S1 Sains Data
FMIPA UNESA

Analisis Multivariat – Materi 01

Introduction to Multivariate Analysis

Prodi S1 Sains Data

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

Assessment Ratio

Komponen	Ratio
Partisipasi+Tugas	20%
Tes (UTS)	30%
Project	50%
Total	**100**

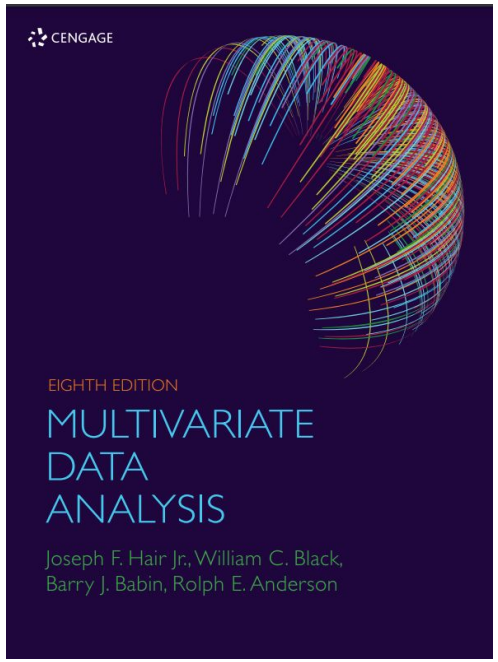
Class Rule

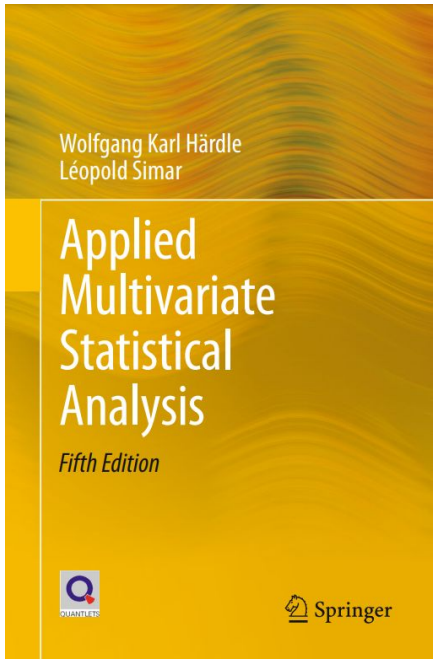
1. Attendance
   calculated at a minimum of **80%** of the total class
2. Permission
   Must **contact** the lecturer before the class
3. Plagiarism
   prohibited (grade E who violate this rule)
4. Make-up exams

   students can submit a follow-up request to the lecturer if students cannot attend a quiz, assignment, mid-exam, or final exam at a predetermined time

