Applied Multivariate Data Analysis
(應用多變量分析)

By Yen-I Chiang (江彥逸),

Department of Information Management

CHAPTER

Introduction

Course Overview (1)

- ✓ 課程負責教師:　江彥逸 Yen-I Chiang
- ✓ Email: yenichn@mail.cgu.edu.tw
- ✓ Background: Algorithms, Data Analysis
- ✓ **上課教室(資管系):**電腦4 (Computer lab 4)
- ✓ 上課時間: 09:10 to 12:00
- ✓ Consultation Time:

  星期三下午二點至四點

  星期五下午二點至四點

Motivations

- More complex data sets.
- Availability of high performance computers.
- More difficult problems that need to be modeled and solved.
- Traditional Statistic method unable to solve, feasibility in applying some complex algorithms

Data Mining History

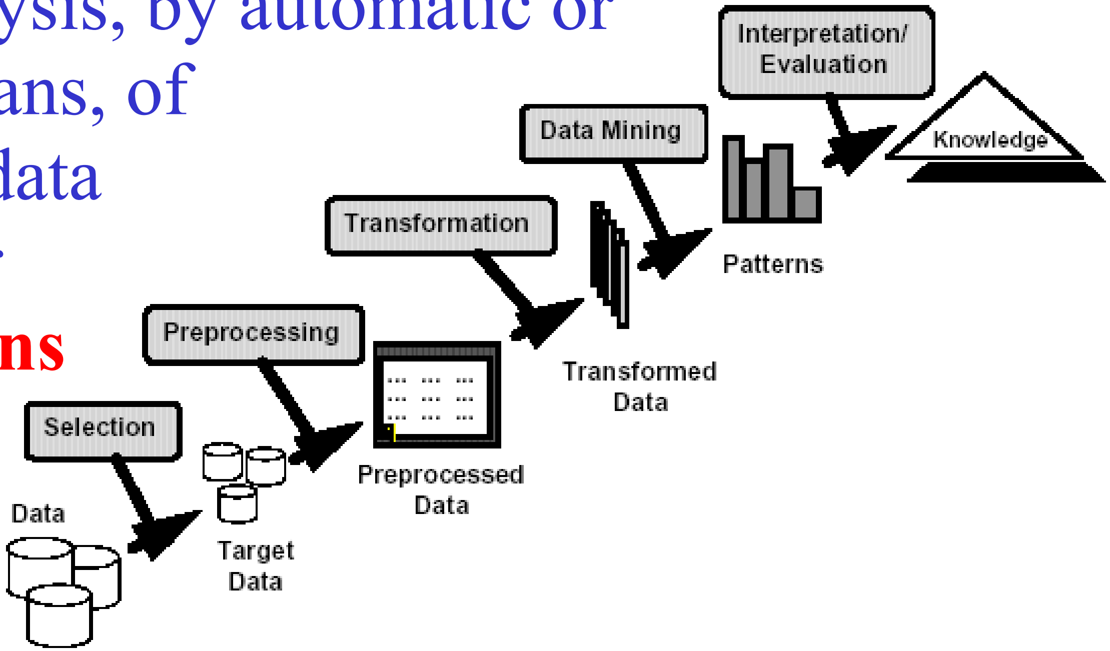
- The approach has roots in practice dating back over 50 years.
- In the early 1960s, data mining was called statistical analysis, and the pioneers were statistical software companies such as SAS and SPSS.
- By the late 1980s, the traditional techniques had been augmented by new methods such as fuzzy logic, heuristics and neural networks.
- Currently, methods such as deep learning etc..

Why Doing Multivariate Analysis?

- Statistical theory was developed several decades ago.
- Widespread use more recent
  - Increased computing power
  - Increased availability of software (SAS, SPSS, Minitab)
- Easy to analyze large quantities of complex data.

What is Data Mining?

- **Many Definitions**
  - Non-trivial **extraction of** implicit, previously unknown and potentially **useful information from data**
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover **meaningful patterns**



Core competence:

- Interdisciplinary Learning: Learning and Integration Capabilities for Interdisciplinary and New Knowledge. (0.3)
- Data Analysis Capabilities: Data Collection, Analysis, Organization, Interpretation, and Self-Learning. (0.7)

核心能力	權重	國際化能力	領導能力	執行能力	創新能力	決策能力
應用多變量分析	**1**	0	0	0	**0.3**	**0.7**

AACSB

- **EMBA-LG2-LO2: Alternative Development**

Program Learning Goal: Our students will be able to solve problems effectively.

Program Learning Objective: Our students will be able to develop alternatives from strategic perspectives.

- **EMBA-LG2-LO3: Alternative Assessment**

Program Learning Goal: Our students will be able to solve problems effectively.

Program Learning Objective: Our students will be able to assess alternatives and make the decision.

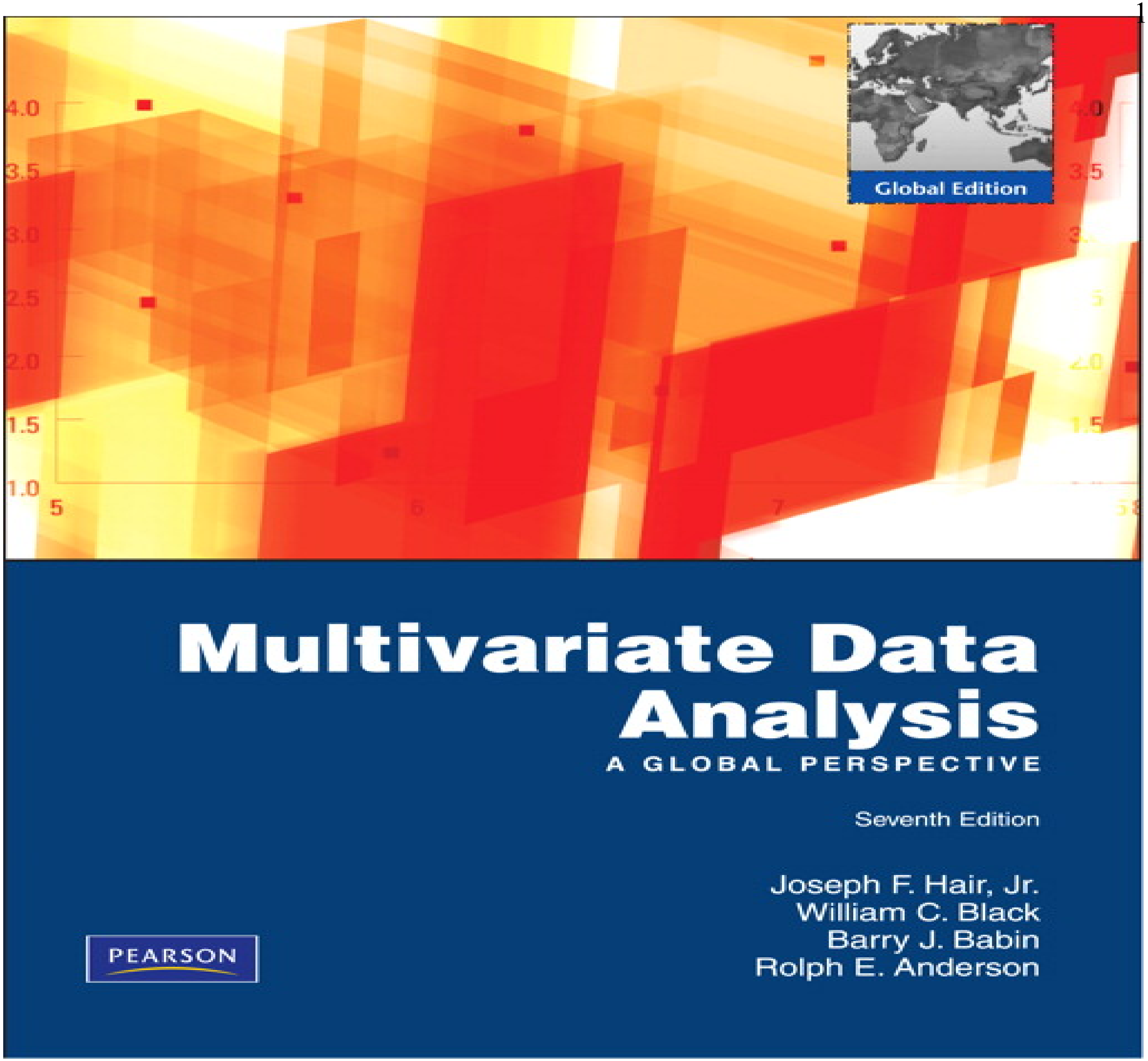
Course Overview 教科書:

1. Multivariate Data Analysis - A global perspective, 7/e, Prentice Hall International, J. F. Hair, W. C. Black, B. J. Babin and R. E. Anderson. (華泰文化)
2. Analyzing Multivariate Data, by J. Lattin, J. D. Carroll and P. E. Green. (滄海書局)

- Course Materials: I will update the course materials weekly and it can be downloaded at following ftp site(課程講義網址):

數位學習 E-Learning: **https://el.cgu.edu.tw/mooc/**

Multivariate Data Analysis, 7/e, Prentice Hall International, Hair, Black, Babin and Anderson.



Global Edition

Multivariate Data Analysis

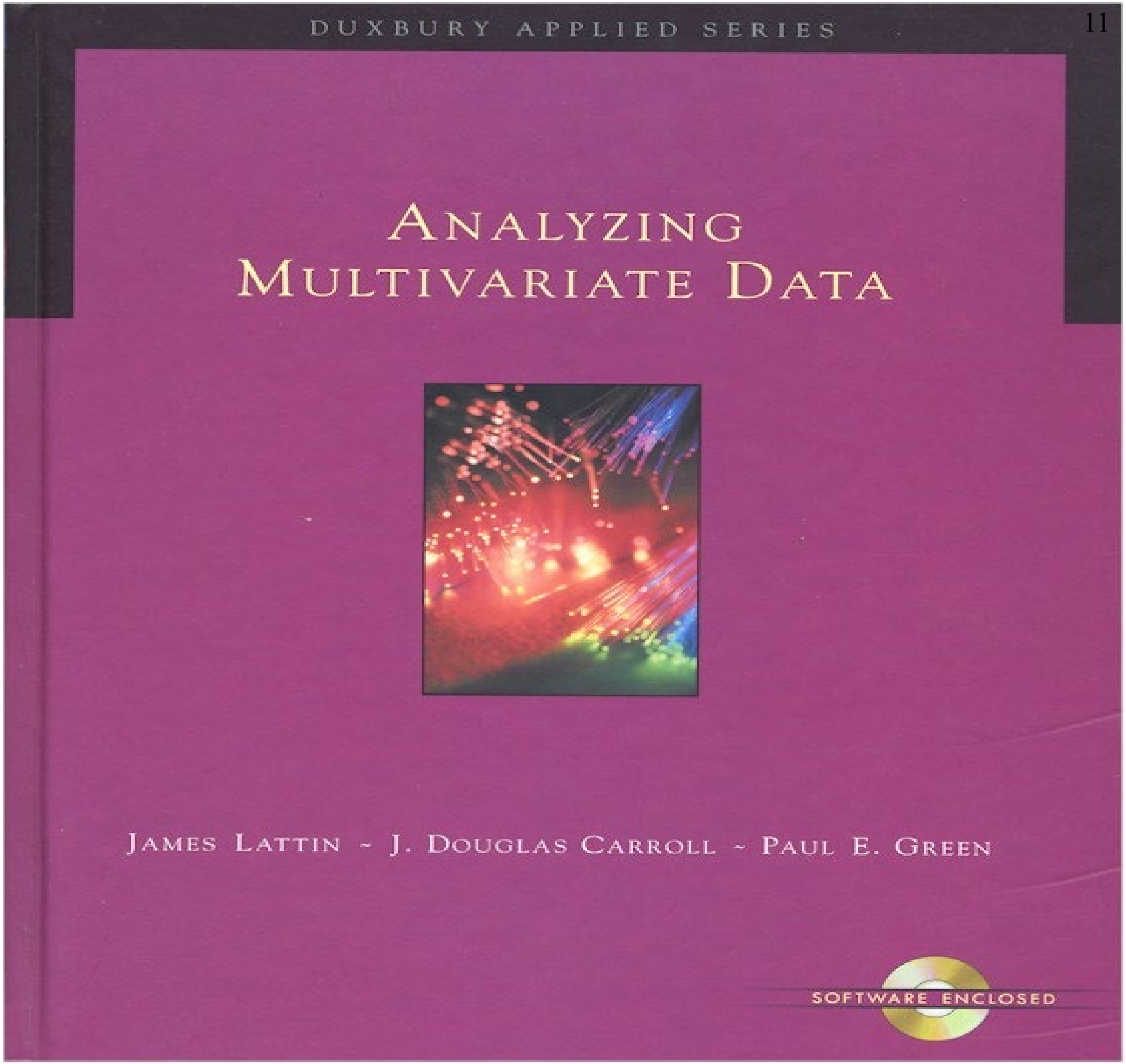
A GLOBAL PERSPECTIVE

Seventh Edition

Joseph F. Hair, Jr.
William C. Black
Barry J. Babin
Rolph E. Anderson

PEARSON

Analyzing Multivariate Data, by J Lattin, J D Carroll and P E Green.



DUXBURY APPLIED SERIES

ANALYZING MULTIVARIATE DATA

JAMES LATTIN ~ J. DOUGLAS CARROLL ~ PAUL E. GREEN

SOFTWARE ENCLOSED

.多變量統計分析(最佳入門實用書)第二版 SPSS＋LISREL;蕭文龍編

多變量分析

最佳入門實用書 第二版

SPSS+LISREL

Course Overview (3)

參考書:

1. 多變量統計分析(最佳入門實用書)第二版 SPSS＋LISREL;蕭文龍編，基峯資訊出版，
2. SPSS 統計應用實務; 吳明隆編，文魁資訊出版, 2005 ，松崗.
3. SPSS 多變量統計分析;張紹勳，林秀娟，2003，滄海書局.
4. Applied Multivariate Techniques, Wiley, Subhash Sharma. (雙葉書局)

Course Overview (4)

Provide students with an understanding the theoretical and fundamentals of multivariate techniques. Then to learn to use software and computers to solve problems and analysis the output (in order listed below, based on Hair’s textbook):

Course Overview (5)

Chapter 1. Introduction: methods and model building

Chapter 2. Preparing for Multivariate analysis

Chapter 4. Simple and Multiple Regression

Chapter 3. Factor Analysis

Chapter 9. Cluster analysis

Chapter 8. MANOVA and MACOVA.

Chapter 5. Canonical Correlation

Chapter 7. Multiple Discriminant Analysis and Logistic Regression

Chapter 10. Multidimensional Scaling

Potential Topics Covered

- Learn how to use SPSS.
- Review Associative Topics
- Principal Components Analysis (PCA)主成分分析
- Factor Analysis (FA) 因素分析
- Discriminate Analysis (DA)兩群的判別分析(Logit)
- Multidimensional Scaling (MDS)多元尺度化
- Cluster Analysis (CA) 群集分析
- Canonical Correlations Analysis (CCA)典型相關
- Multivariate Regression 多元迴歸
- Multivariate Analysis of Variance (MANOVA)

  共變數分析,多變量變異數分析

No.	日期Date	教學進度Outline
1	2021-09-24	Introduction and Overview of Applied Multivariate Analysis概述及課程介紹
2	2021-10-01	Introduction: SPSS and Basic Statistical methods SPSS和基本統計方法介紹
3	2021-10-08	Simple Regression 迴歸
4	2021-10-15	Multiple Regression 多元迴歸 Part 1
5	2021-10-22	Multiple Regression 多元迴歸 Part 2
6	2021-10-29	Factor Analysis: PCA 主成分分析
7	2021-11-05	Factor Analysis: PAF 因素分析
8	2021-11-12	General Linear Model (ANOVA and ANCOVA) 變異數分析 Part 1
9	2021-11-19	General Linear Model (k-ANOVA) 變異數分析 Part 2
10	2021-11-26	Midterm Exam. 期中考
11	2021-12-03	Cluster analysis (1) hierarchical 群集分析 Part 1 (階層式)
12	2021-12-10	Cluster analysis (2) non-hierarchical 群集分析 Part 2 (非階層式)
13	2021-12-17	Logistic Regression (1) 兩群的判別分析 Part 1
14	2021-12-24	Logistic Regression (2) 兩群的判別分析 Part 2
15	2021-12-31	Holiday 元旦放假
16	2022-01-07	MANOVAR 變異數的多變數分析
17	2022-01-14	Multidimensional scaling (1) 多元尺度化 Part 1
18	2022-01-21	Final Exam 期末考 ???

Extra data materials

Materials for practice can be found

- from the textbook (Analyzing Multivariate Data, by J Lattin, J D Carroll and P E Green);
- from **The Data and Story Library**
  - http://www.statsci.org/data/index.html
  - http://lib.stat.cmu.edu/DASL/DataArchive.html

修課限制與需求:
Knowledge in basic Statistic.

Grading

1. Class attendances: 10% 出席表現 10%
2. Class Assignments and Participation: 20% 課堂作業、表現 20% (at the end of each chapter, class assignments with questions similar to the class examples will be given to the students for to do them at home or in class individually； must participate in class discussions, homework discussions etc.)
3. Practical presentation: 20% 課堂報告 20% (students will select one of the question from the Class Assignments and then students will present their work in class and be graded accordingly)
4. Midterm Exam. 25% 期中考 25%
5. Final Exam: 25% 期末考 25%

A quick introduction

What is Multivariate Analysis?

- What is it?

Multivariate Data Analysis = all statistical methods that **simultaneously analyze multiple measurements** on each individual or object under investigation.

