies

- Most often some form of data transformation.

Homoscedasticity Assumption

Impact of Heteroscedasticity – inflates/deflates standard errors

Sources of heteroscedasticty:

- Variable type – common in percentages or proportions.
- Skewed distribution – one or both variables.

Tests for homoscedasticity:

- Graphical test.
- Statistical tests.
  - Levene test (univariate)
  - Box’s M (multivariate)

Remedies for heteroscedasticity:

- Transformation of variable(s).
- Use of heteroscedasticity-consistent standard errors (HCSE).

Linearity and Absence of Correlated Errors Assumptions

Nonlinear relationships:

- can be very well defined, but seriously understated unless:
  - data is transformed to a linear pattern, or
  - explicit model components are used to represent the nonlinear portion of the relationship.

Correlated errors:

- arise from a process that must be treated much like missing data:
  - Researcher must first identify and define the “causes” among variables, either internal or external to the dataset (e.g., grouping or time series).
  - If they are not found and remedied, serious biases can occur in the results, many times unknown to the researcher.
- Remedies:
  - Inclusion of omitted causal factor underlying correlation of errors.
  - Apply specialized model forms (e.g., multi-level linear models, see Chapter 5).

Transformations

- ❑ Transformations Related to Statistical Properties
- ❑ Transformations Related to Interpretation
- ❑ Transformations Related to Specific Relationship Types
- ❑ Transformations Related to Simplification
- ❑ General Guidelines for Transformations

Data Transformations

Provide a means of modifying variables for **one of four reasons**:

1. Enhancing statistical properties.
   - Primarily to achieve normality, homoscedasticity or linearity.
2. Ease of interpretation.
   - Standardization – performed across cases to provide common metric for comparison
   - Centering – performed within-case to allow for comparison across variables.
3. Representing specific relationship types.
   - Transformed variables represent unique relationships – (e.g., elasticity).
4. Simplification.
   - Binning – categorization of values into a smaller number of categories (i.e., reduce cardinality).
     - Dichotomization – frequently employed to form two groups (e.g., mean-split).
     - Extreme groups – define three categories, eliminate middle group to accentuate differences.
   - Smoothing – use of response surface methods or other techniques to represent generalized patterns in the data.

Guidelines for Transforming Data

When explanation is important, beware of transformation

- To judge the potential impact of a transformation, calculate the ratio of the variable’s mean to its standard deviation:
  - Noticeable effects should occur when the ratio is less than 4.
  - When the transformation can be performed on either of two variables, select the variable with the smallest ratio .
- Generally applied to the independent variables except in the case of heteroscedasticity.
  - Heteroscedasticity can be remedied only by the transformation of the dependent variable in a dependence relationship. If a heteroscedastic relationship is also nonlinear, the dependent variable, and perhaps the independent variables, must be transformed.
- Transformations may change the interpretation of the variables.
  - For example, transforming variables by taking their logarithm translates the relationship into a measure of proportional change (elasticity). Always be sure to explore thoroughly the possible interpretations of the transformed variables.
  - Use variables in their original (untransformed) format when profiling or interpreting results.

Incorporating Nonmetric Data with Dummy Variables

- ❑ Concept of Dummy Variables
- ❑ Dummy Variable Coding
- ❑ Using Dummy Variables

Concept of Dummy Variable . . .

Definition: a nonmetric independent variable that has two distinct levels that are coded 0 and 1. These variables act as replacement variables to enable multi-category (3 or more) nonmetric variables to be used as metric variables.

Dummy Variable Coding of 3-category nonmetric variable with 2 dummy variables (X_1 and X_2):

Categories	X_1	X_2
Physician	1	0
Attorney	0	1
Professor	0	0