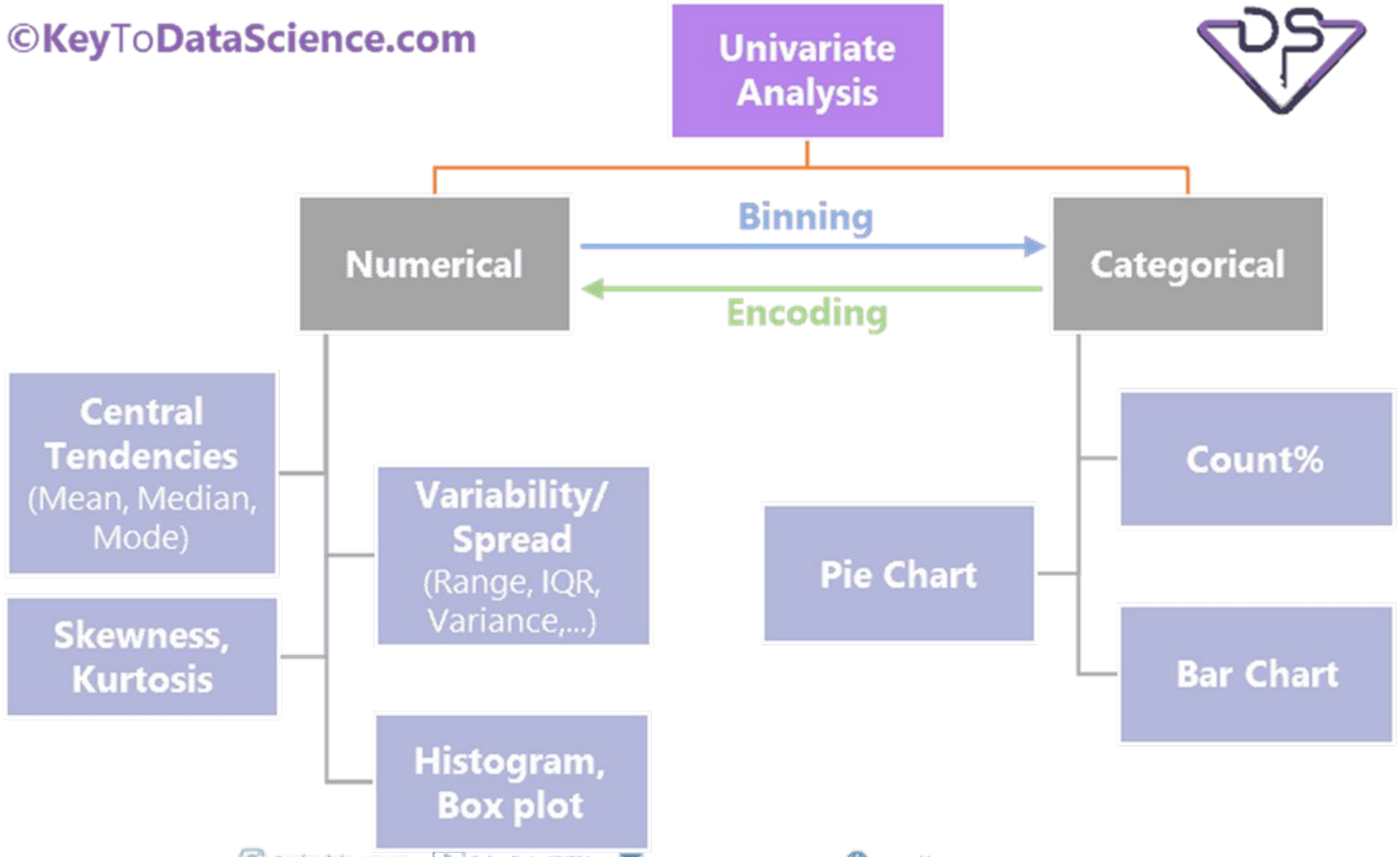
Reference

- Joseph F. Hair, Multivariate Data Analysis, 8th Ed. Cengage, 2018
- Härdle, Wolfgang Karl, and Léopold Simar. Applied multivariate statistical analysis. Springer Nature, 2019.





Univariate Analysis



Univariate Analysis

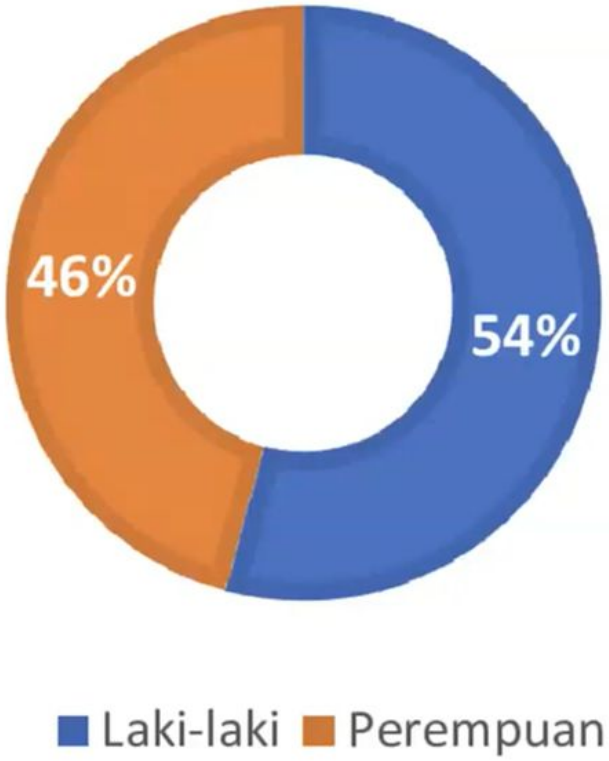
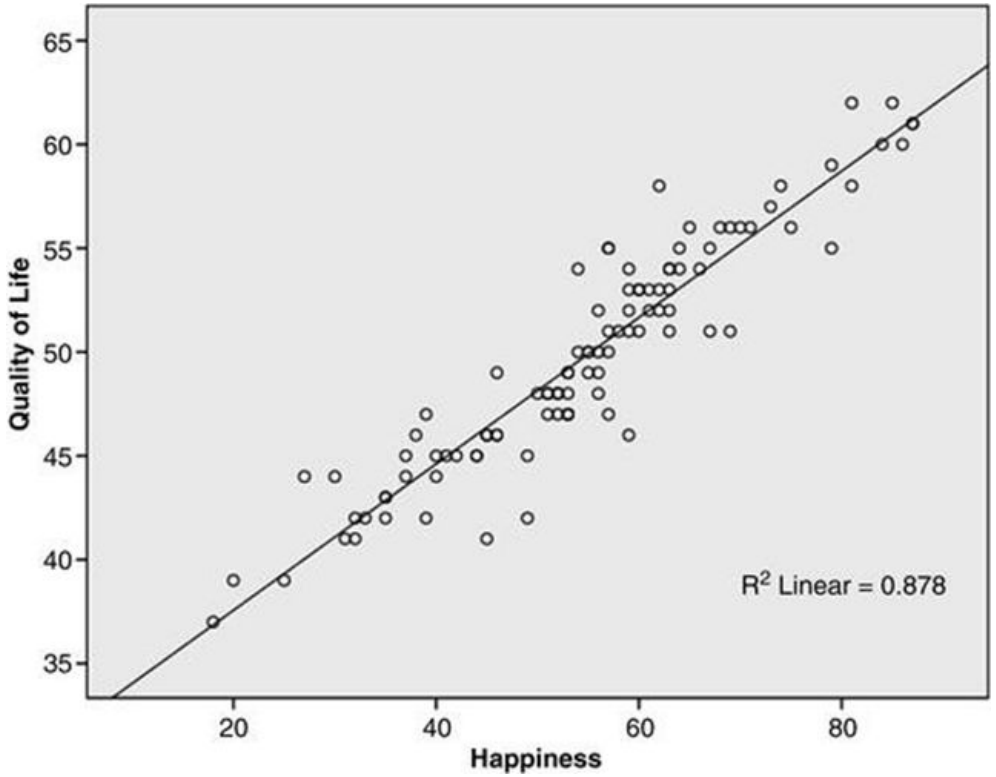
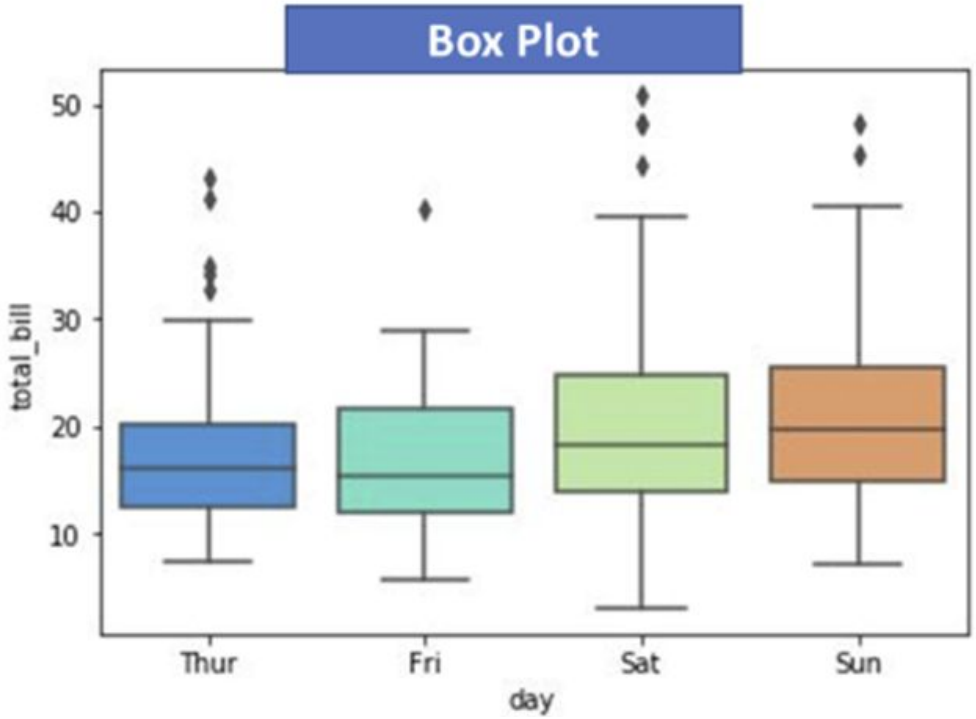