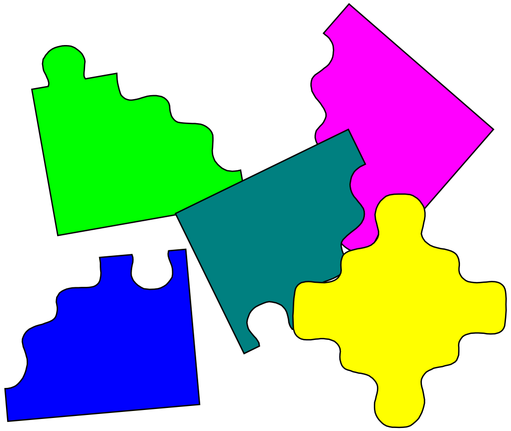
- Why use it?
    - Explanation & Prediction
    - Hypothesis Testing

What is Multivariate Analysis?

- Some techniques are extensions to the familiar Statistical analyses
  - analysis of single-variable distributions
  - simple regression
  - analysis of variance (ANOVA)
  - correlation

Nature of Multivariate Analysis

- Typically exploratory, not confirmatory.
- Often focused on simplification.
- Often focused on revealing structure in dimensions that our eyes and imaginations don′t fully support.



Our Approach

- Emphasize
  - Intuition
  - Geometry
  - Interpretations
  - Data Analysis

- De-emphasize
  - Theoretical basis
  - Formal proofs

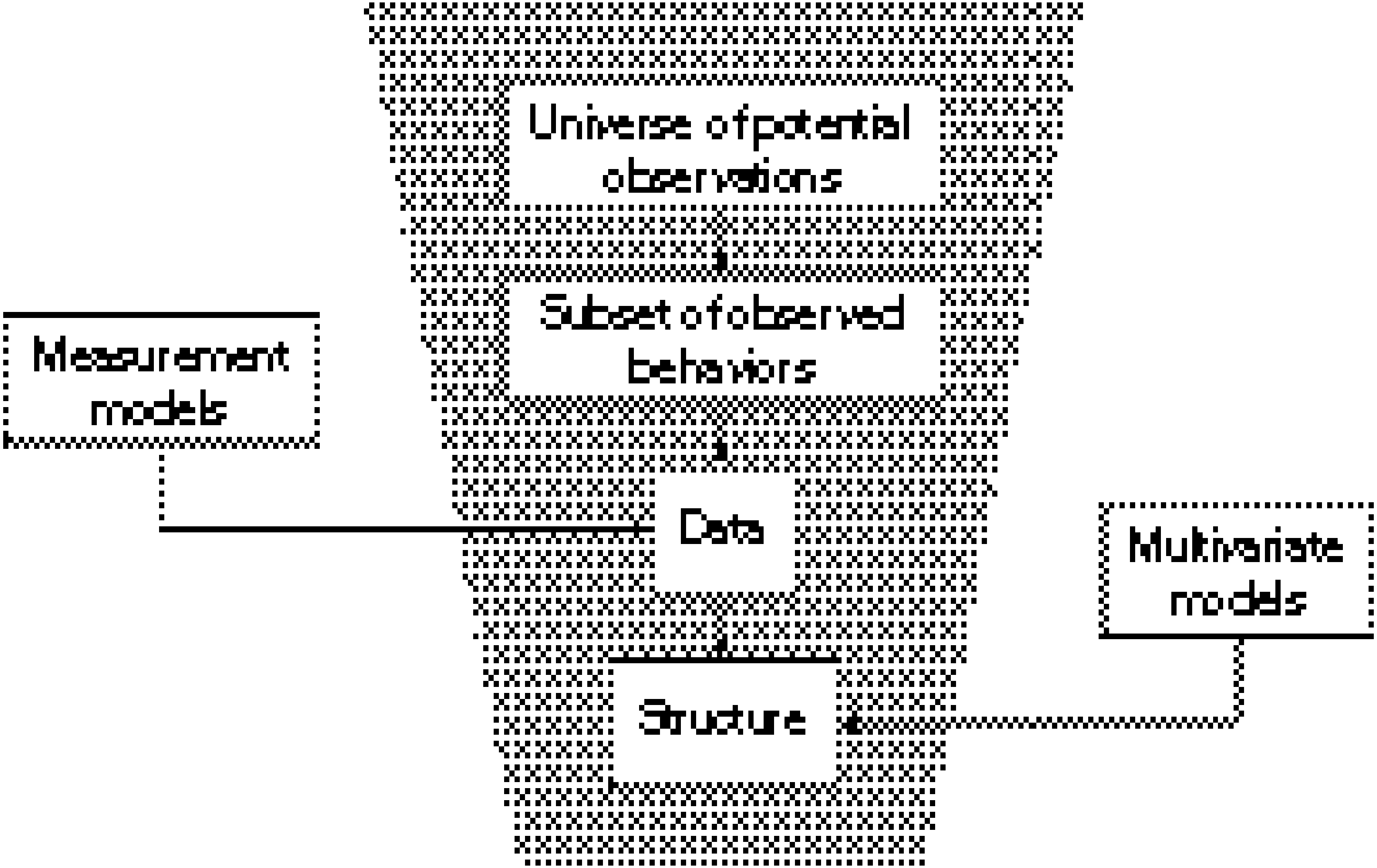
Why Multivariate?

- Typically more than one measurement is taken on a given experimental unit

- Need to consider all the measurements together so that one can understand how they are related

- Need to consider all the measurements together so that one can extract essential structure

The process of searching for structure


Universe of potential observations
Subset of observed behaviors
Measurement models
Data
Multivariate models
Structure

Should you become a professional photographer?

We can use factor analysis to find the attributes that contribute to become a professional photographer.

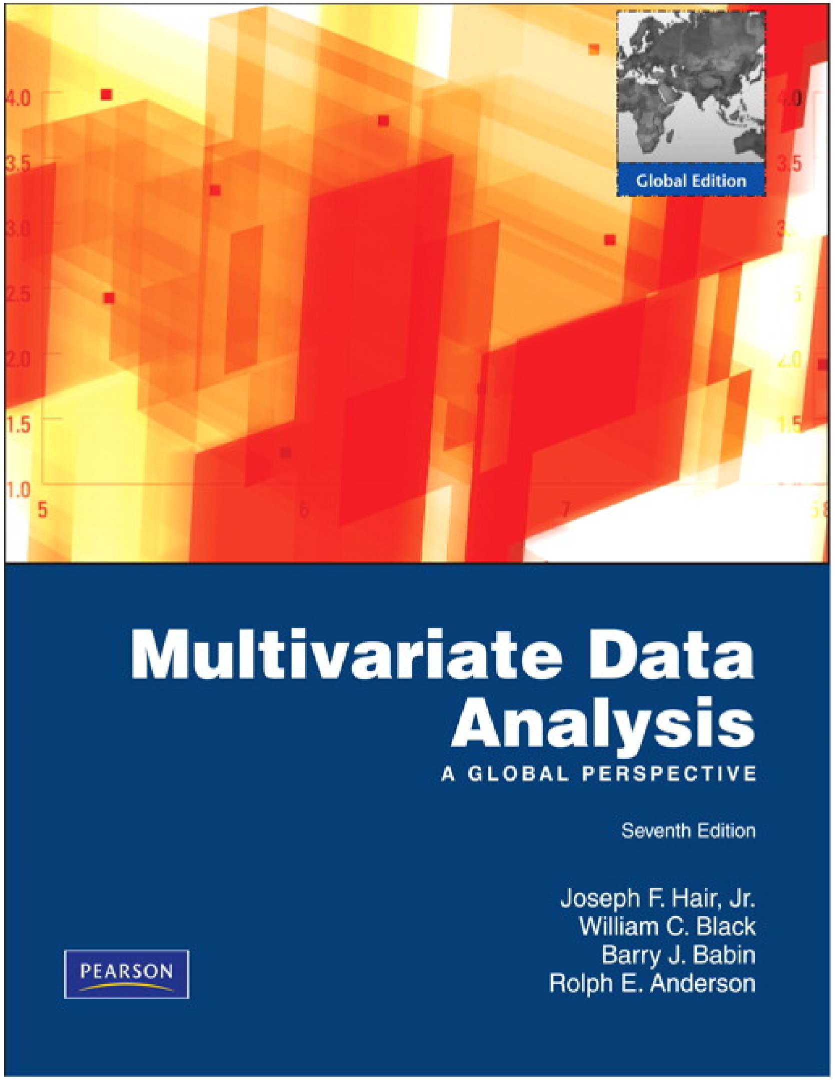
Can your car sale well in Africa?

We can use Multidimensional Scaling to do a research in marketing or product development.

What will you learned in the end of this course?

- How to do Data Mining using Statistics
- Expand your repertoire of analytical options.
- Understand enough theory to appreciate appropriate and inappropriate application.
- Be able to use and interpret data.
- Know where to go for additional help.

Chapter 1 Introduction: Methods and Model Building



Multivariate Data Analysis

A GLOBAL PERSPECTIVE

Seventh Edition

Joseph F. Hair, Jr.
William C. Black
Barry J. Babin
Rolph E. Anderson

By Yen-I Chiang (江彥逸),
Department of Information Management

Chapter 1 Introduction: Methods and Model Building

LEARNING OBJECTIVES:

Upon completing this chapter, you should be able to do the following:

1. Explain what multivariate analysis is and when its application is appropriate.
2. Define and discuss the specific techniques included in multivariate analysis.
3. Determine which multivariate technique is appropriate for a specific research problem.
4. Discuss the nature of measurement scales and their relationship to multivariate techniques.
5. Describe the conceptual and statistical issues inherent in multivariate analyses.

What is Multivariate Analysis?

- What is it? Multivariate Data Analysis = all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation.

- Why use it?
  - Measurement
  - Explanation & Prediction
  - Hypothesis Testing

What is Multivariate Analysis?

- Some techniques are extensions to the familiar uni-variate and bi-variate analyses
  - analysis of single-variable distributions
  - simple regression
  - analysis of variance
  - correlation
- e.g. extend simple regression to look at several predictor variables or extend ANOVA to include several dependent variables

Basic Concepts of Multivariate Analysis

- The Variate
- Measurement Scales
    - Nonmetric
    - Metric
- Multivariate Measurement
- Measurement Error
- Types of Techniques

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.
- Variate equation: $(Y') = {}_{w1} X_1 + {}_{w2} X_2 + \ldots + {}_{wn} X_n$
- 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.
- Potential Independent Variables:

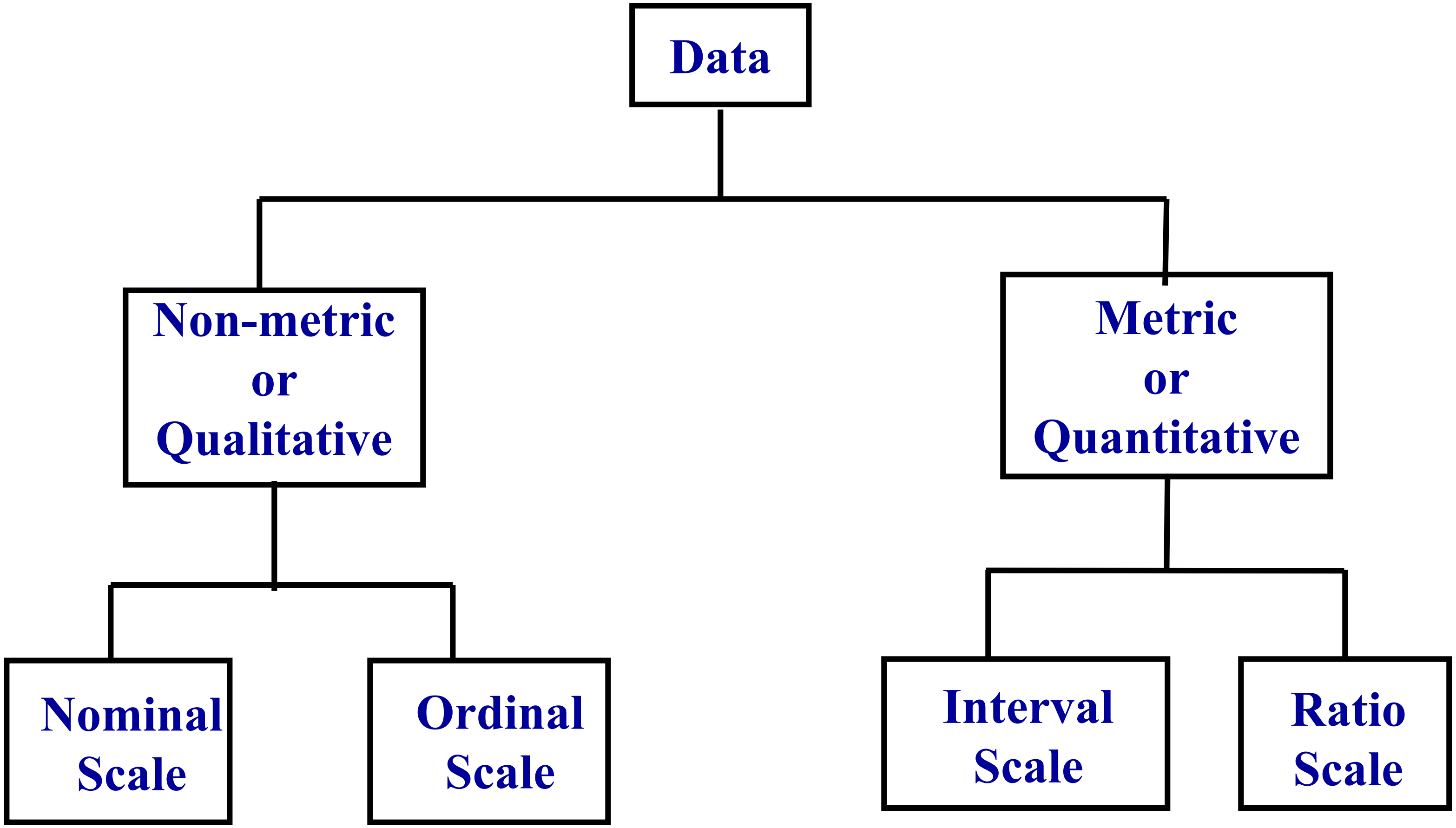
  X1 = income

  X2 = education

  X3 = family size

  X4 = ??

Types of Data and Measurement Scales



Data

Non-metric 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!

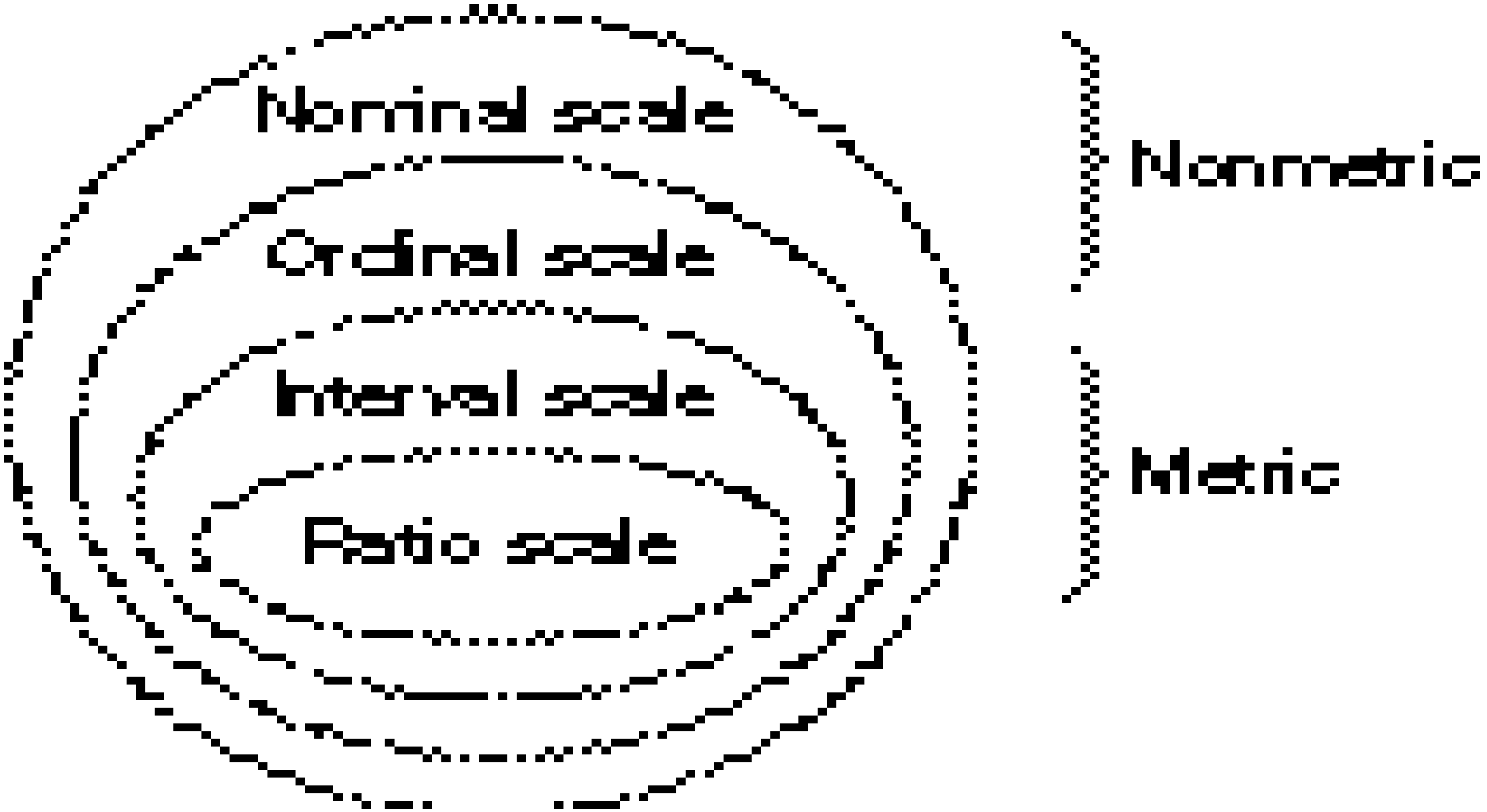
Measurement Scales

- Nonmetric (qualitative)
  - Nominal   eg Male or Female
  - Ordinal   eg level of satisfaction with a course

- Metric (quantitative)
  - Interval   eg Temperature measured in C
  - Ratio   eg Weight, height

變數資料之類型

- 名目尺度(nominal scale)
- 序列尺度(ordinal scale)
- 區間尺度(interval scale)
- 比率尺度(ratio scale)



Measurement Errors

- Difference between observed and true values
  - Observed values include noise
  - Validity - Does the measure represent what it's supposed to ?
  - Reliability - How much error is present?