Table 1: Frequency distribution of research variables (n = 30133)

Variables	n (%)	Mean
Status of diabetes mellitus		
Suffering from diabetes mellitus	229 (0.8)	
Not suffering from diabetes mellitus	29904 (99.2)	
Cholesterol level status		
High cholesterol	236 (0.8)	
Normal cholesterol	29897 (99.2)	
Gender		
Man	14118 (46.9)	
Women	16015 (53.1)	
Hypertension status		
Hypertension	2371 (7.9)	
Not hypertension	27762 (92.1)	
Overweight		
Overweight/ Obesity	6536 (21.7)	21.03 Kg / m ^{2}
Non-overweight	23597 (78.3)	
Age		
≥ 40 years old	12275 (40.7)	26.05 years
<40 years	17858 (59.3)	

Bivariate Analysis





Bivariate Analysis

Favorite Flavor	Boys		Girls	
Vanilla	8	32%	9	26%
Chocolate	10	40%	6	17%
Strawberry	5	20%	14	40%
Mint Chip	2	8%	6	17%
Total	25	100%	35	100%

Table II: Correlation between cholesterol level status. gender. hypertension status. overweight and age to diabetes mellitus (n = 30133)

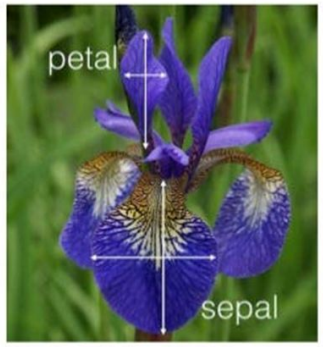
Variables	Status of diabetes mellitus		p value *	OR ** (95% CI)
	Sick n (%)	Painless n (%)		
Cholesterol status				
High cholesterol	52 (22.0)	184 (78.0)	0.000	47.453 (33.727 - 66.765)
Normal cholesterol	177 (0.6)	29720 (99.4)		
Gender				
Man	108 (0.8)	14010 (99.2)	0.925	-
Women	121 (0.8)	15894 (99.2)		
Hypertension status				
Hypertension	57 (2.4)	2314 (97.6)	0.000	3.951 (2.920 - 5.347)
Not hypertension	172 (0.6)	27590 (99.4)		
Overweight				
Overweight/obese	110 (1.7)	6426 (98.3)	0.000	3.377 (2.602 - 4.383)
Non-overweight	119 (0.5)	23478 (99.5)		
Age				
≥ 40 years old	97 (0.8)	12178 (99.2)	0.664	-
<40 years	132 (0.7)	17726 (99.3)		

*Significance $p \leq 0.05$; ** Crude OR; OR: Odds Ratio; CI: Confidence Interval

A dataset

Fisher's iris flower data set:
This famous dataset was collected to quantify the morphologic variation of iris flowers of three species - setosa, versicolor, virginica.



Sepal length	Sepal width	Petal length	Petal width	Species
5.0	2.0	3.5	1.0	*I. versicolor*
6.2	2.2	4.5	1.5	*I. versicolor*
6.0	2.2	5.0	1.5	*I. virginica*
6.0	2.2	4.0	1.0	*I. versicolor*
6.3	2.3	4.4	1.3	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
5.0	2.3	3.3	1.0	*I. versicolor*
4.5	2.3	1.3	0.3	*I. setosa*
5.5	2.4	3.8	1.1	*I. versicolor*
5.5	2.4	3.7	1.0	*I. versicolor*
4.9	2.4	3.3	1.0	*I. versicolor*
6.7	2.5	5.8	1.8	*I. virginica*
6.3	2.5	5.0	1.9	*I. virginica*