- Summated Scales - using a composite measure to represent a concept

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.

Measurement Error

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

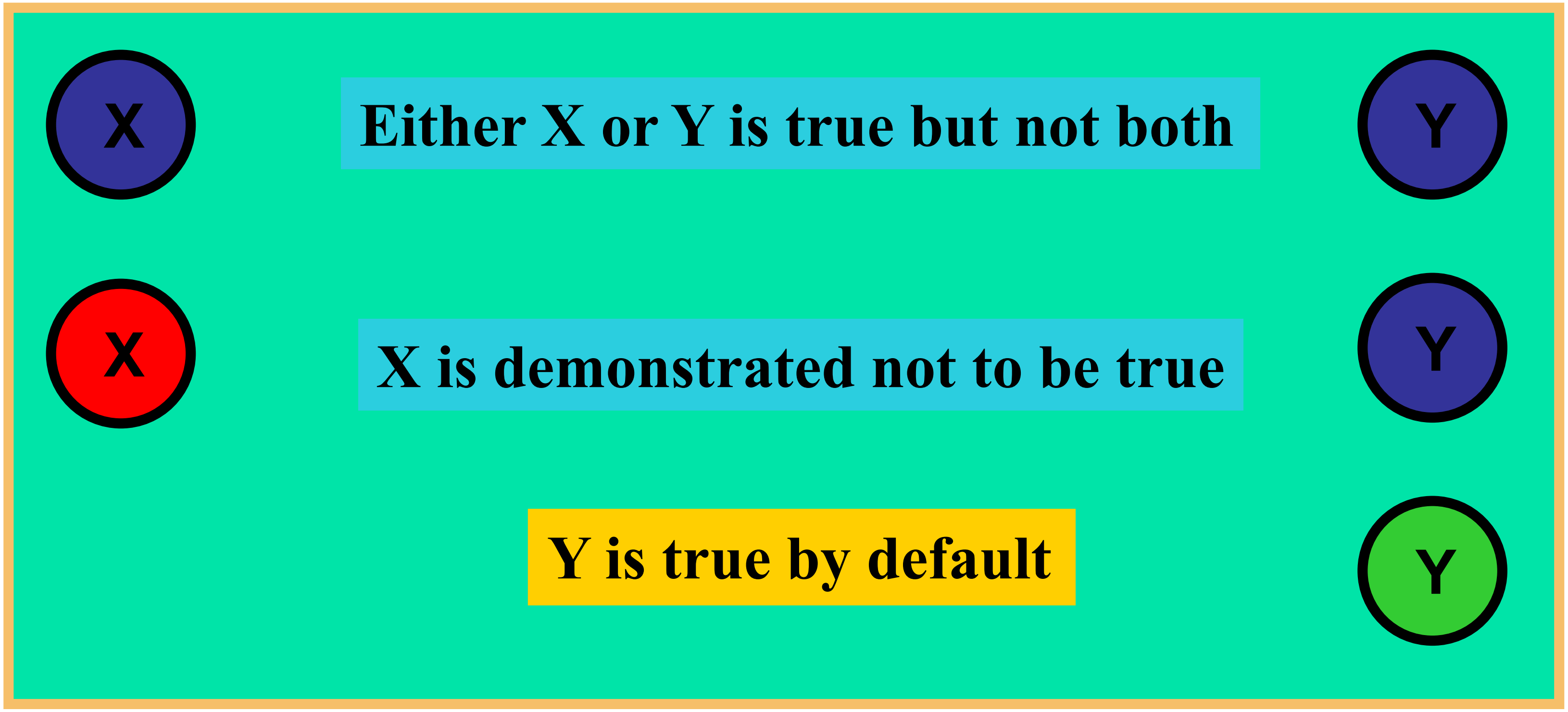
- Validity (效度) =

  the degree to which a measure accurately represents what it is supposed to. (指測量的工具能夠精確地反映出要測量的概念，就是「我們想要測量的是什麼？」)

- Reliability (信度) =

  the degree to which the observed variable measures the “true” value and is thus error free. (即可靠性，指測驗結果的一致性或穩定性。一個測驗的信度在於表示測驗內部問題間是否相互符合與同一現象進行重複觀察之後是否前後一致得到相同資料值)

Method of Indirect Proof



Either X or Y is true but not both

X is demonstrated not to be true

Y is true by default

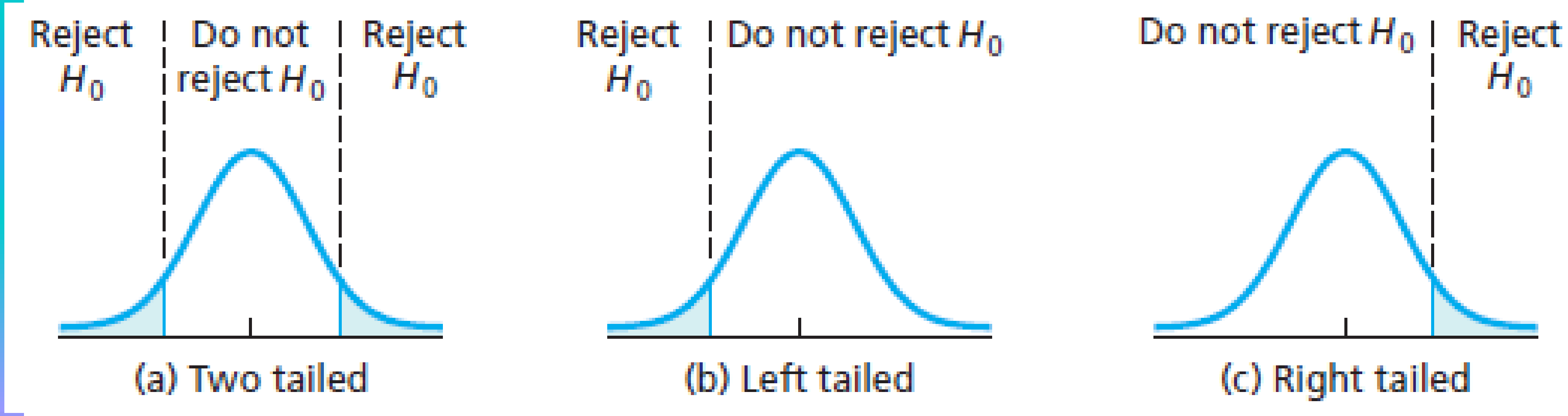
Definition

Significance Level

The probability of making a Type I error, that is, of rejecting a true null hypothesis, is called the **significance level,** α **,** of a hypothesis test.

Idea of Hypothesis Test

Graphical display of rejection regions for two-tailed, left-tailed, and right-tailed tests



(a) Two tailed

(b) Left tailed

(c) Right tailed

Statistical Significance and Power

- **Type I error**, or α, is the probability of rejecting the null hypothesis when it is true.
- **Type II error**, or β, is the probability of failing to reject the null hypothesis when it is false.
- **Power**, or 1-β, is the probability of rejecting the null hypothesis when it is false.

	H_0 true	**H_0 false**
Fail to Reject H_0	**$1-\alpha$**	**β** **Type II error**
Reject H_0	**α** **Type I error**	**$1-\beta$** **Power**

Power is Determined by Three Factors

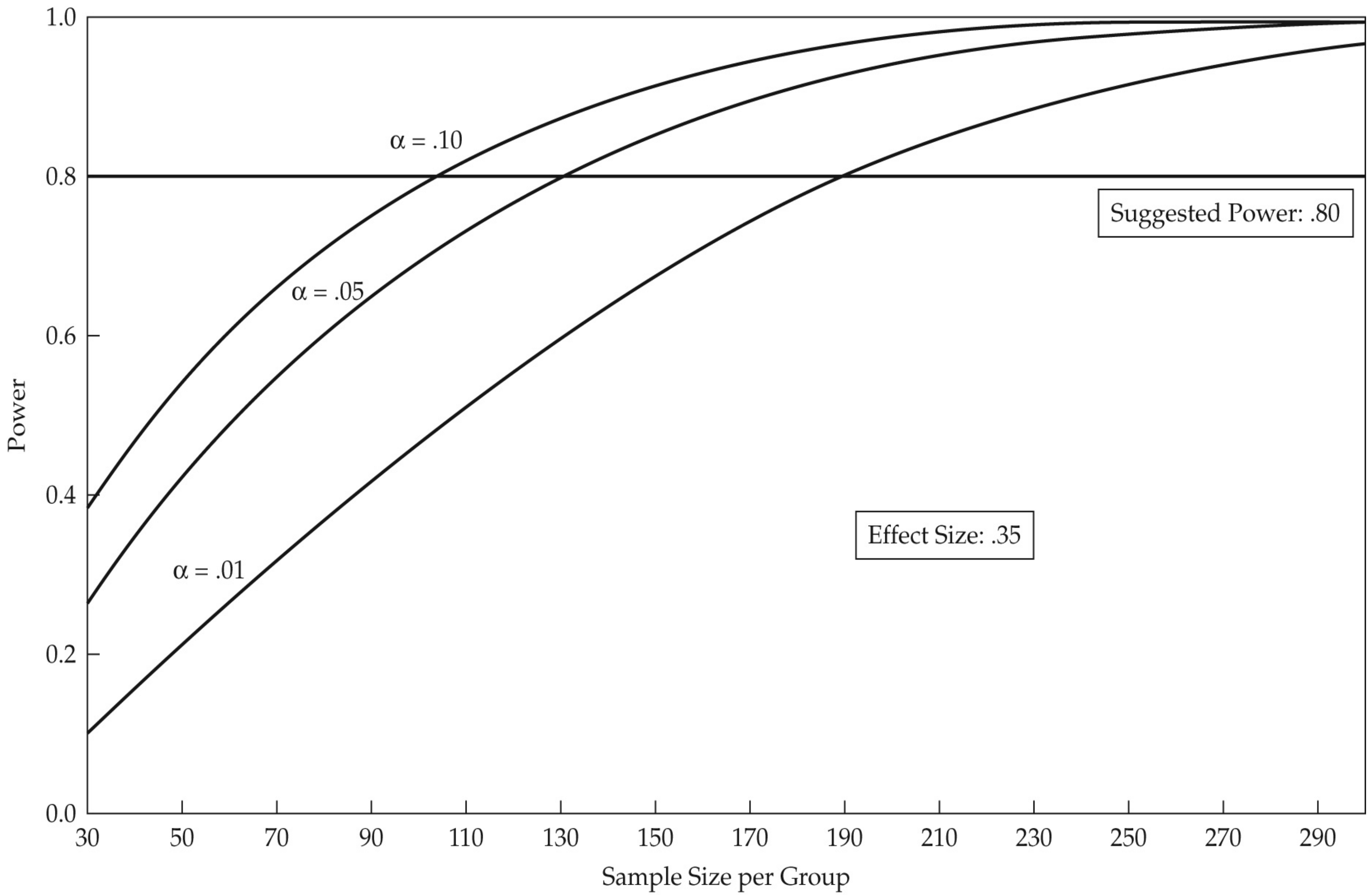
- Effect size: the actual magnitude of the effect of interest (e.g., the difference between means or the correlation between variables).
- Alpha (α): as α is set at smaller levels, power decreases. Typically, α = .05.
- Sample size: as sample size increases, power increases. With very large sample sizes, even very small effects can be statistically significant, raising the issue of practical significance vs. statistical significance.

Rules of Thumb 1–1

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.

Impact of Sample Size on Power



Definition

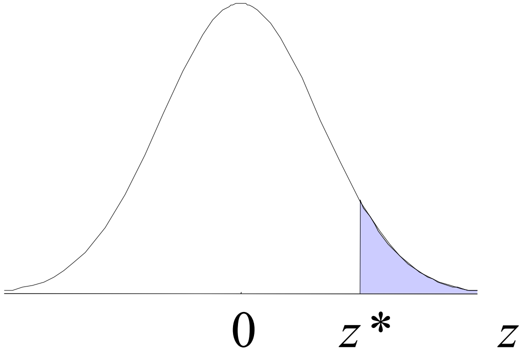
***P*-Value**

The ***P*-value** of a hypothesis test is the probability of getting sample data at least as inconsistent with the null hypothesis (and supportive of the alternative hypothesis) as the sample data actually obtained. We use the letter ***P*** to denote the *P*-value.

Finding *p*-Values

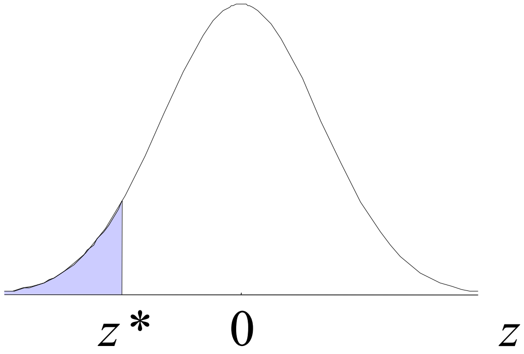
1. H_a contains $>$ (Right tail)

   p-value = $P(z > z^*)$



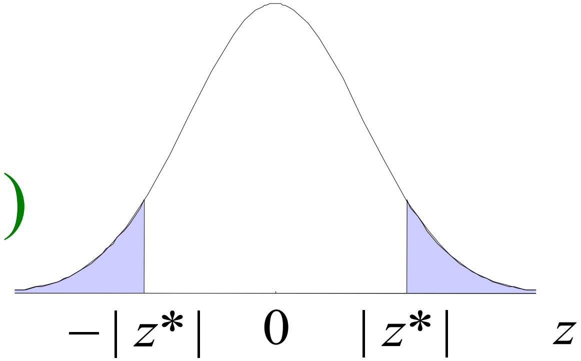
2. H_a contains $<$ (Left tail)

   p-value = $P(z < z^*)$



3. H_a contains $\neq$ (Two-tailed)

   p-value = $P(z < -|z^*|) + P(z > |z^*|)$



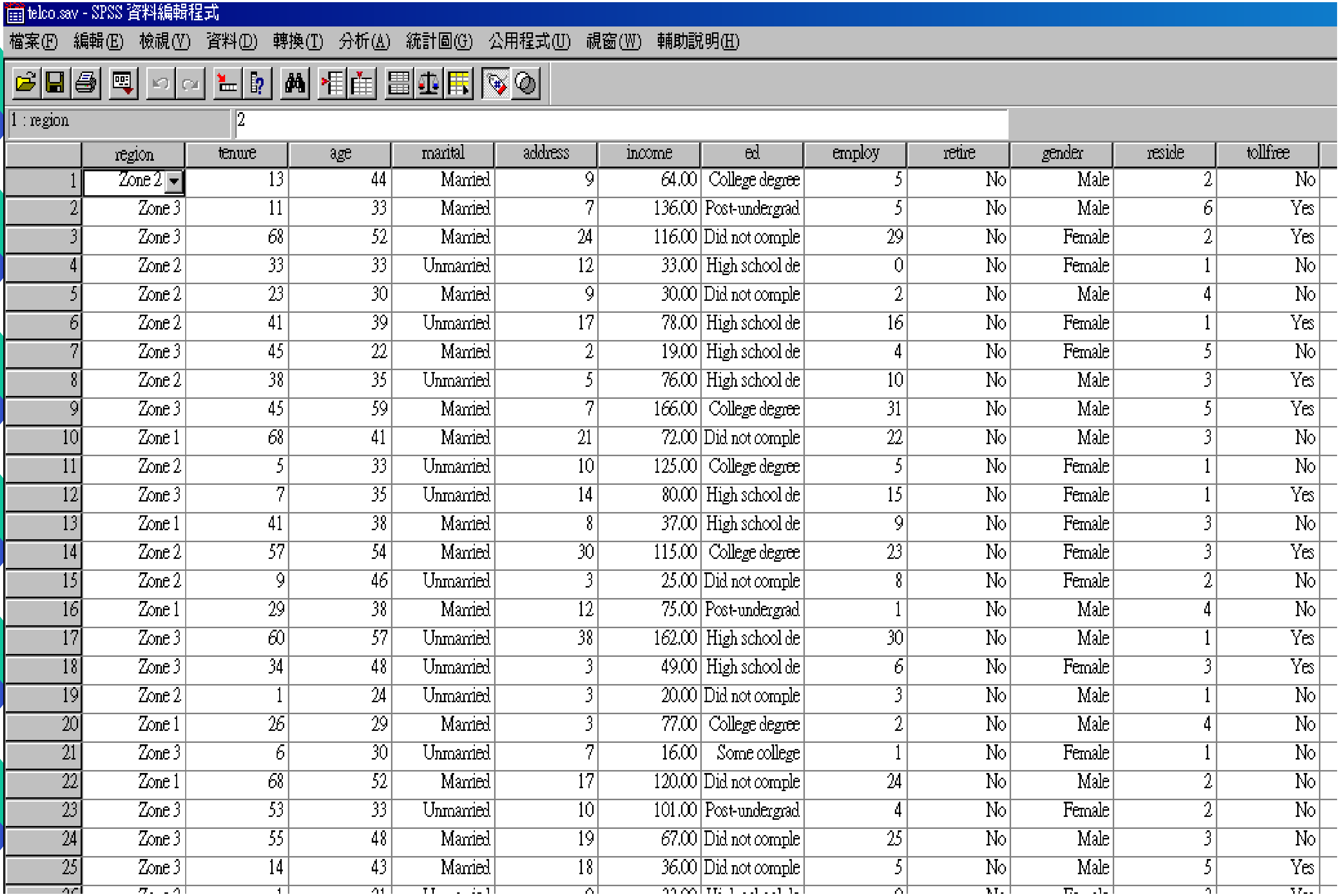
How to interpret P-values

Significance level	Reported probability	description
> 20%	P > 0.2	not significant
< 20%	P < 0.2	possibly significant
< 10%	P < 0.1	nearly significant
< 5%	P < 0.05	significant
< 1%	P < 0.01	very significant
< 0.5%	P < 0.005	highly significant
< 0.1%	P < 0.001	very highly significant

Examples

The basic starting point for any statistical analysis is a matrix of data. For most applications in the social sciences, this matrix will be a ***People*** × ***Variables*** array. But, the objects of measurement need not be people—they could be animals, work groups, cities, etc.

A Typical Example of a Data Set



1 : region 2

	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	tollfree
1	Zone 2	13	44	Married	9	64.00	College degree	5	No	Male	2	No
2	Zone 3	11	33	Married	7	136.00	Post-undergrad	5	No	Male	6	Yes
3	Zone 3	68	52	Married	24	116.00	Did not comple	29	No	Female	2	Yes
4	Zone 2	33	33	Unmarried	12	33.00	High school de	0	No	Female	1	No
5	Zone 2	23	30	Married	9	30.00	Did not comple	2	No	Male	4	No
6	Zone 2	41	39	Unmarried	17	78.00	High school de	16	No	Female	1	Yes
7	Zone 3	45	22	Married	2	19.00	High school de	4	No	Female	5	No
8	Zone 2	38	35	Unmarried	5	76.00	High school de	10	No	Male	3	Yes
9	Zone 3	45	59	Married	7	166.00	College degree	31	No	Male	5	Yes
10	Zone 1	68	41	Married	21	72.00	Did not comple	22	No	Male	3	No
11	Zone 2	5	33	Unmarried	10	125.00	College degree	5	No	Female	1	No
12	Zone 3	7	35	Unmarried	14	80.00	High school de	15	No	Female	1	Yes
13	Zone 1	41	38	Married	8	37.00	High school de	9	No	Female	3	No
14	Zone 2	57	54	Married	30	115.00	College degree	23	No	Female	3	Yes
15	Zone 2	9	46	Unmarried	3	25.00	Did not comple	8	No	Female	2	No
16	Zone 1	29	38	Married	12	75.00	Post-undergrad	1	No	Male	4	No
17	Zone 3	60	57	Unmarried	38	162.00	High school de	30	No	Male	1	Yes
18	Zone 3	34	48	Unmarried	3	49.00	High school de	6	No	Female	3	Yes
19	Zone 2	1	24	Unmarried	3	20.00	Did not comple	3	No	Male	1	No
20	Zone 1	26	29	Married	3	77.00	College degree	2	No	Male	4	No
21	Zone 3	6	30	Unmarried	7	16.00	Some college	1	No	Female	1	No
22	Zone 1	68	52	Married	17	120.00	Did not comple	24	No	Male	2	No
23	Zone 3	53	33	Unmarried	10	101.00	Post-undergrad	4	No	Female	2	No
24	Zone 3	55	48	Married	19	67.00	Did not comple	25	No	Male	3	No
25	Zone 3	14	43	Married	18	36.00	Did not comple	5	No	Male	5	Yes

Examples

cont.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1												
P_2												
P_3												
P_4												
P_5												
. . .												
P_N												

The variables (V_i's) can be continuous measures, categories represented by numbers, and transformations, products or combinations of other variables.

Determining the Weights

- Multiple Regression : Weights are determined to best correlate with the variable being predicted

- Discriminant Analysis: Weights are determined so to maximally differentiates amongst groups of observations.

Examples cont.

Nearly all statistical procedures—univariate and multivariate—are based on ***linear combinations***. Understanding that basic fact has far-reaching implications for using statistical procedures to their fullest advantage.

A linear combination (LC) is nothing more than a weighted sum of variables:

$$LC = W_1V_1 + W_2V_2 + . . . + W_KV_K$$

A very simple example is the total score on a questionnaire. The individual items on the questionnaire are the variables V_1 ,V_2 ,V_3 , etc. The weights are all set to a value of 1 (i.e., $W_1 = W_2 = . . . W_k = 1$).

Examples cont.

The items combined in a linear combination need not be variables. In statistics, the items combined are often people.

$$LC = W_1P_1 + W_2P_2 + . . . + W_KP_N$$

A good example is the sample mean. In this case the weights are set to the reciprocal of the sample size (i.e., $W_1 = W_2 = . . . W_k = 1/N$).

Examples cont.

Different statistical procedures derive the weights (***W***) in a linear combination to either maximize some desirable property (e.g., a correlation) or to minimize some undesirable property (e.g., error).

The weights are sometimes *assigned* rather than *derived* to produce linear combinations of particular interest.

An approach to MV modeling

1. Define the problem
   - Conceptual foundation comes first
   - Then choose the appropriate MV technique
2. Develop the analysis plan
   - Sample size
   - Allowable or required data types
3. Test the MV assumptions
4. Estimate the MV model & assess the overall fit
5. Interpret the variate
6. Validate the MV model

Outline

- Basics of Multivariate Data
- Exploration of Multivariate Data
- Methods for Multivariate Data

Basics of Multivariate Data

- Simultaneous measurements on multiple variables at the same spatial location
- Issues of multivariate data:
  - Understand relationship (distances)
  - Visualization
  - Redundant information? (correlations)
  - Models
  - Units …

Basics of Multivariate Data

- Data reduction or structural simplification
- Sorting and grouping
- Investigation of the dependence among variables
- Prediction
- Hypothesis construction and testing

Basics of Multivariate Data

- Simultaneous measurements on multiple variables at the same spatial location:
  - Observations y_{i1}, …,y_{ip} on p attributes at each of n sites s_i.
  - Matrix Y set of n points in p-dimensional data or attribute space (rather than geographic space)
  - Standardizations from Y to Z also referred to as ‘z scores’

Basics of Multivariate Data

- Standardizations for each variable are based on:
  1. Mean centered
  2. Divide by its standard deviation to achieve a comparable quantitative scale
- Issues:
  - Weighting imposed by original scale
  - Reduce numerical problems in computation

Outline

- ✓ Basics of Multivariate Data
- Exploration of Multivariate Data
- Methods for Multivariate Data

Exploration of Multivariate Data

- Descriptive statistics
- Visualization
- Distance calculations

Exploration of Multivariate Data

- Descriptive statistics with $j, k = 1,2,\ldots, p$:
  - Mean:

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$$

  - Variance:

$$\sigma_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ik} - \bar{x}_k\right)^2$$

  - Correlation:

$$r_{ik} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}} = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_{ik} - \bar{x}_k\right)^2}}$$

Exploration of Multivariate Data

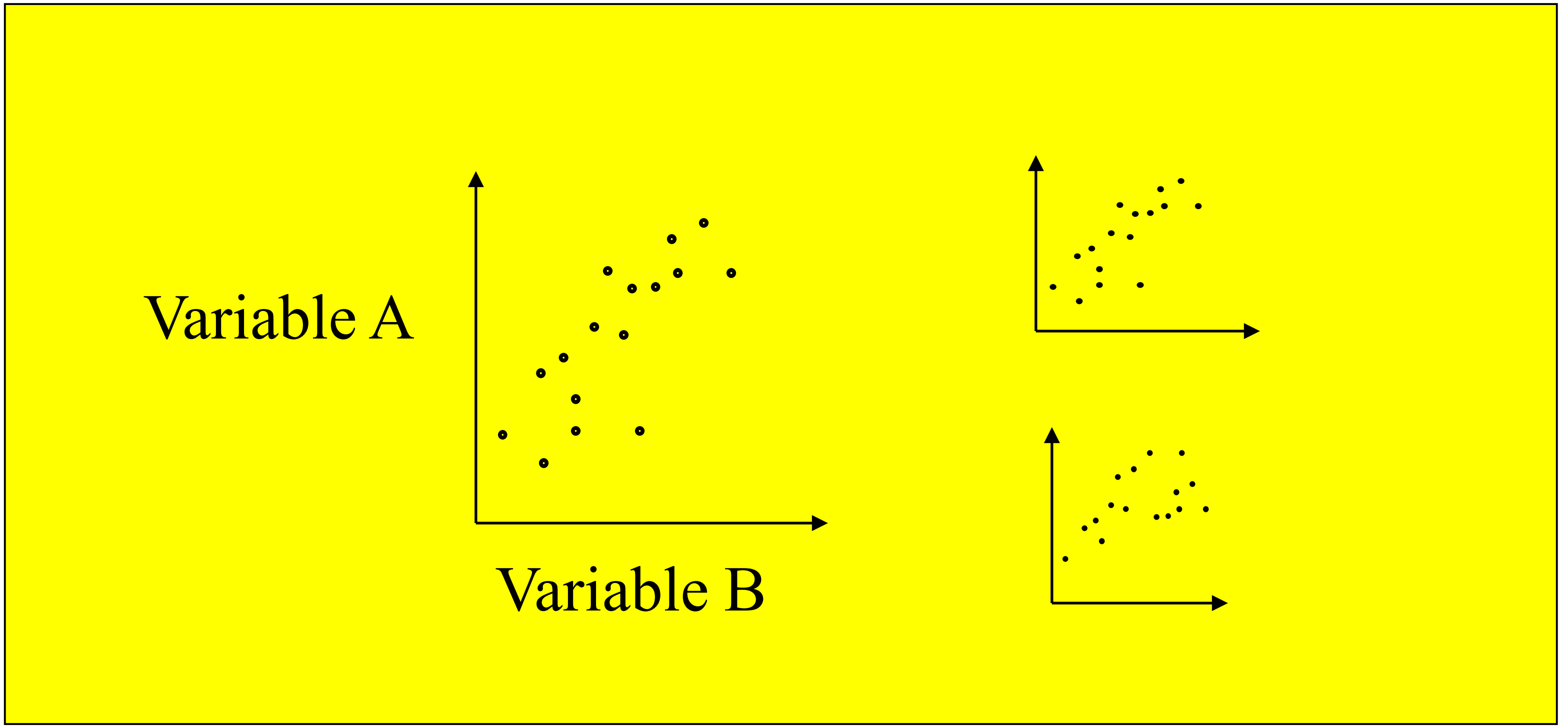
- Descriptive statistics hide non-linear relationships
- Are susceptible to outliers
- May indicate association when in reality little exists

Exploration of Multivariate Data

- Visualization:
  - Scatter plots between two variables at a time
  - Multiple scatter plots in array form
  - Multidimensional scatter plots

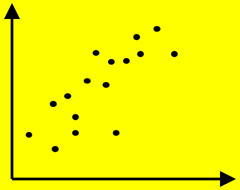
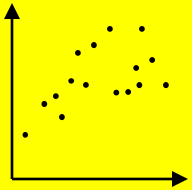
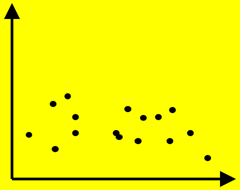
Exploration of Multivariate Data

- Visualization:
  - Scatter plots between two variables at a time



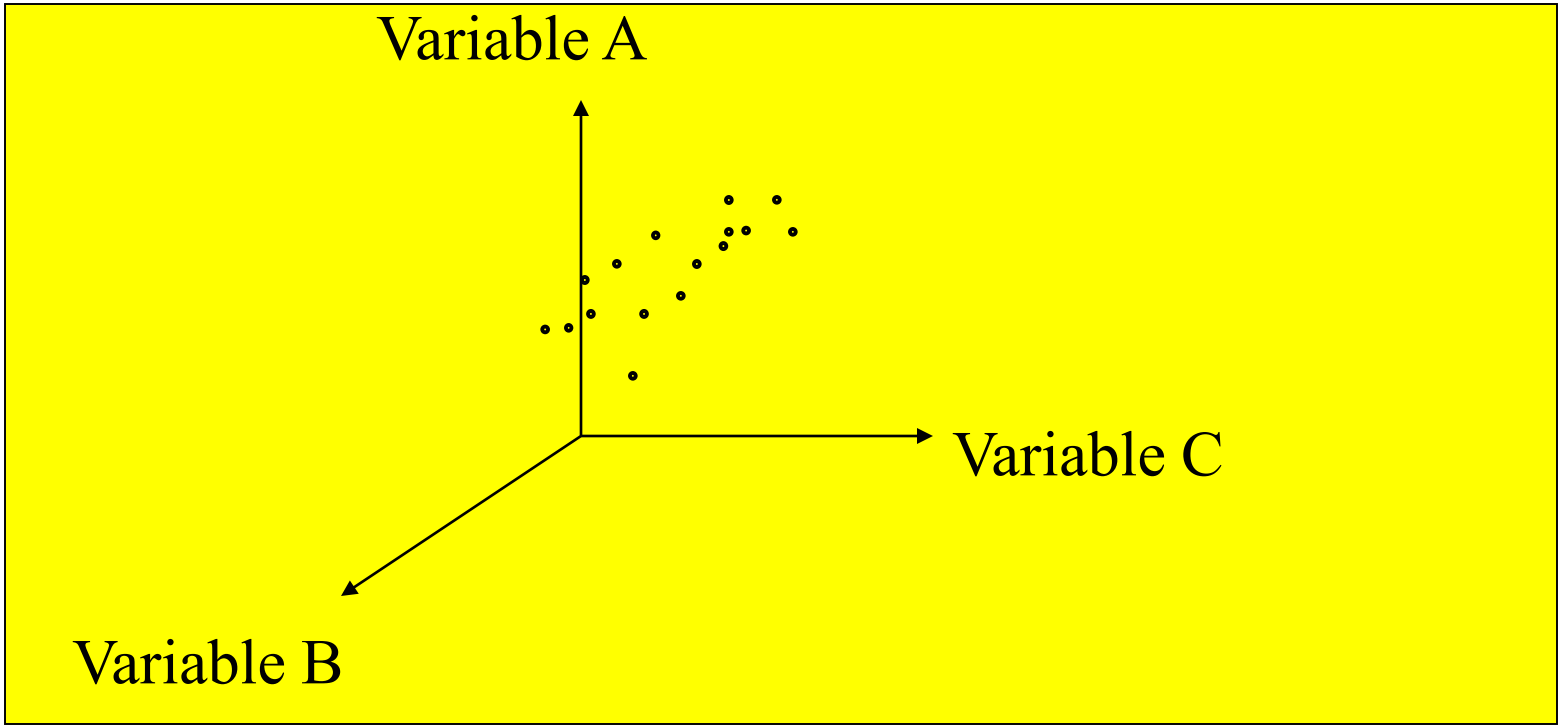
Exploration of Multivariate Data

- Visualization:
  - Multiple scatter plots in array form

	Variable A	B	C
A	Box plot		
B		or range	
C			or empty

Exploration of Multivariate Data

- Visualization:
  - Multidimensional scatter plots



Exploration of Multivariate Data

- Distance calculation based on Euclidian distance or?
  - Weighting to incorporate standard deviation
  - Rotation to incorporate dependence (correlation)

Outline

- ✓ Basics of Multivariate Data
- ✓ Exploration of Multivariate Data
- • Methods for Multivariate Data

Two Broad Types of Multivariate Methods:

1. Dependence – analyze dependent and independent variables at the same time.

2. Interdependence – analyze dependent and independent variables separately.

Classification of MV Techniques

1. Can the variables be divided into dependent and independent variables ?
2. If they can be divided, how many variables are to be treated as dependent ?
3. How are the variables, both dependent and independent, measured ?

- Dependence techniques apply if 1 or more dependent variables can be identified.
- Interdependence techniques apply if no single or group of variables can be identified as dependent

Types of Multivariate 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
- Canonical Correlation
- Conjoint Analysis
- Structural Equations Modeling (SEM)

Types of Multivariate Techniques

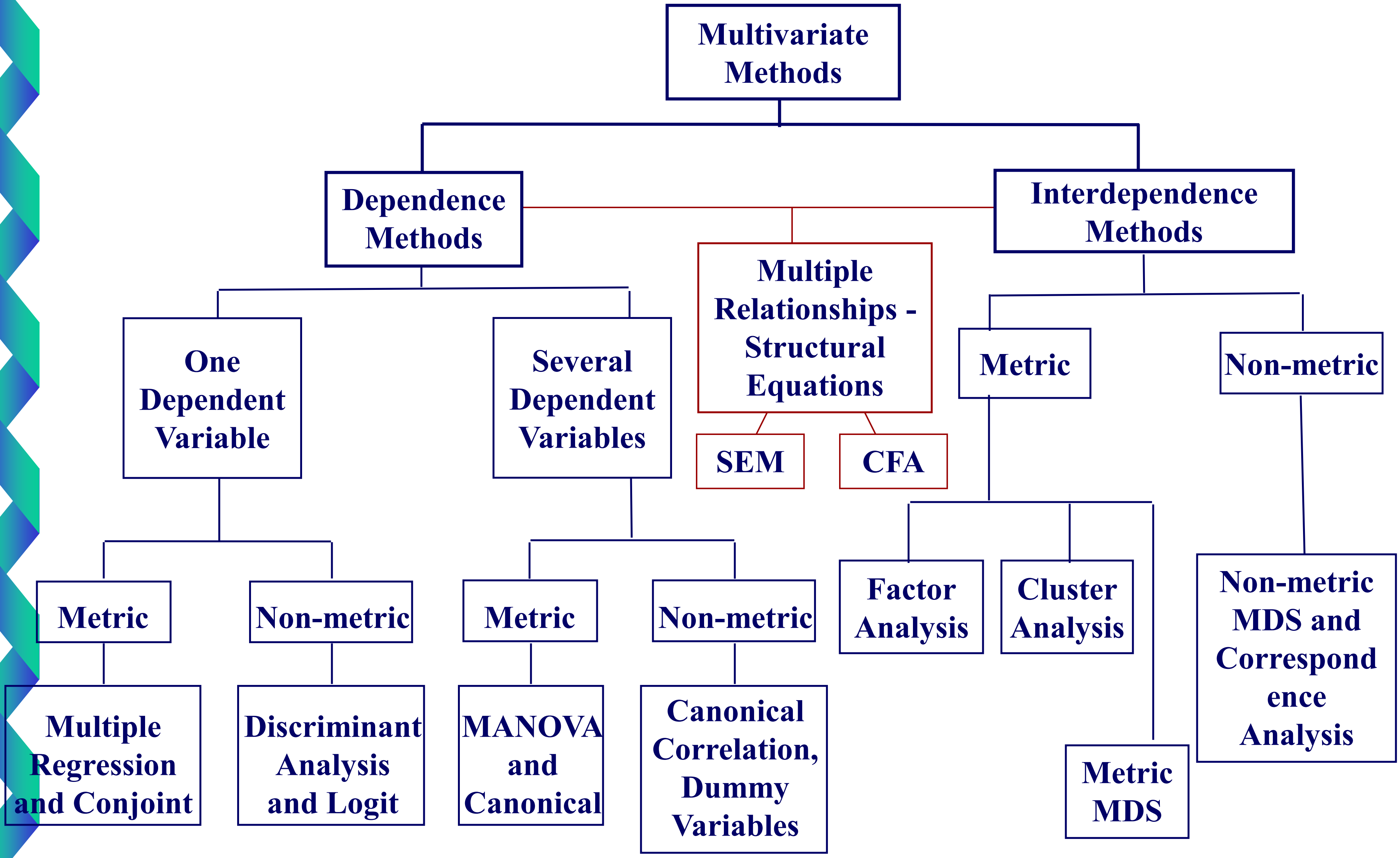
Interdependence techniques: involve the simultaneous analysis of all variables in the set, without distinction between dependent variables and independent variables.