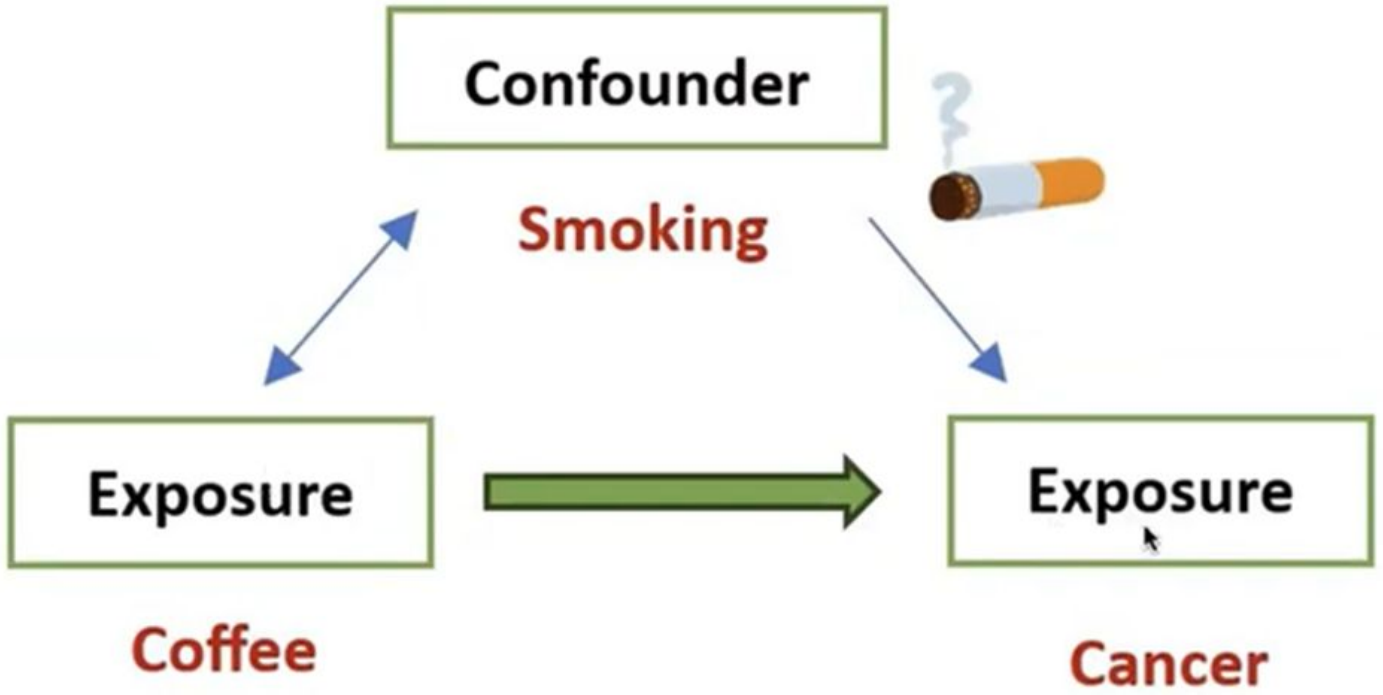
What is Multivariate Analysis?

how do you predict the weather ?

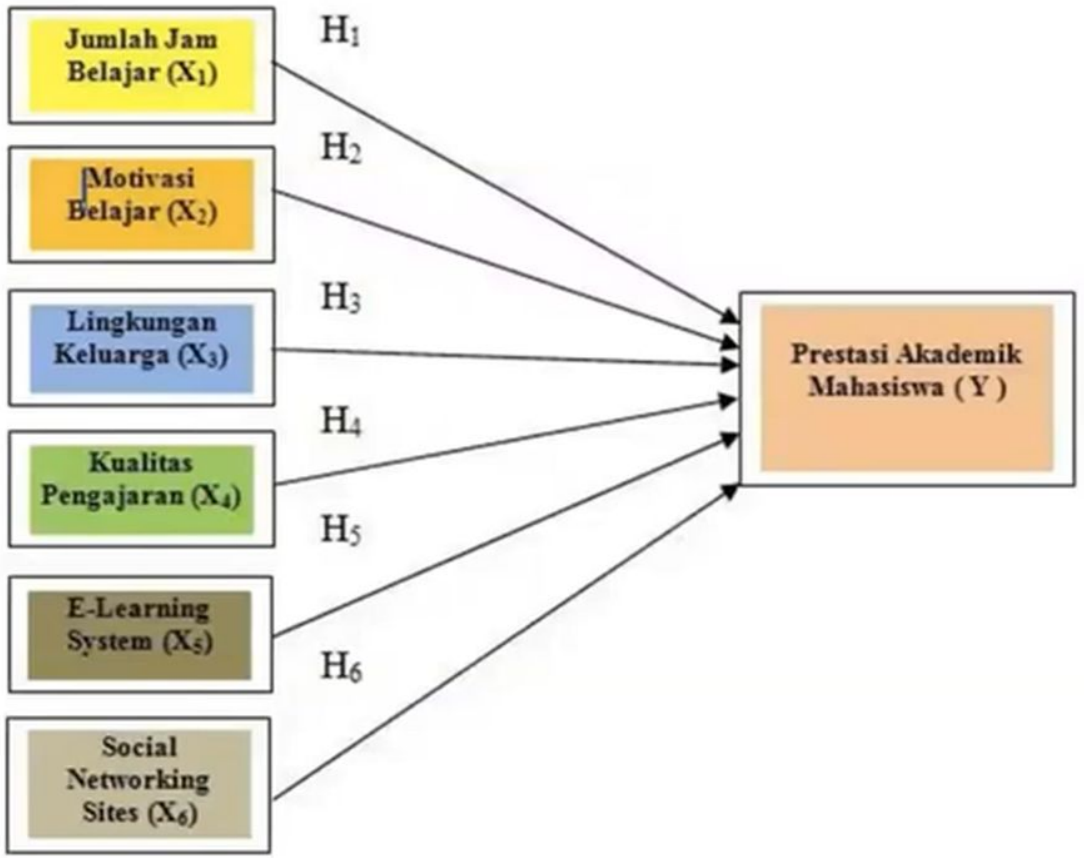
- humidity
- pollution
- precipitation

Multivariate means the involvement of multiple variables!

Multivariate Analysis



Multivariate Analysis

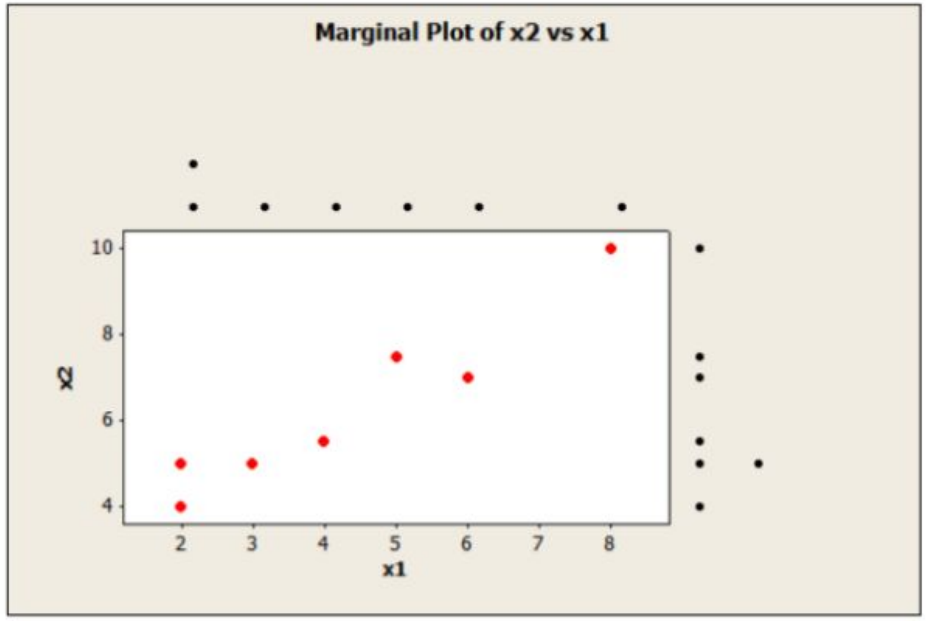


Multivariate Analysis

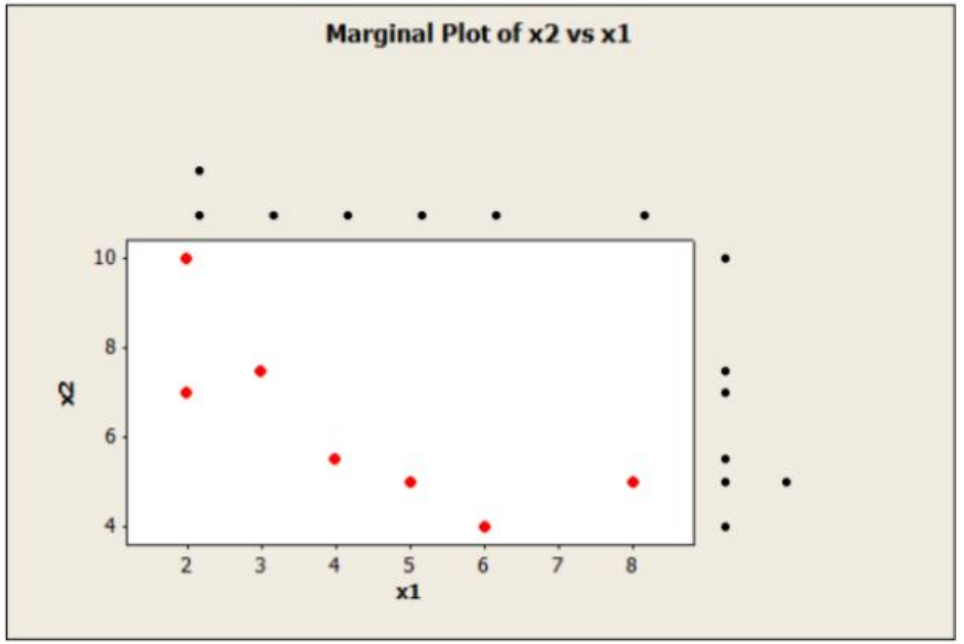
- **Multivariate analysis** involves evaluating multiple variables (more than two) to identify any possible association among them. Key takeaways: Multivariate analysis offers a more complete examination of data by looking at all possible independent variables and their relationships to one another.
- **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 loosely considered multivariate analysis

Why we learn Multivariate Analysis?