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

Selecting a Multivariate Technique

1. What type of relationship is being examined – dependence or interdependence?
2. 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?
3. 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

SEM

CFA

One Dependent Variable

Several Dependent Variables

Metric

Non-metric

Metric

Non-metric

Metric

Non-metric

Multiple Regression and Conjoint

Discriminant Analysis and Logit

MANOVA and Canonical

Canonical Correlation, Dummy Variables

Factor Analysis

Cluster Analysis

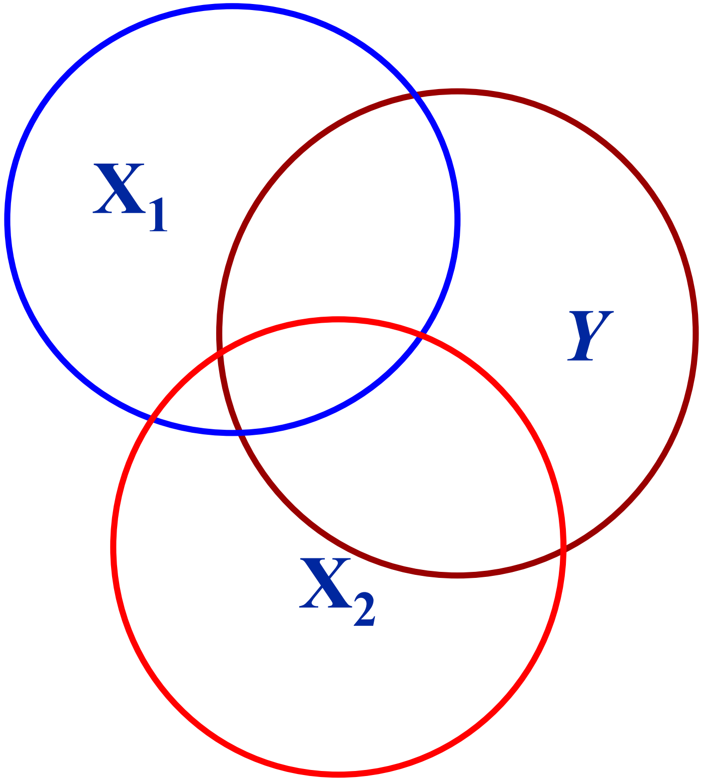
Metric MDS

Non-metric MDS and Correspond ence Analysis

Dependence techniques

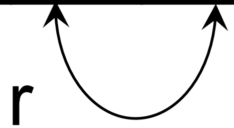
Multiple Regression

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

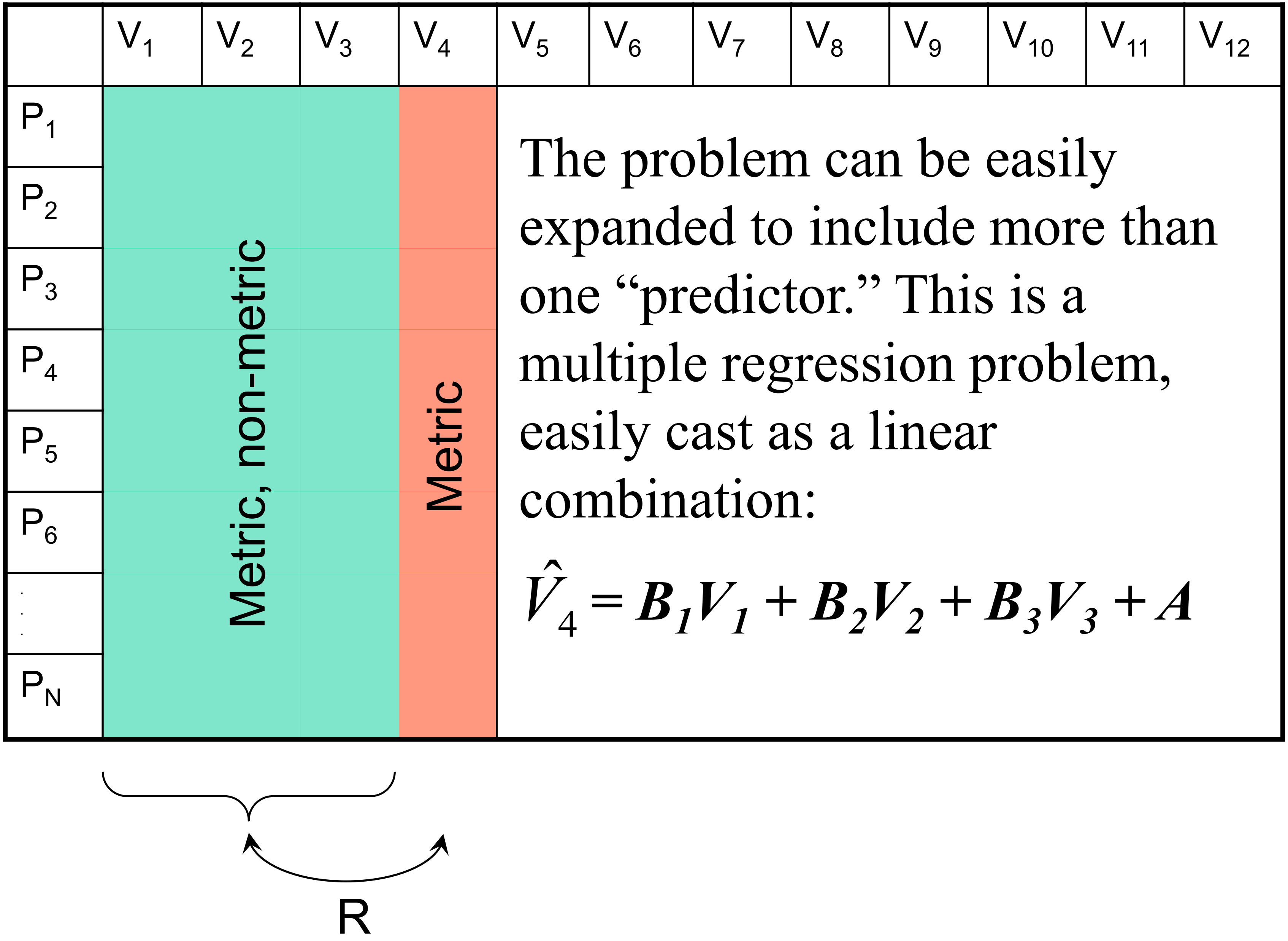


Remember Bivariate Correlation:

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1	Metric (continuous)	Metric (continuous)										
P_2												
P_3												
P_4												
P_5												
P_6												
⋮												
P_N												



The simplest possible inferential statistic: the bivariate correlation-involves just two variables.



The problem can be easily expanded to include more than one “predictor.” This is a multiple regression problem, easily cast as a linear combination:

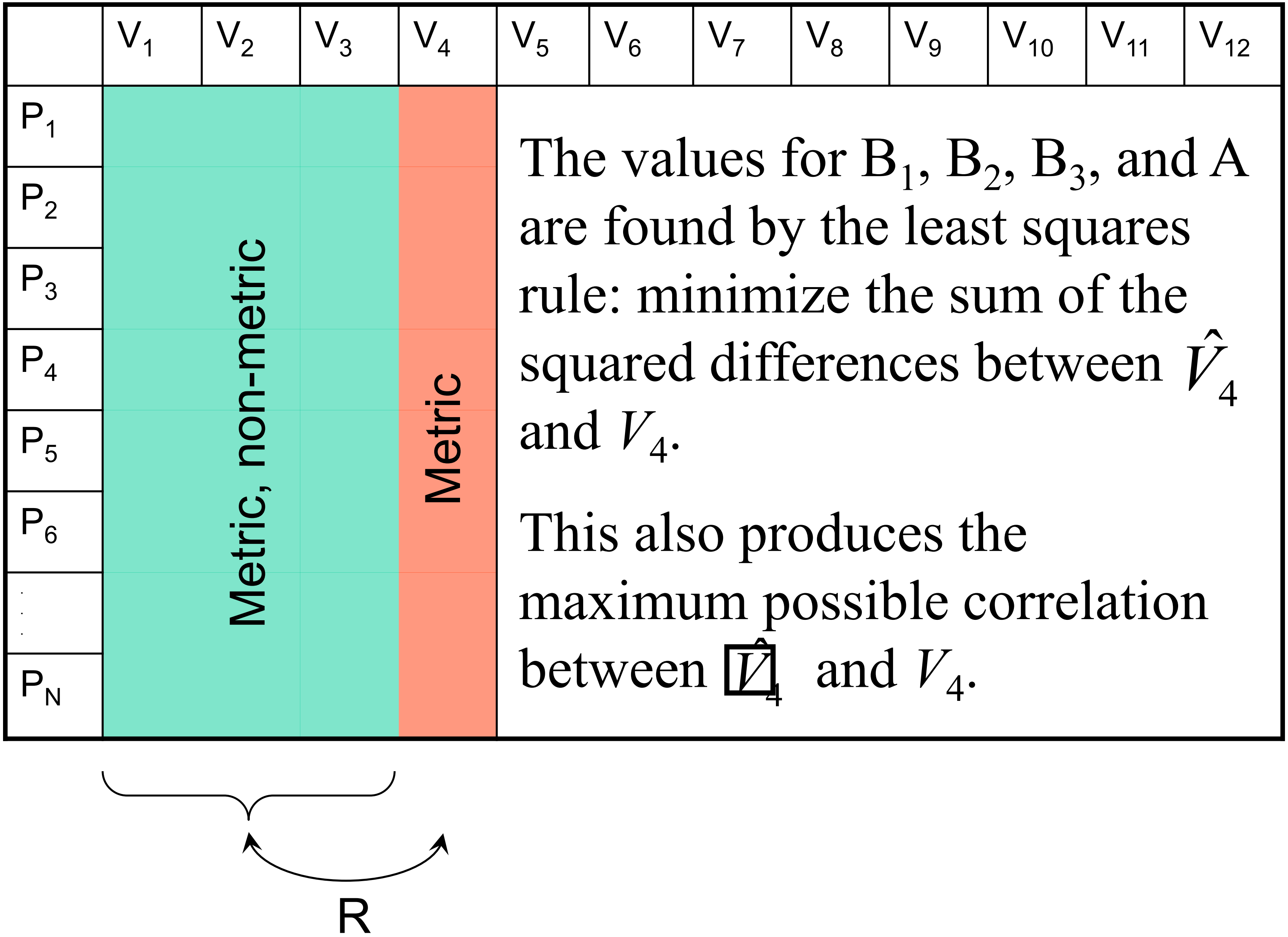
$$\hat{V}_4 = \boldsymbol{B_1 V_1} + \boldsymbol{B_2 V_2} + \boldsymbol{B_3 V_3} + \boldsymbol{A}$$

Once a linear combination is created, we also need to know something about its variability. Everything we need to know about the variability of a linear combination is contained in the

variance-covariance matrix of the original variables (S). The weights that are applied to create a linear combination can also be applied to S to get hat LC

$$r_{12} = \frac{\sigma_{12}}{(\sigma^2{}_1\sigma^2{}_2)^{½}}$$

	V_1	V_2	V_3	V_4	V_5
V_1	$s^2{}_1$	s_{12}	s_{13}	s_{14}	s_{15}
V_2	s_{21}	$s^2{}_2$	s_{23}	s_{24}	s_{25}
V_3	s_{31}	s_{32}	$s^2{}_3$	s_{34}	s_{35}
V_4	s_{41}	s_{42}	s_{43}	$s^2{}_4$	s_{45}
V_5	s_{51}	s_{52}	s_{53}	s_{54}	$s^2{}_5$



The values for B_1, B_2, B_3, and A are found by the least squares rule: minimize the sum of the squared differences between $\hat{V}_4$ and V_4.

This also produces the maximum possible correlation between $\boxed{\hat{V}}_4$ and V_4.

Discriminant Analysis

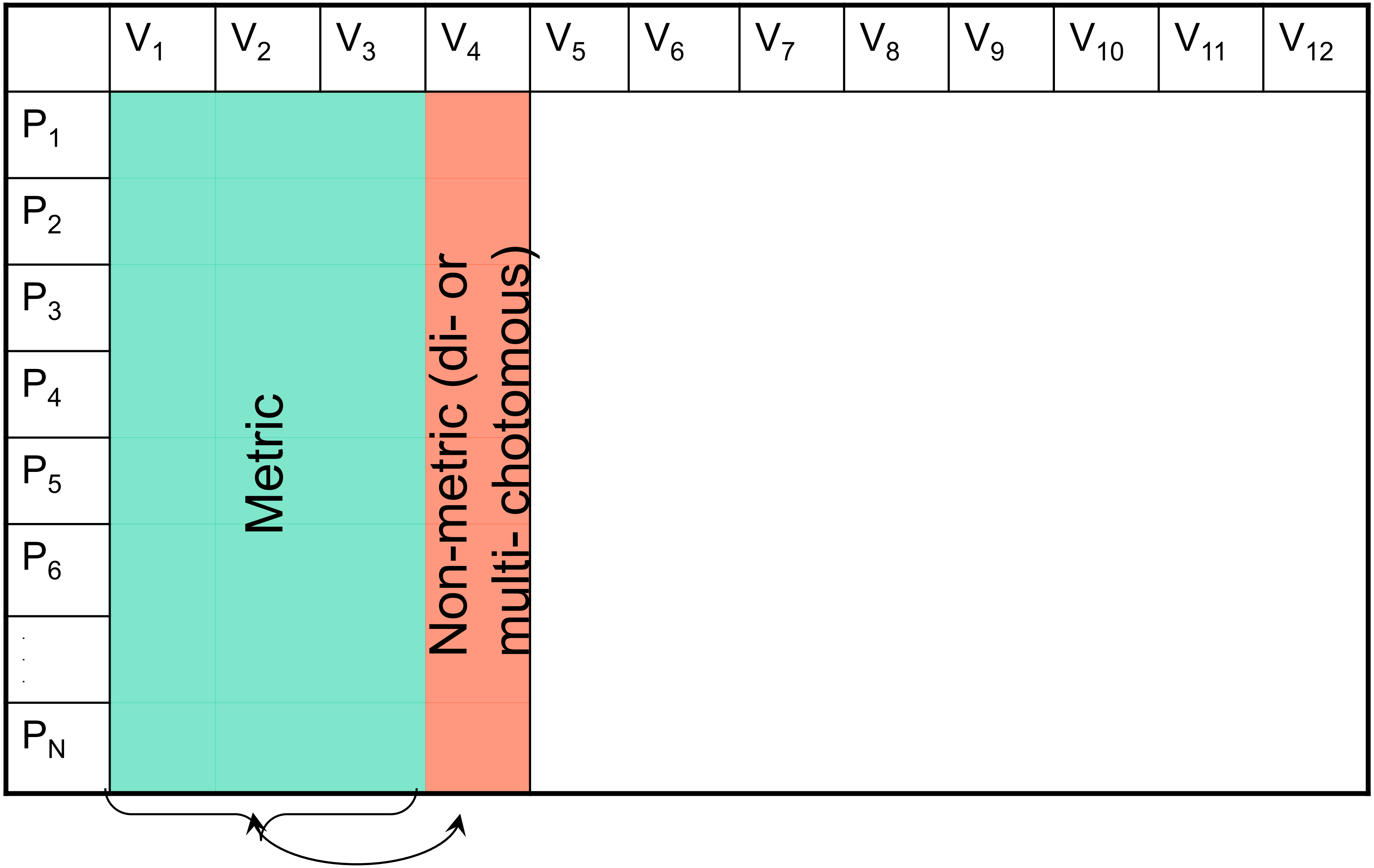
- What is it?