x1	3	4	2	6	8	2	5
x2	5	5,5	4	7	10	5	7,5



x1	5	4	6	2	2	8	3
x2	5	5,5	4	7	10	5	7,5



Basic Concept

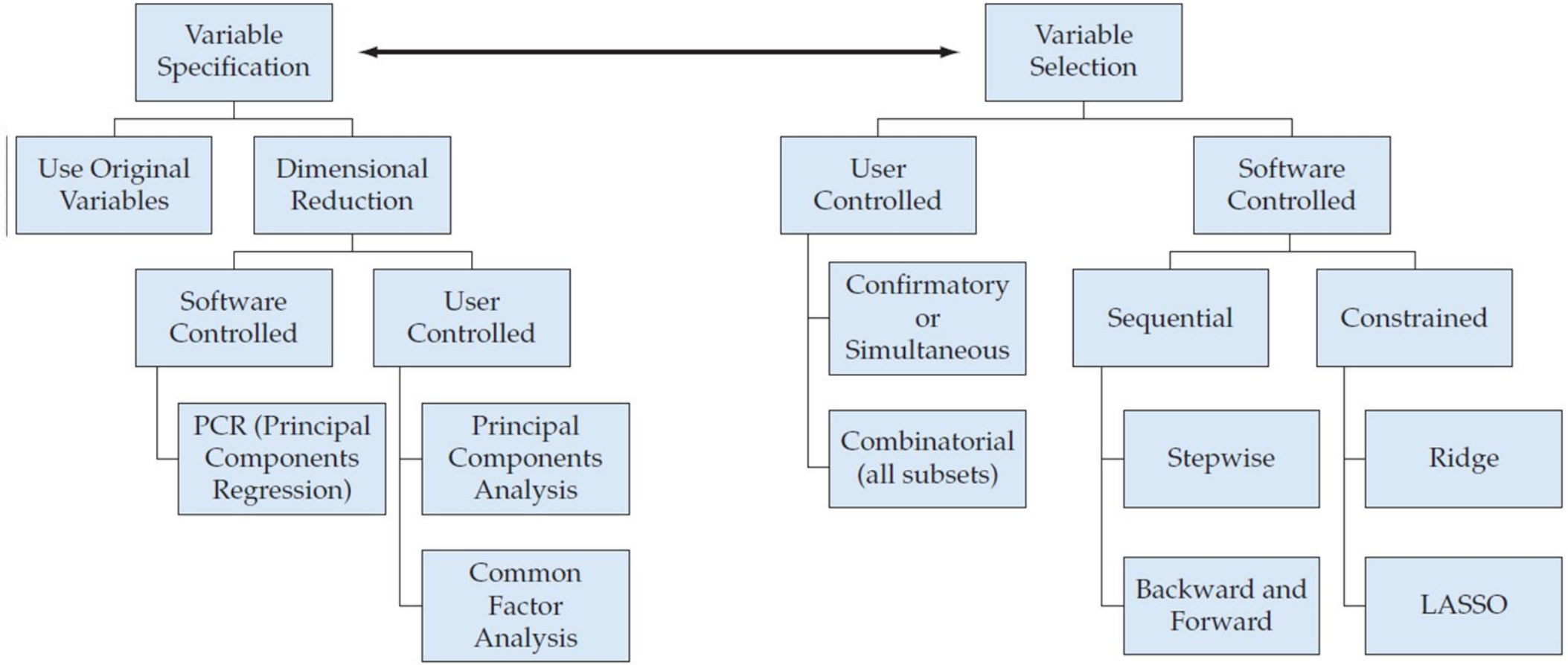
- The Variate
  - Dependent vs independent variables
- Measurement Scales
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Measurement error and Multivariate measurement
  - Validity and reliability

Objectives of Multivariate Analysis

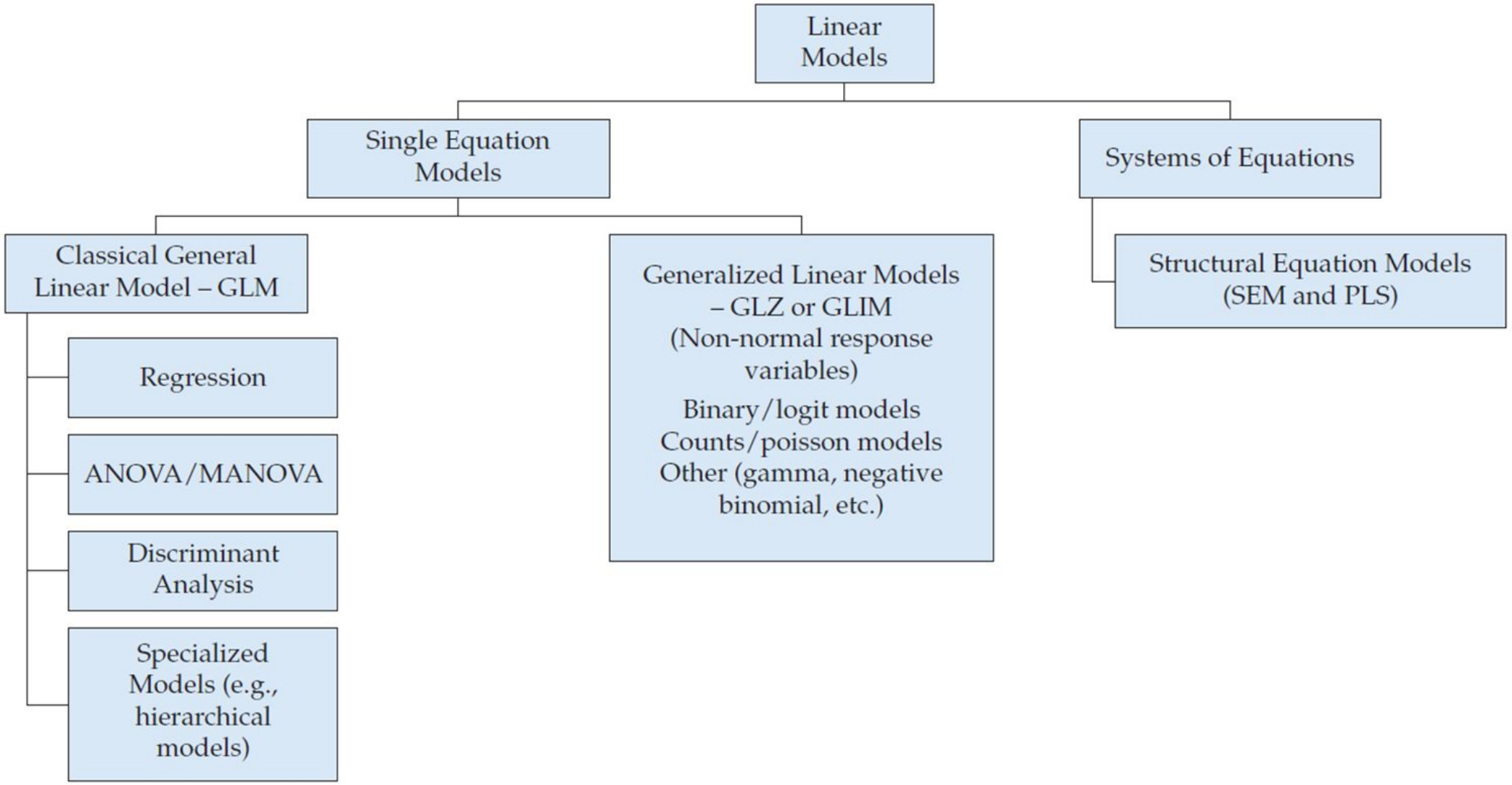
- Data reduction
- Grouping
- Relationship Among Variables
- Prediction
- Hypothesis Construction & Testing