. . . 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



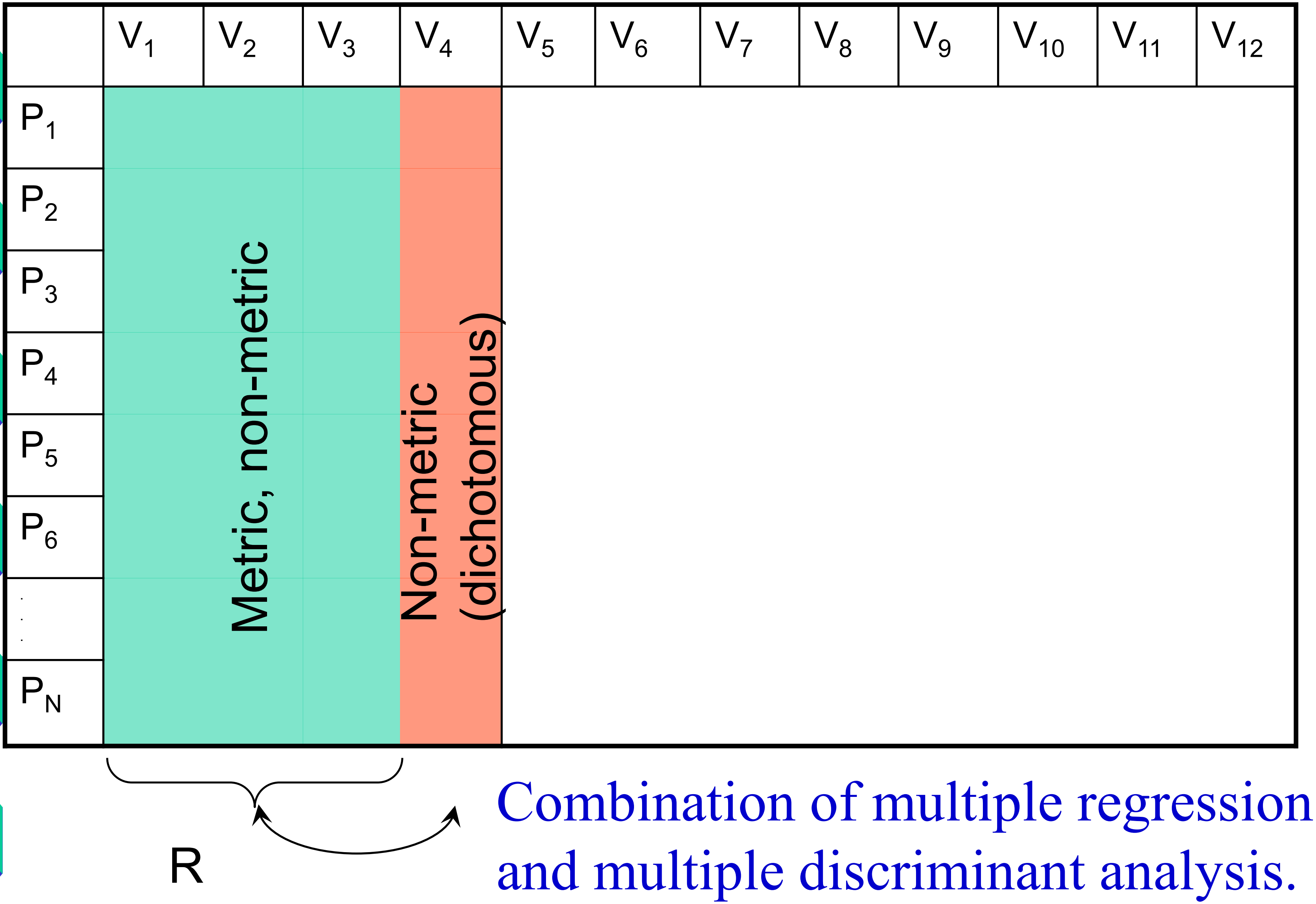
	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1	Metric			Non-metric (di- or multi- chotomous)								
P_2												
P_3												
P_4												
P_5												
P_6												
· · ·												
P_N												

Data can be grouped in terms of the dependent variable, base on the independent variables.

Logistic Regression (Logit analysis)

A single non-metric dependent variable is predicted by several metric independent variables. This technique is similar to discriminant analysis, but relies on calculations more like regression.

Logit: remember multiple regression



V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12

P1 P2 P3 P4 P5 P6 . . . PN

Metric, non-metric

Non-metric (dichotomous)

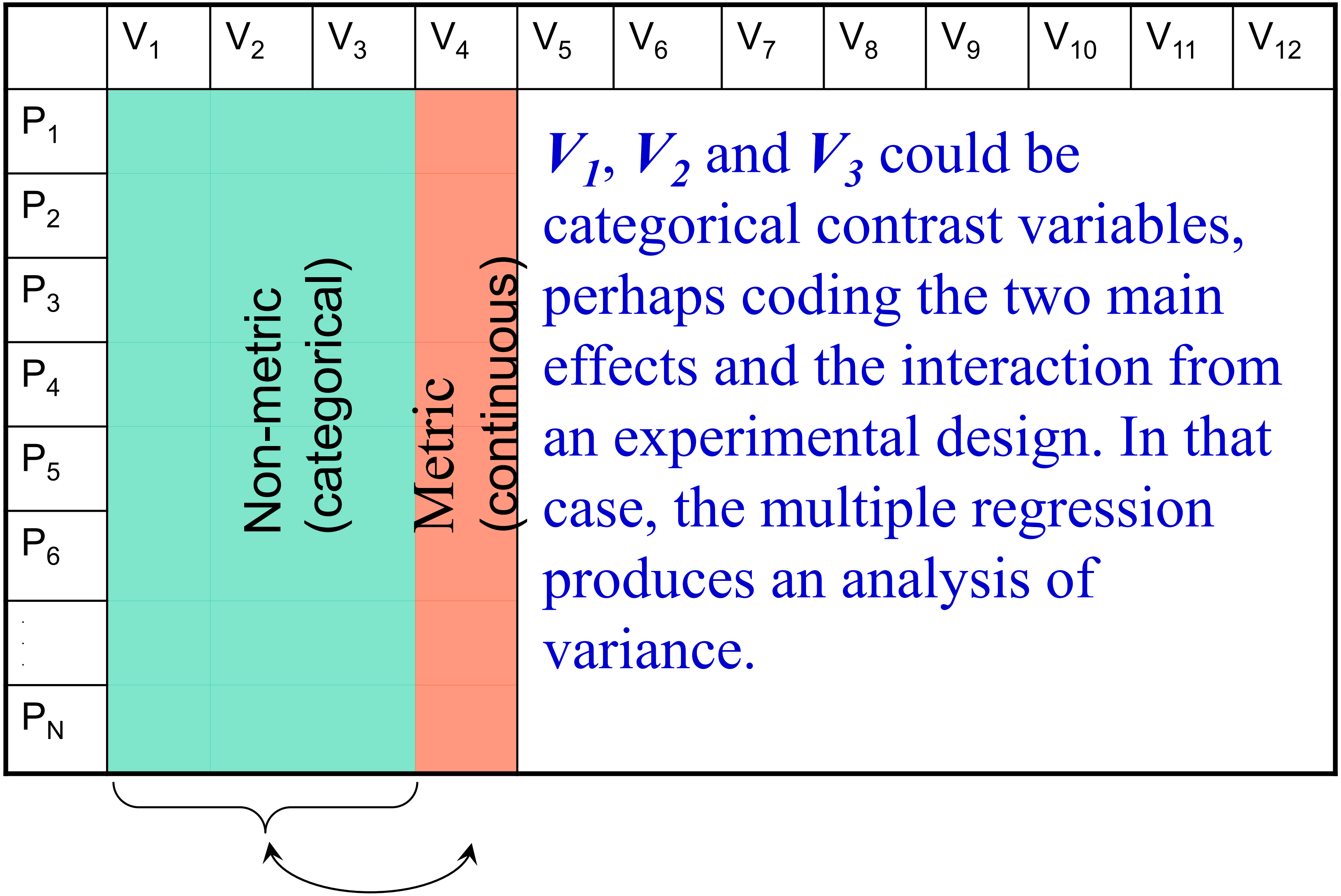
R

Combination of multiple regression and multiple discriminant analysis.

MANOVA

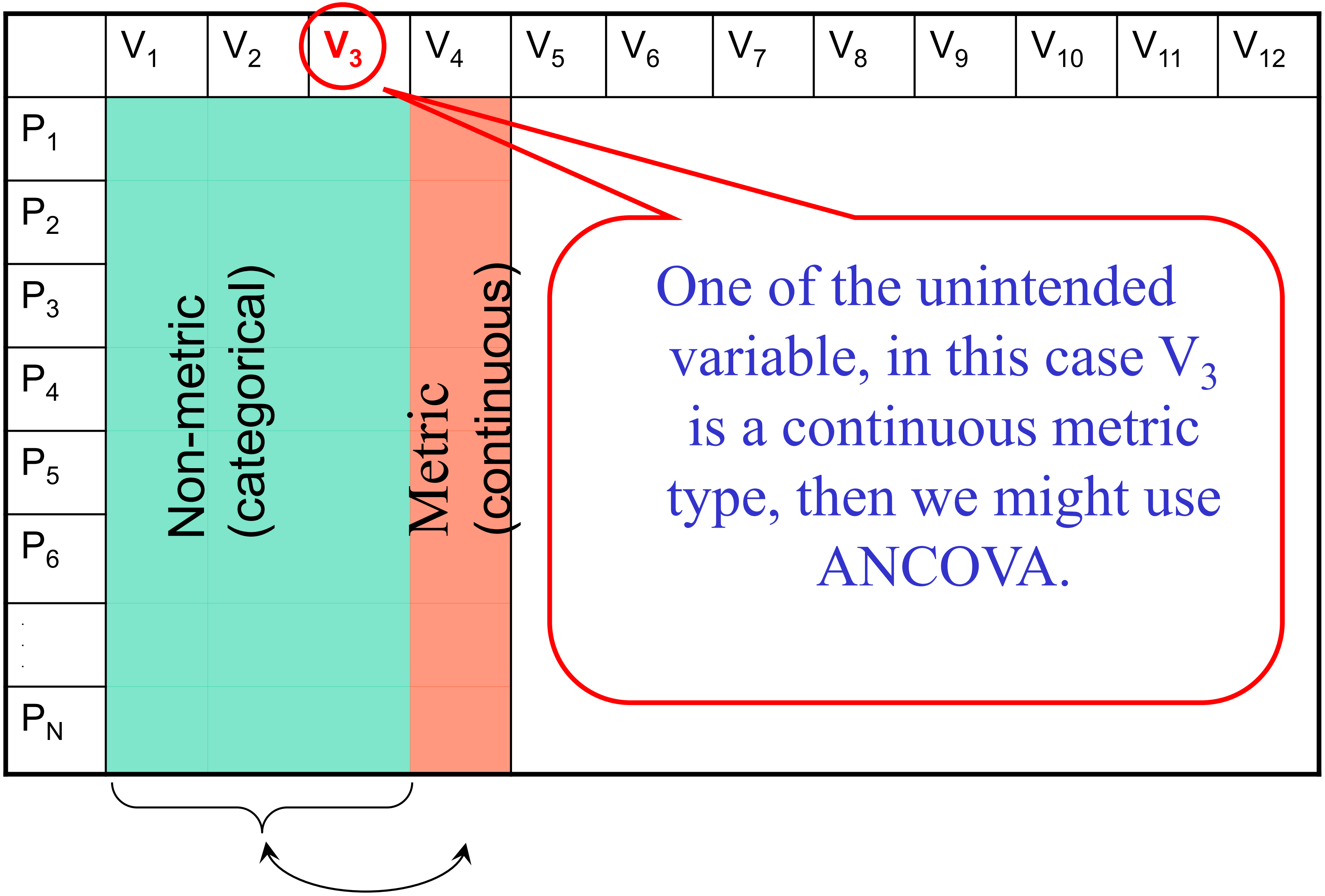
Several metric dependent variables are predicted by a set of non-metric (categorical) independent variables.

Remember k-ANOVA:



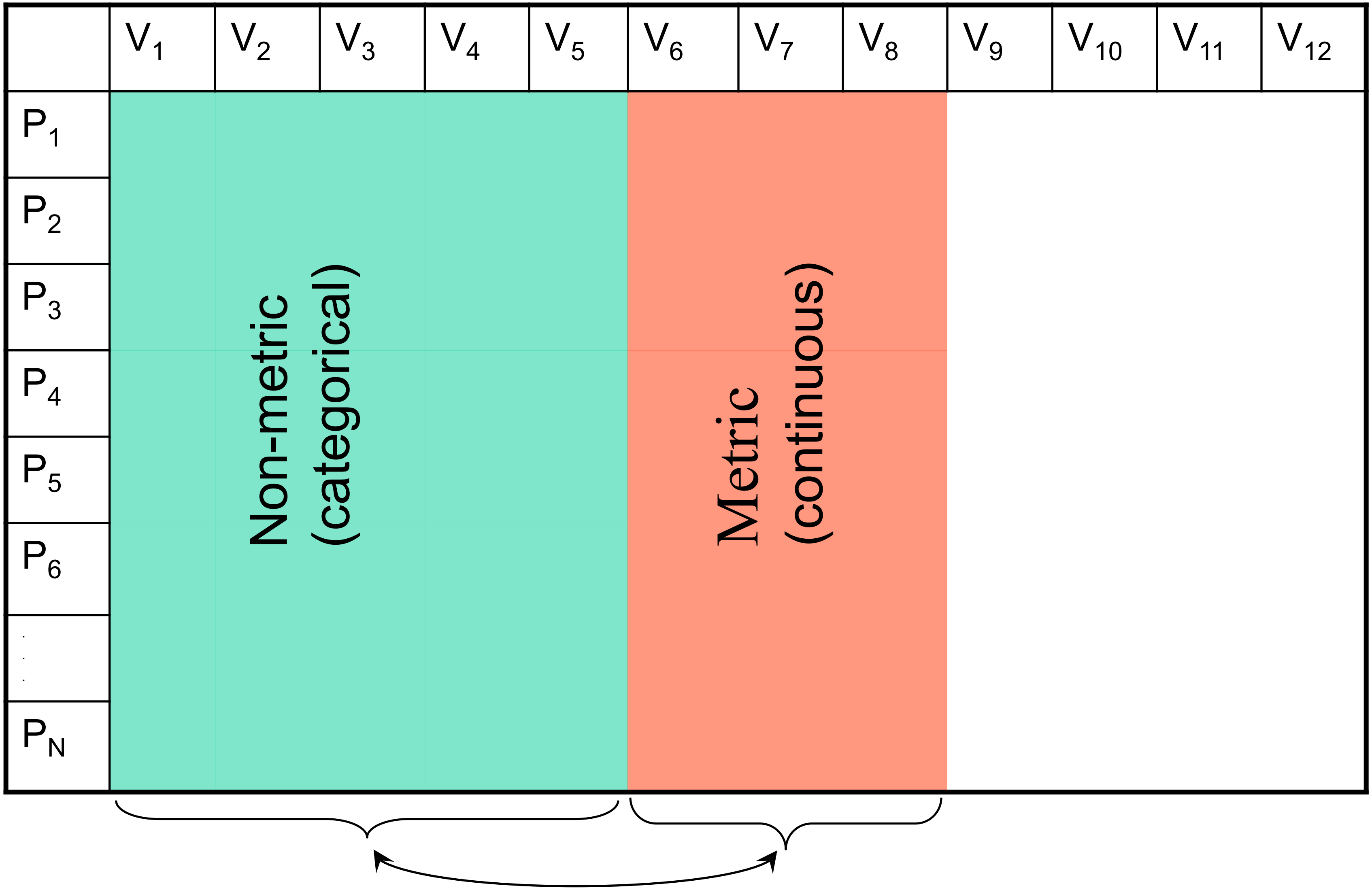
V_1, ***V_2*** and ***V_3*** could be categorical contrast variables, perhaps coding the two main effects and the interaction from an experimental design. In that case, the multiple regression produces an analysis of variance.

ANCOVA:



One of the unintended variable, in this case V_3 is a continuous metric type, then we might use ANCOVA.

MANOVA



	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1												
P_2												
P_3												
P_4												
P_5												
P_6												
. . .												
P_N												

Non-metric (categorical)

Metric (continuous)