Managing the Multivariate Model

Managing the Variate



Managing the Dependence Model



Statistical Significance vs Statistical Power

		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

Figure 1.4
Relationship of Error Probabilities in Statistical Inference

Figure 1.5
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

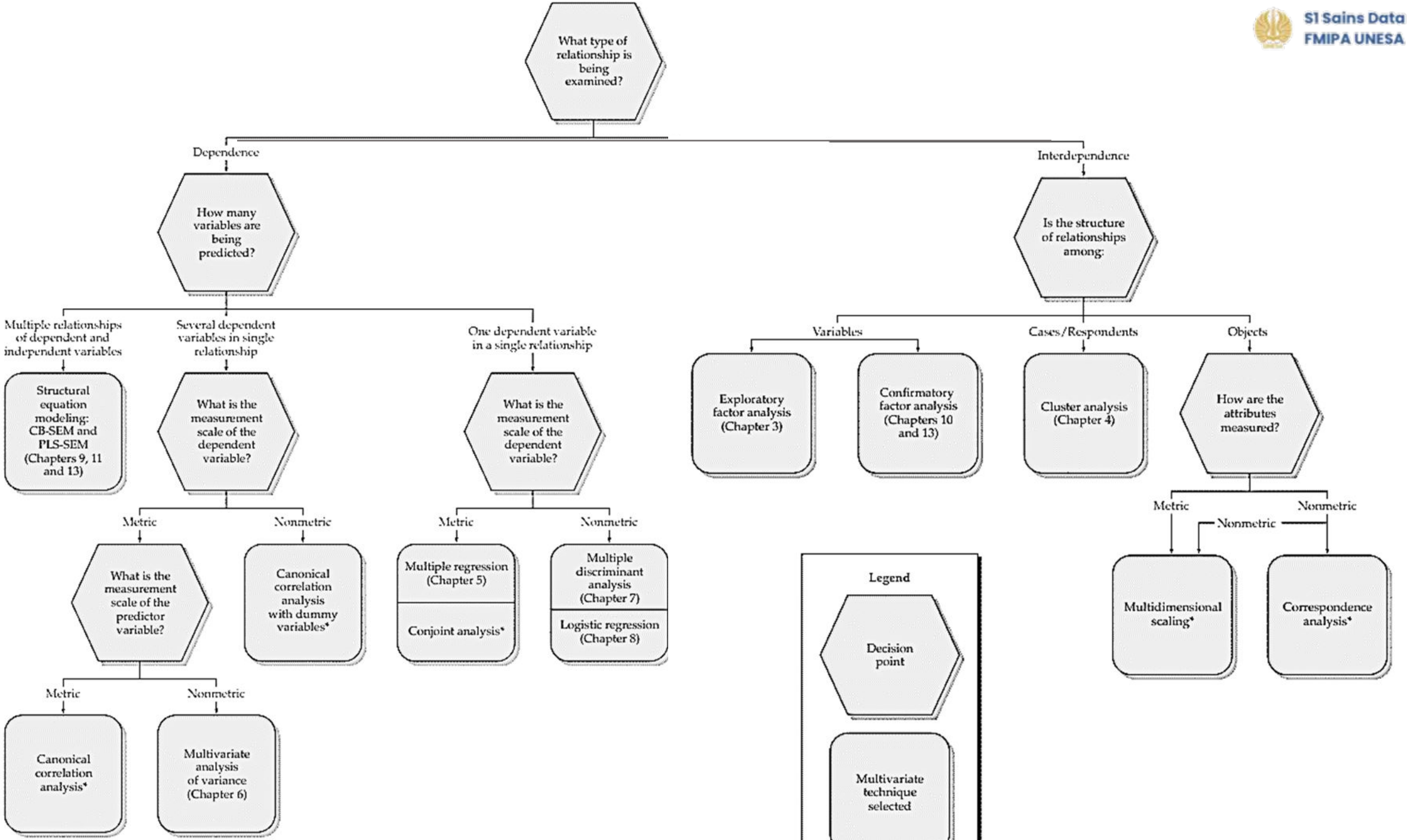
Source: *SOLO Power Analysis*, BMDP Statistical Software, Inc. [4]

A Classification of Multivariate Techniques

A Classification of Multivariate Techniques

- This 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 independent and dependent classifications 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?


What type of relationship is being examined?
Dependence
How many variables are being predicted?
Multiple relationships of dependent and independent variables
Structural equation modeling: CB-SEM and PLS-SEM (Chapters 9, 11 and 13)
Several dependent variables in single relationship
What is the measurement scale of the dependent variable?
Metric
What is the measurement scale of the predictor variable?
Metric
Canonical correlation analysis*
Nonmetric
Multivariate analysis of variance (Chapter 6)
Nonmetric
Canonical correlation analysis with dummy variables*
One dependent variable in a single relationship
What is the measurement scale of the dependent variable?
Metric
Multiple regression (Chapter 5)
Conjoint analysis*
Nonmetric
Multiple discriminant analysis (Chapter 7)
Logistic regression (Chapter 8)
Interdependence
Is the structure of relationships among:
Variables
Exploratory factor analysis (Chapter 3)
Confirmatory factor analysis (Chapters 10 and 13)
Cases/Respondents
Cluster analysis (Chapter 4)
Objects
How are the attributes measured?
Metric
Multidimensional scaling*
Nonmetric
Correspondence analysis*
Legend
Decision point
Multivariate technique selected

Canonical Correlation

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

Multivariate Analysis of Variance

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

Analysis of Variance

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

Multiple Discriminant Analysis

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

Multiple Regression alysis

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

Conjoint Analysis

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

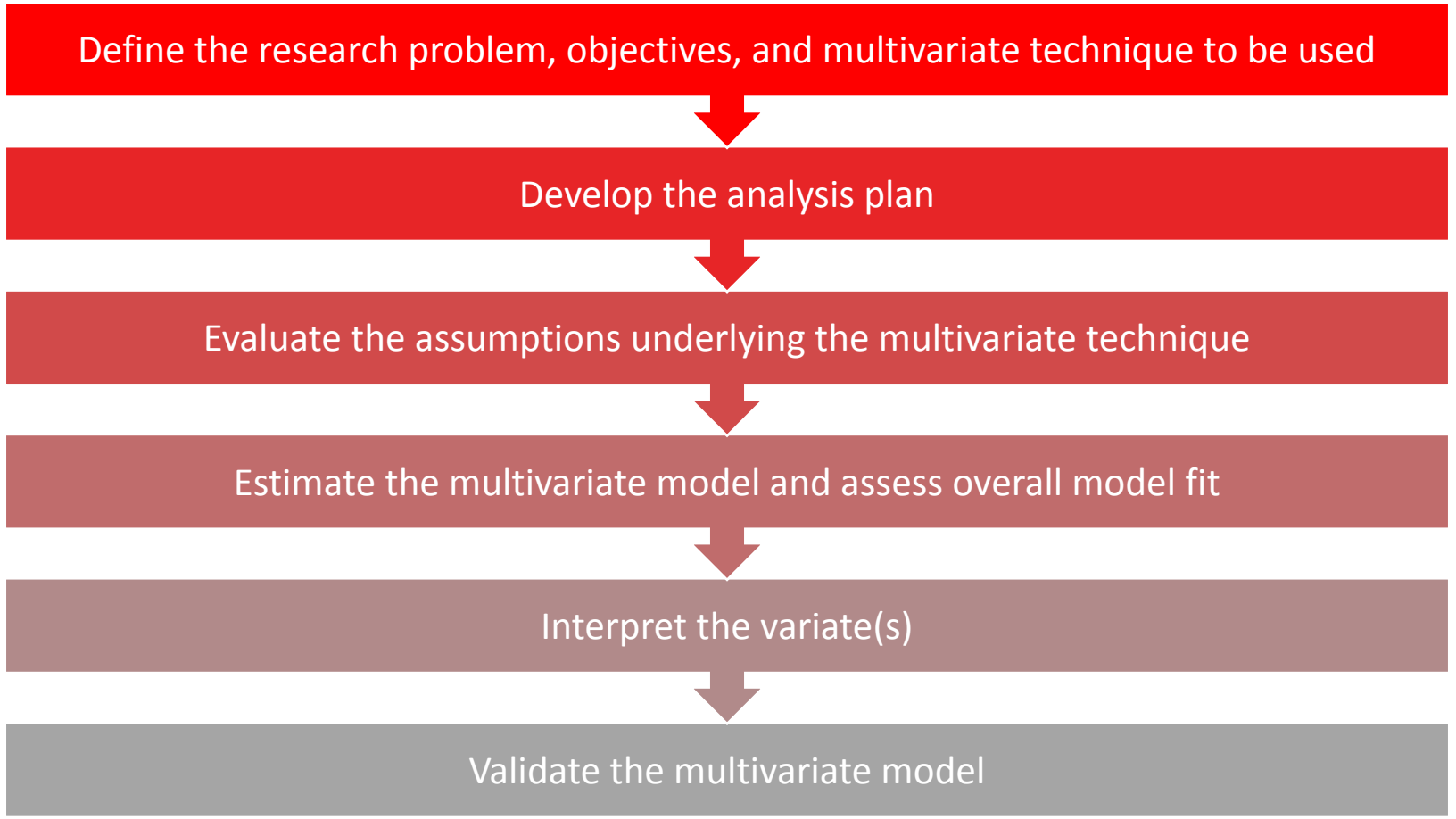
St uctural Equation Modeling

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

Guidelines for Multivariate Analyses and Interpretation

- Establish practical significance as well as statistical significance
- Recognize that sample size affects all results
- Know your data
- Strive for model parsimony
- Look at your errors
- Simplify your models by separation
- Validate your results

A Structured Approach to Multivariate Model Building



Multivariate Analysis Application

- Marketing
- Banking
- Finance
- Insurance
- Healthcare
- Molecular biology
- Astronomy
- Sports

Multivariate Analysis Application

- **Marketing:** Predict new purchasing trends. Identify "loyal" customers. Detect potential customers. Segment markets. Precise marketing.
- **Banking:** Evaluate loan policies using customer characteristics. Predict credit card switch.
- **Finance:** Identify relationships between financial indicators. Track changes in an investment portfolio and predict price turning points. Analyze volatility patterns in high-frequency stock transactions.
- **Insurance:** Identify characteristics of buyers of new policies. Find unusual claim patterns. Identify "risky" customers.

Multivariate Analysis Application

- **Healthcare:** Early warning of diseases. Predict doctor visits from patient characteristics. Precise medical care.
- **Molecular Biology:** Gene detection. Analyze DNA microarrays. Characterize biological function. Predict protein structure.
- **Astronomy:** Catalogue (as stars, galaxies, etc.) objects in the sky. Identify patterns and relationships of objects.
- **Forensic Accounting:** Detect fraud in insurance, credit card and medical claims. Identify instances of tax evasion. Identify insider-trading behaviors in stock market.
- **Sports:** Identify most effective stretagies. Discover hidden game patterns.

Next

- Multivariate normal distribution
- Canonical correlation
- Factor analysis
- ANOVA, ANCOVA, MANOVA, MANCOVA
- Clustering
- Conjoint analysis
- Multidimensional Scaling
- Discriminant analysis
- Logistic regression
- SEM dan PLS

Thank you