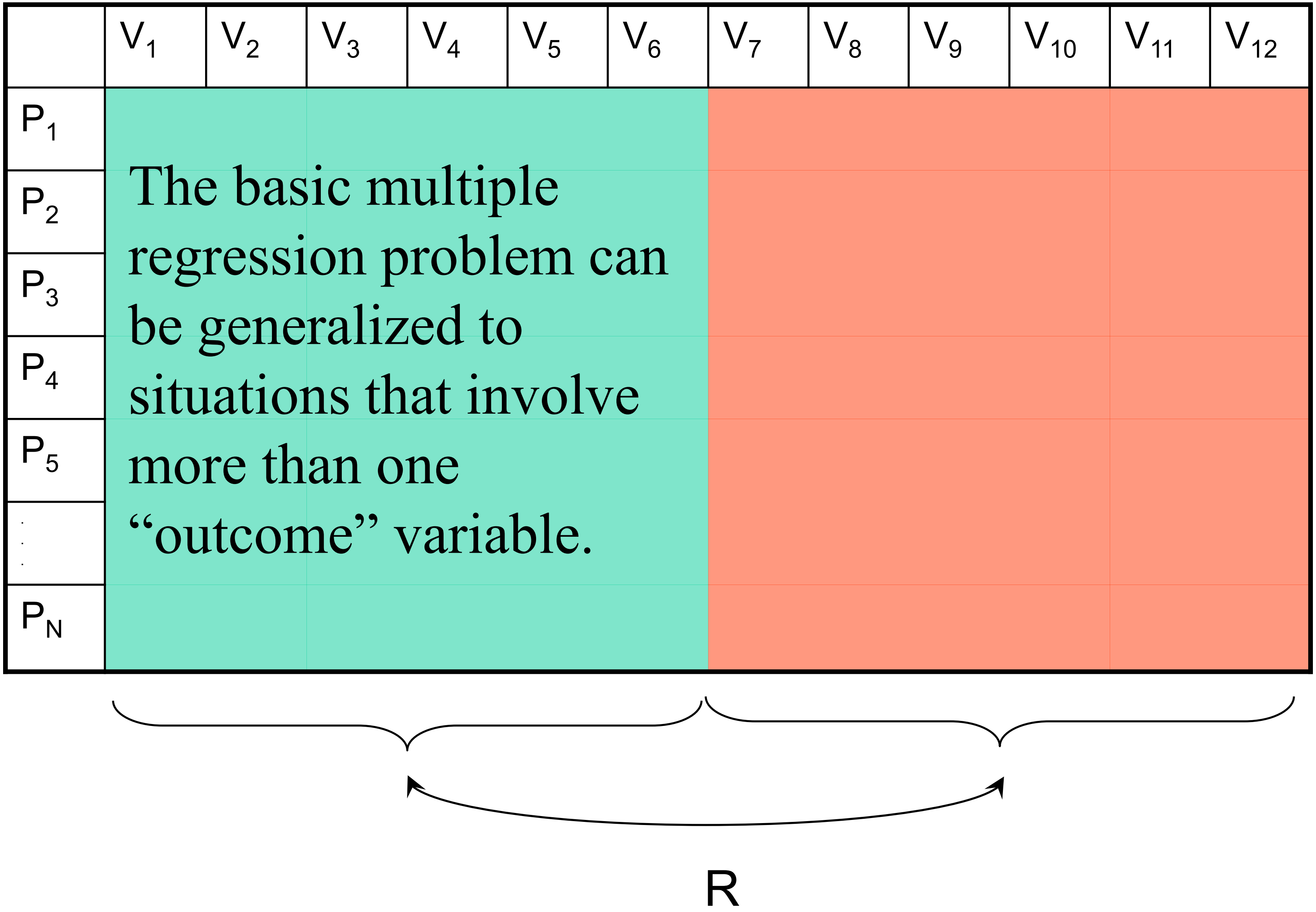
CANONICAL ANALYSIS

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

Canonical Variates

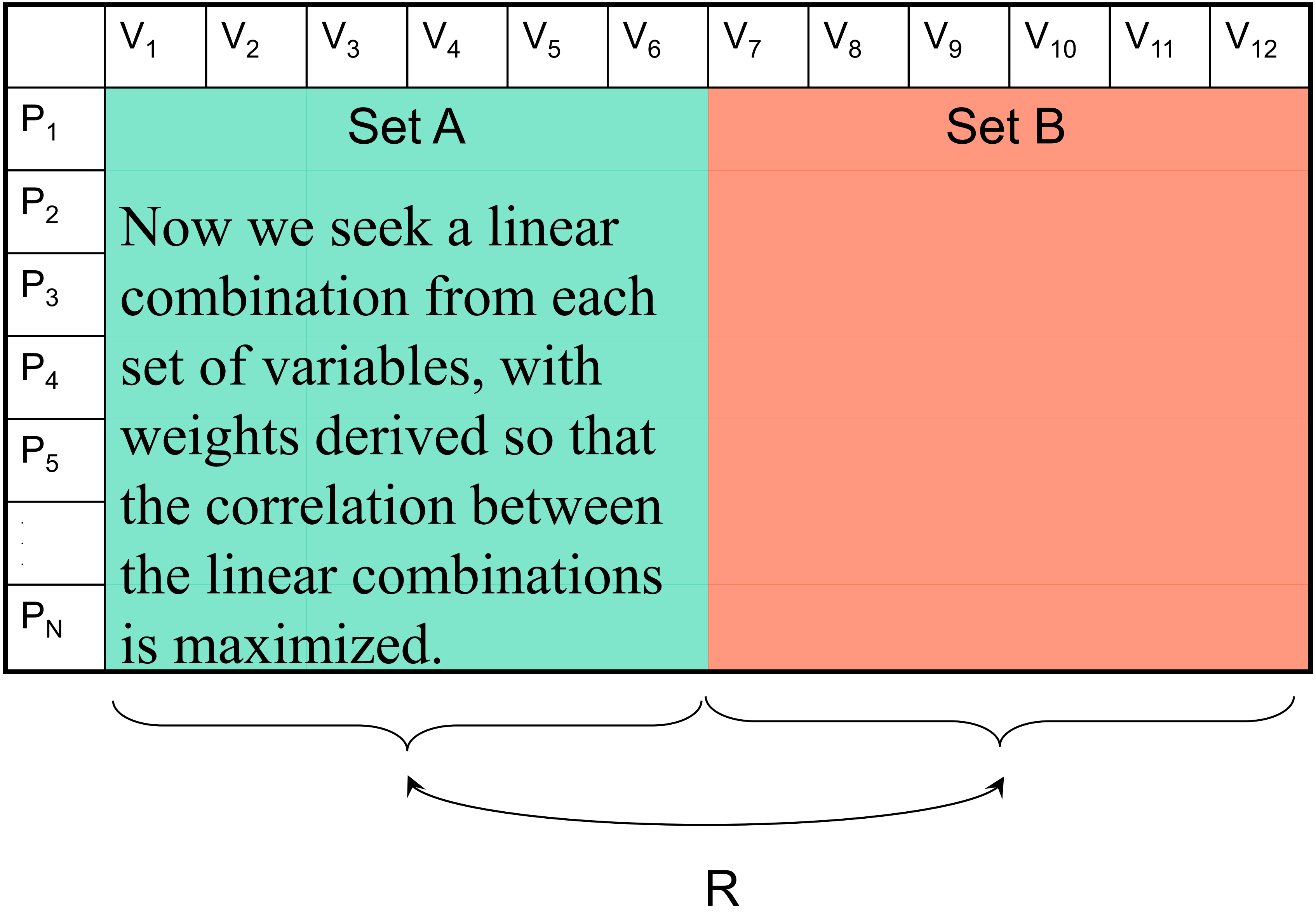
- Rather than the structure (correlation) canonical variates aims at highlighting differences in a pre-defined known structure.
- Objective: Group original data into groups and make them as statistically distinct as possible.

$$u_r = a_{r1} y_1 + a_{r2} y_2 + \ldots + a_{rp} y_p$$

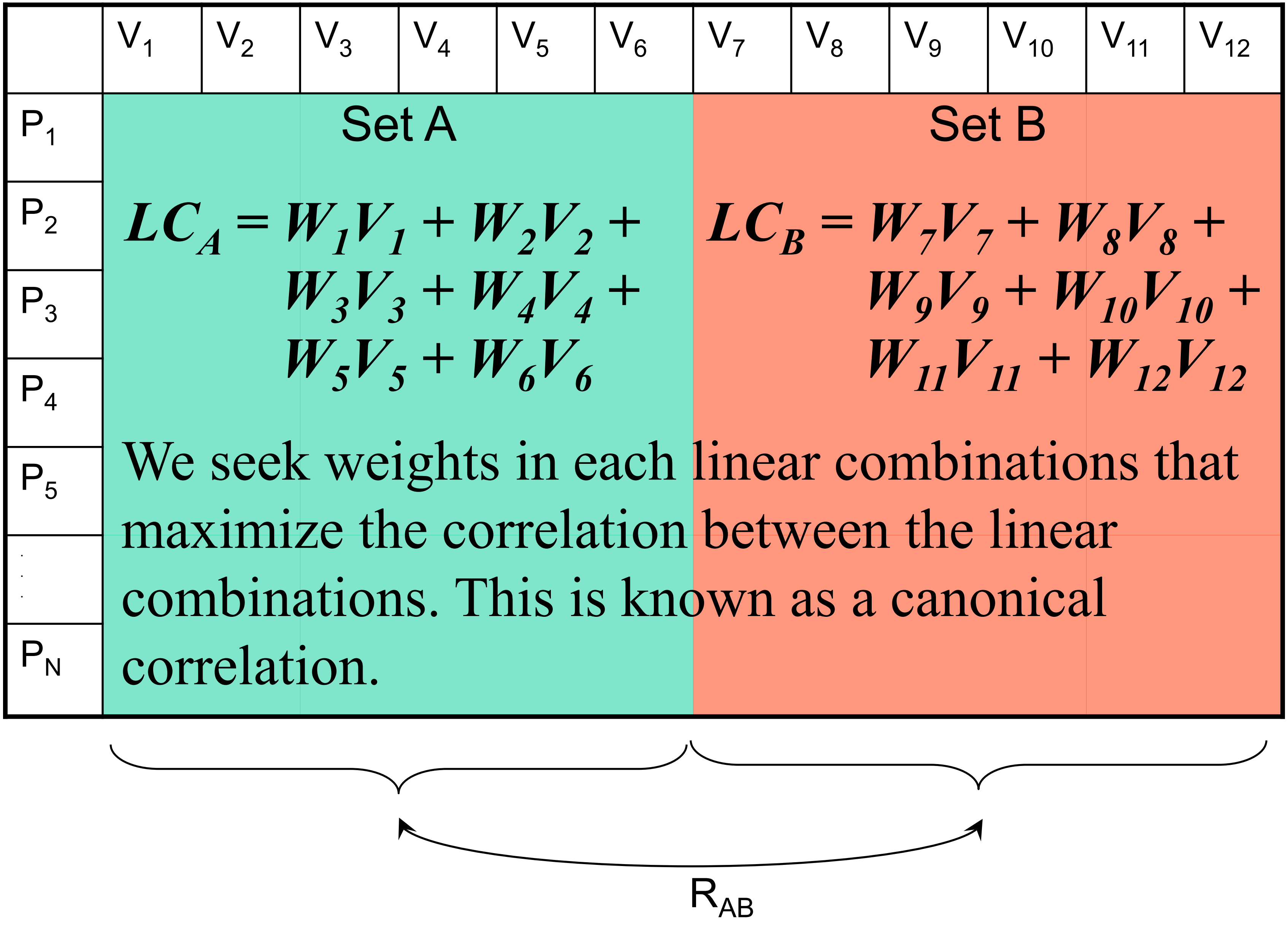


	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1												
P_2												
P_3												
P_4												
P_5												
⋮												
P_N												

The basic multiple regression problem can be generalized to situations that involve more than one “outcome” variable.



Now we seek a linear combination from each set of variables, with weights derived so that the correlation between the linear combinations is maximized.



	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1												
P_2												
P_3												
P_4												
P_5												
. . .												
P_N												

$$LC_A = W_1V_1 + W_2V_2 + W_3V_3 + W_4V_4 + W_5V_5 + W_6V_6$$

$$LC_B = W_7V_7 + W_8V_8 + W_9V_9 + W_{10}V_{10} + W_{11}V_{11} + W_{12}V_{12}$$

We seek weights in each linear combinations that maximize the correlation between the linear combinations. This is known as a canonical correlation.

CONJOINT ANALYSIS

. . . is used to understand respondents' preferences for products and services.

In doing this, it determines the importance of both:

1. attributes

and

2. levels of attributes

. . . based on a smaller subset of combinations of attributes and levels.

CONJOINT ANALYSIS

Typical Applications:

- Soft Drinks
- Candy Bars
- Cereals
- Beer
- Apartment Buildings; Condos
- Solvents; Cleaning Fluids

Structural Equations Modeling (SEM)

Estimates multiple, interrelated dependence relationships based on two components:

1. Measurement Model
2. Structural Model

Interdependence techniques

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1												
P_2												
P_3												
P_4												
P_5												
⋮												
P_N												

Sometimes we are not interested in relations between sets of variables but instead focus on a single set and seek a linear combination that has desirable properties.

Exploratory Factor Analysis

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

Principle Components Analysis

- Transforms data space into orthogonal space of most variance or “separation”
- More later …

Common Factor Analysis

- Based on the assumption that the observed correlations between the attributes $y_{i1}, \ldots, y_{ip}$ are mainly the result of some *a priori* underlying regularity or structure.
- Structure responsible for common factors—whereas, remainder unique (uncorrelated error terms).

Common Factor Analysis

- Represent the original variables $y_{i1}, \ldots, y_{ip}$ as linear combinations of the r hypothesized common factors $u_1, \ldots, u_r$ plus a residual ε:

$$y_1 = a_{11}u_1 + a_{12}u_2 + \ldots + a_{1r}u_r + \varepsilon_1$$

$$y_2 = a_{21}u_1 + a_{22}u_2 + \ldots + a_{2r}u_r + \varepsilon_2$$

$$\vdots$$

$$y_p = a_{p1}u_1 + a_{p2}u_2 + \ldots + a_{pr}u_r + \varepsilon_p$$

Common Factor Analysis

- Results mathematically more complex than PCA.
- Furthermore, results are not unique.
- Solution can be interpreted similar to PCA:
  - r useful reduction in the system
- Are factors conceptually sound?

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1												
P_2												
P_3												
P_4												
P_5												
⋮												
P_N												

For example, we might seek a linear combination of V_1 through V_{12} that captures most of the key information in those variables. If such a linear combination exists, we could replace 12 variables with 1 new variable, simplifying other analyses.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}
P_1												
P_2												
P_3												
P_4												
P_5												
⋮												
P_N												

Or we might wonder how many “dimensions” underlie the 12 variables. These multiple dimensions also would be represented by linear combinations, perhaps constrained to be uncorrelated.

These questions are addressed in principal components analysis and factor analysis.

Other Independent Techniques

The key idea is that the original data matrix can be transformed using linear combinations to provide useful ways to summarize the data and to test hypotheses about how the data are structured.

Sometimes the linear combinations are of variables and sometimes they are of people.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	. . .	P_N
V_1	🟨	🟨	🟥	🟥	🟥	🟧	🟧					
V_2	🟨	🟨	🟥	🟥	🟥	🟧	🟧					
V_3	🟨	🟨	🟥	🟥	🟥	🟧	🟧					
V_4	🟨	🟨	🟥	🟥	🟥	🟧	🟧					
V_5	🟨	🟨	🟥	🟥	🟥	🟧	🟧					
. . .	🟨	🟨	🟥	🟥	🟥	🟧	🟧					
V_K	🟨	🟨	🟥	🟥	🟥	🟧	🟧					

Approaches such as multidimensional scaling and cluster analysis can address such questions. These are conceptually similar to principal components analysis, but on a transposed matrix.

Cluster Analysis

. . . 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.

Cluster Analysis (numerical taxonomy)

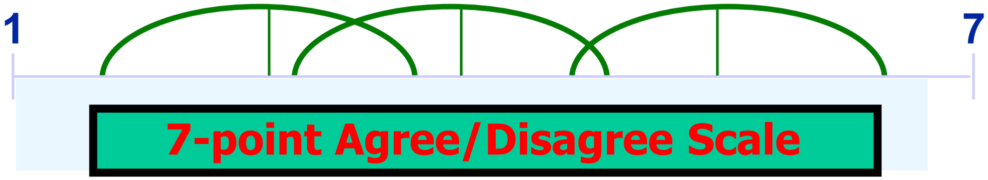
- Grouping into clusters based on dissimilarity measure. For example, first perform PCA, then calculate dissimilarity matrix based on the Euclidian distances.
- Break observations into groups based on:
  - Hierarchical clustering or
  - Minimum spanning tree
  - …

Cluster Analysis of "Eating Out" Questions

1. I eat at fast food restaurants at least once a week.
2. I prefer restaurants with the highest quality food.
3. I prefer restaurants that have quick service.
4. I prefer to eat at restaurants that have a nice atmosphere.

Objective: Identify groups that maximize ratio of

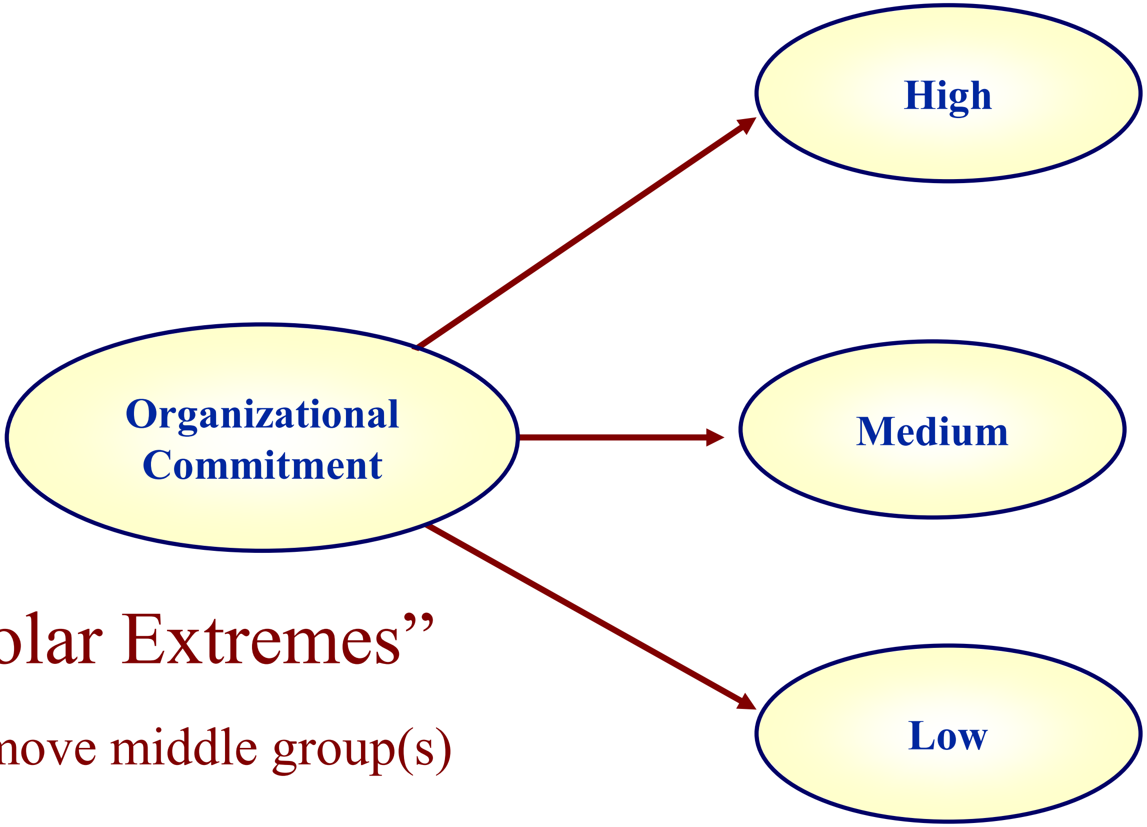
$$\frac{\text{between groups variance}}{\text{within groups variance}} = \frac{\text{large}}{\text{small}}$$



Example: Cluster Analysis . . .

Constructs . . .

- ✓ Trust
- ✓ Commitment
- ✓ Cooperation
- ✓ Locus of Control
- ✓ Job Satisfaction
- ✓ Turnover



“Polar Extremes”

= remove middle group(s)

Multidimensional Scaling

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

- similarities or
- preferences

and transforms these into distances represented as perceptual maps.

Multidimensional Scaling

- Unlike PCA and factor analysis that start with an *n* x *p* matrix, metric and non-metric scaling techniques (generally referred to as multidimensional scaling) start with an *n* x *n* matrix, *D*, representing some sort of dissimilarity for pairs of observations.
- For example, *p*-dimensional Euclidian distance (in this case metric or principal coordinates analysis)

Multidimensional Scaling

- Example:
  - Measured three parameters of different brands of PC:

    Prices , Features, & Calculating powers
  - Represent distances in a map not based on Euclidian distance between sample points but based on dissimilarities in variables—reflecting differing importance of the three characteristics.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	. . .	P_N
V_1												
V_2												
V_3												
V_4												
V_5												
⋮												
V_K												

Transposes of the original matrix and then evaluate.

Sometimes we might shift the status of "people" and "variables" in our analysis. Our interest might be in whether a smaller number of dimensions or clusters might underlie the larger collection of people.

Correspondence Analysis

. . . 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.

A Structured Approach to Multivariate Model Building:

1. Define the Research Problem, Objectives, and Multivariate Technique(s) to be Used
2. Develop the Analysis Plan
3. Evaluate the Assumptions Underlying the Multivariate Technique(s)
4. Estimate the Multivariate Model and Assess Overall Model Fit
5. Interpret the Variate(s)
6. Validate the Multivariate Model

Guidelines for Multivariate Analysis

- Establish Practical Significance as well as Statistical Significance.
- Sample Size Affects All Results.
- Know Your Data.
- Strive for Model Parsimony (Don't add variables indiscriminately).
- Look at Your Errors.
- Validate Your Results.
  - Split the sample
  - Gather extra data to check that the model also matches the new data
  - Resimpling methods

How to Assess the Validity of the Solution?

- Gather extra data to check that the model also matches the new data
- Split the sample – Holdout sample (k-fold cross validation):
  - For large samples, use 3-fold cross validation: use 2/3 of the data for the analysis and 1/3 as the holdout sample (to test our model)
  - For small samples, use 10-fold cross validation: use 9/10 of the data for the analysis and 1/10 as the holdout sample (to test our model)

Holdout Sample: Split Sample Validation

- To test the generalizablity of findings from a principal component analysis, we could conduct a second research study to see if our findings are verified.
- A less costly alternative is to split the sample randomly into two halves, do the principal component analysis on each half and compare the results.
- If the communalities and the factor loadings are the same on the analysis on each half and the full data set, we have evidence that the findings are generalizable and valid because, in effect, the two analyses represent a study and a replication.

How to Assess the Validity of the Solution?

- Resampling Techniques:

1. Jackknife Validation.
2. Bootstrap Validation.

Introduction on Resampling Techniques

Problems: A statistic has been estimated from a sample, we want to

- know how confident we can be in the estimator and what its standard error and bias are, and
- gauge the estimator against a null distribution we want to discount

Introduction on Resampling Techniques

What, why, and how.

- Rather than using classical and/or analytical statistics we use brute force (Monte Carlo) computations to generate huge numbers of synthetic or fake samples. These samples form the basis for constructing sampling distributions of either the estimator itself or its null distribution to address respectively the two problems.

Introduction on Resampling Techniques

- What, *why*, and how.
  - It is not clear assumptions for usual approaches are satisfied.
  - Sample sizes are too small for satisfactory application of usual approaches.
  - It is not easy or possible to derive analytical descriptions of distributions for the estimator.
  - The inference problem is complicated.

Introduction on Resampling Techniques

- What, why, and *how*.
  - Resampling/rerandomization: Using the available sample to generate additional samples.
  - Statistical modeling: Fitting a model to the available sample and using the model to generate additional samples.

Versions of Resampling

Resampling	Procedure	Applications
Permutation	**Samples are drawn at random from original pool without replacement**	**Tests of hypotheses**
Bootstrap[1]	**Samples are drawn with replacement**	**Tests of hypotheses** **AND** **Standard error, bias, and confidence intervals of estimator**
Jack Knife[2]	**Samples consist of original pool with one at a time withheld**	**Standard error, bias, and confidence intervals of estimator**

[1] Most versatile.

[2] Generally outperformed by others.

Jacknife

Jacknife is used for bias removal. As we know that mean-square error is sum of squared bias and variances of the estimator. If bias is much higher than variance then under some circumstances Jacknife could be used.

Description of Jacknife: Let us assume that we have sample of size n. We estimate some sample statistics using all data – t_n. Then by removing one point at a time we estimate $t_{n-1,i}$, where subscript indicates size of the sample and index of removed sample point. Then new estimator is derived as:

$$t'_n = nt_n - (n-1)\bar{t}_{n-1}, \text{ where } \bar{t}_{n-1} = \frac{\sum_{i=1}^{n} t_{n-1,i}}{n}$$

Jackknife cross-validation

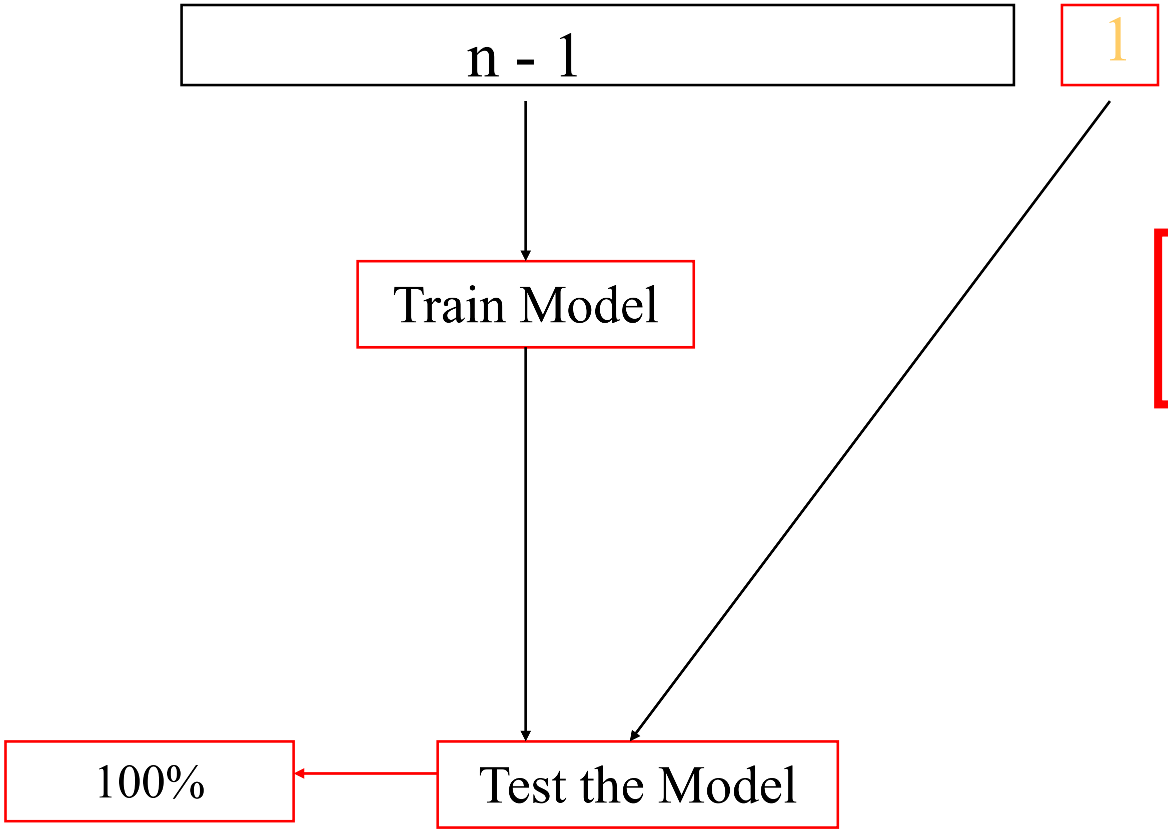
- Calculate the statistic by eliminating one sample, and then proceed as usual

- Repeat *n* times, where *n* is the number of available samples

- Bias estimate: $bias_{jack} = (n-1)(\hat{\theta}_{(.)} - \hat{\theta})$

- Variance estimate: $Var_{jack}[\hat{\theta}] = \dfrac{n-1}{n}\sum_{i=1}^{n}[\hat{\theta}_{(i)} - \hat{\theta}_{(.)}]^2$

Cross-validation setup



n - 1

1

Train Model

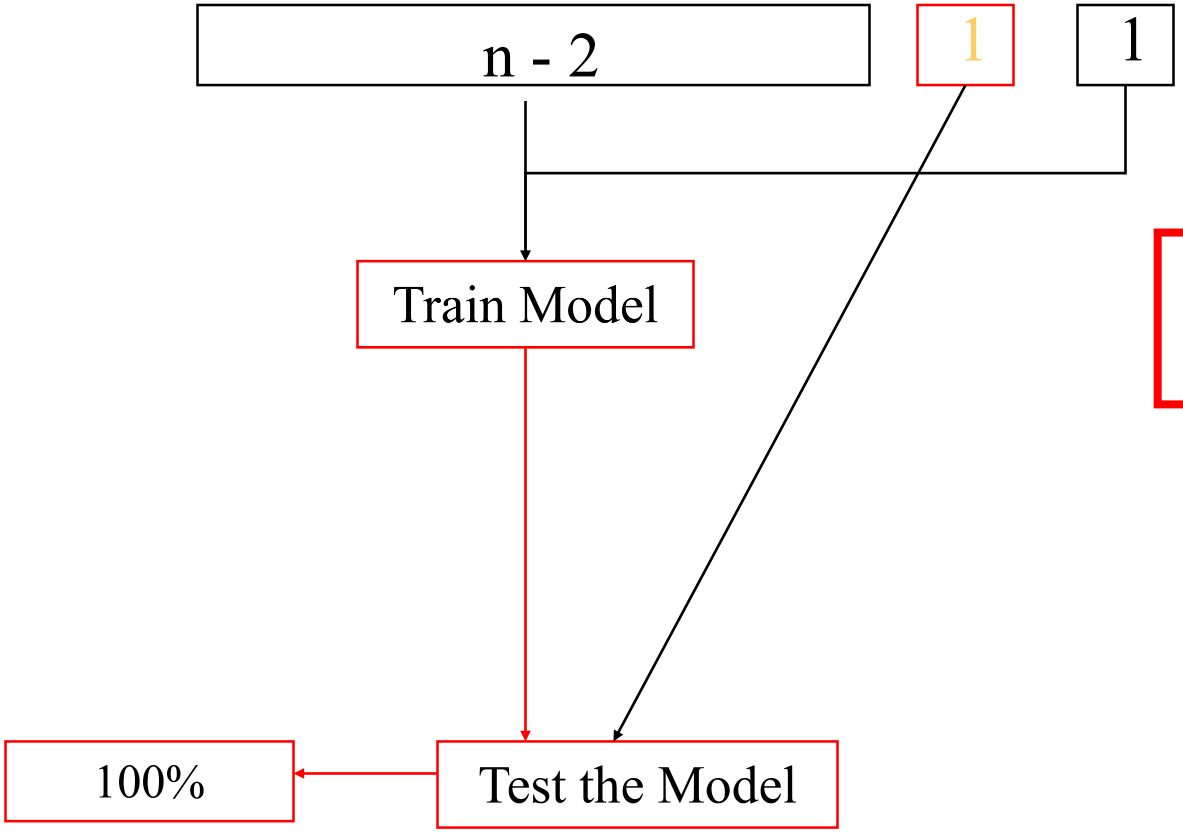
Test the Model

100%

Jackknife scores

1st validation set: 100%

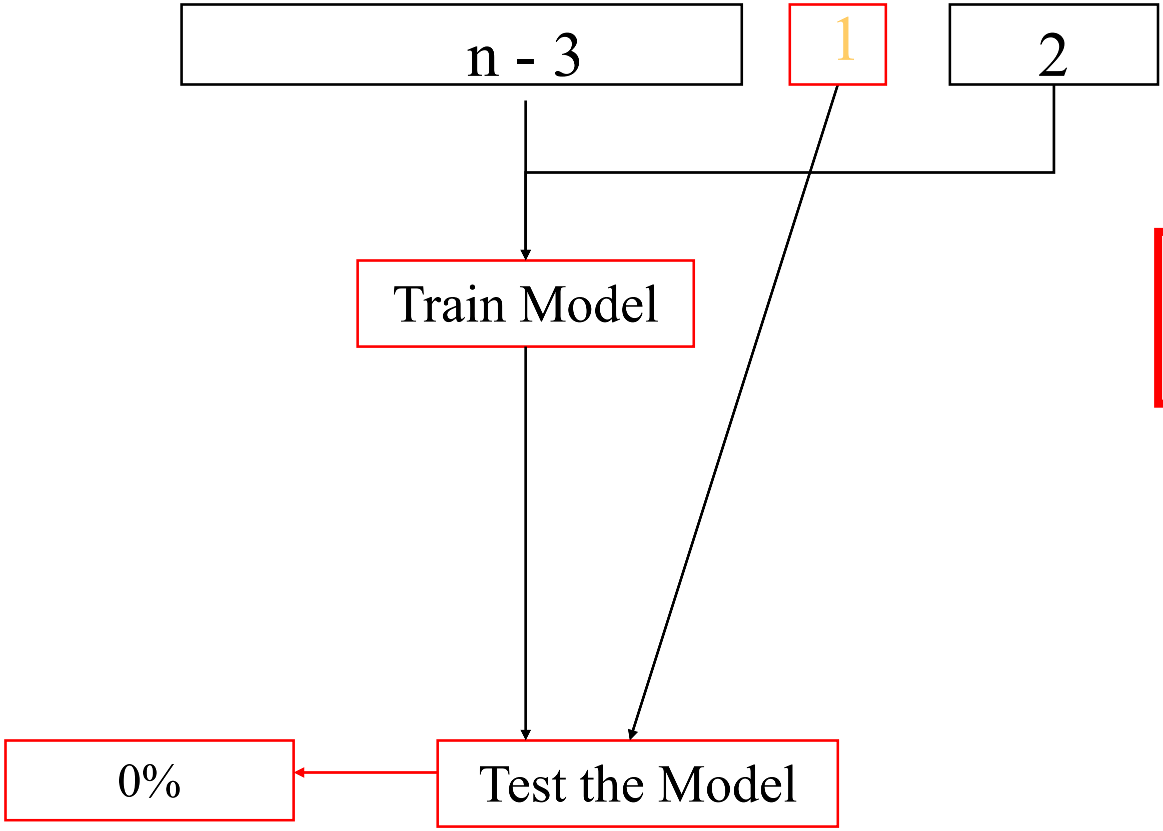
Cross-validation setup



Jackknife scores

1st validation set: 100%

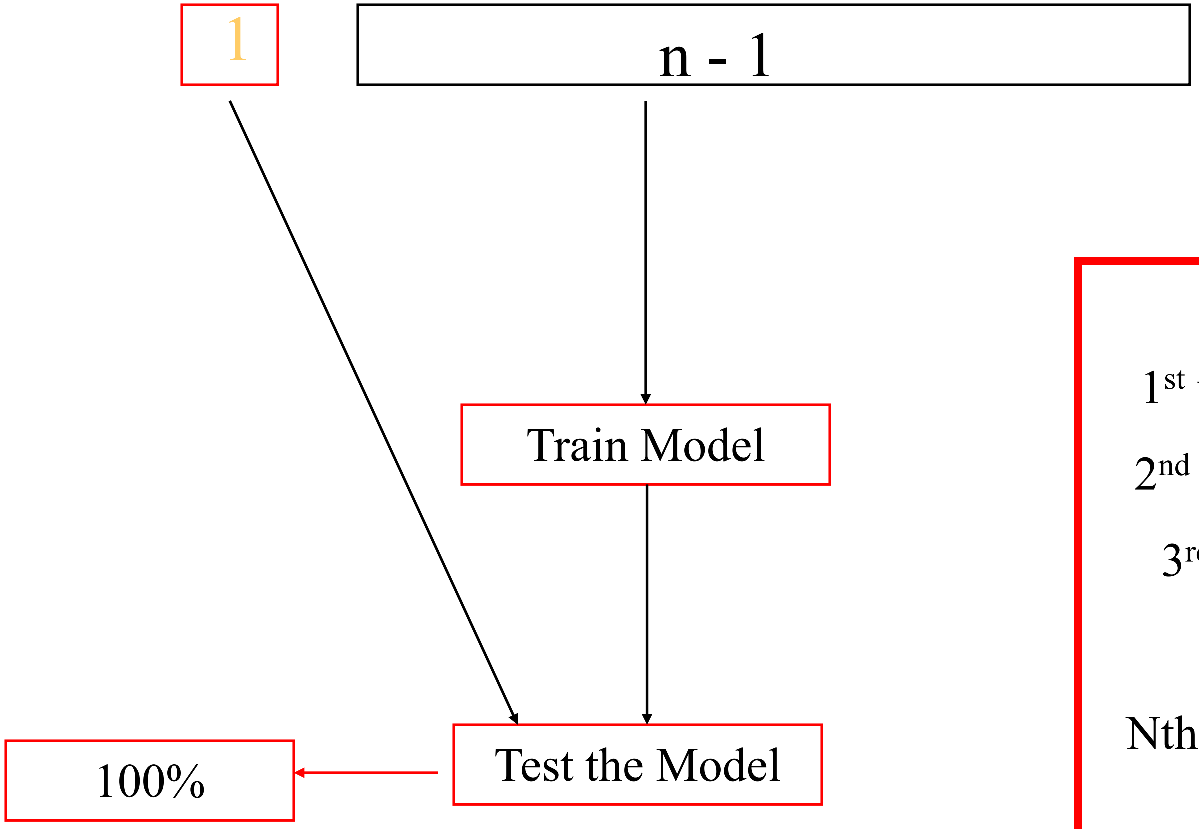
Cross-validation setup



Jackknife scores

1st validation set: 100%

Cross-validation setup



Jackknife scores

1st validation set: 100%

2nd validation set: 100%

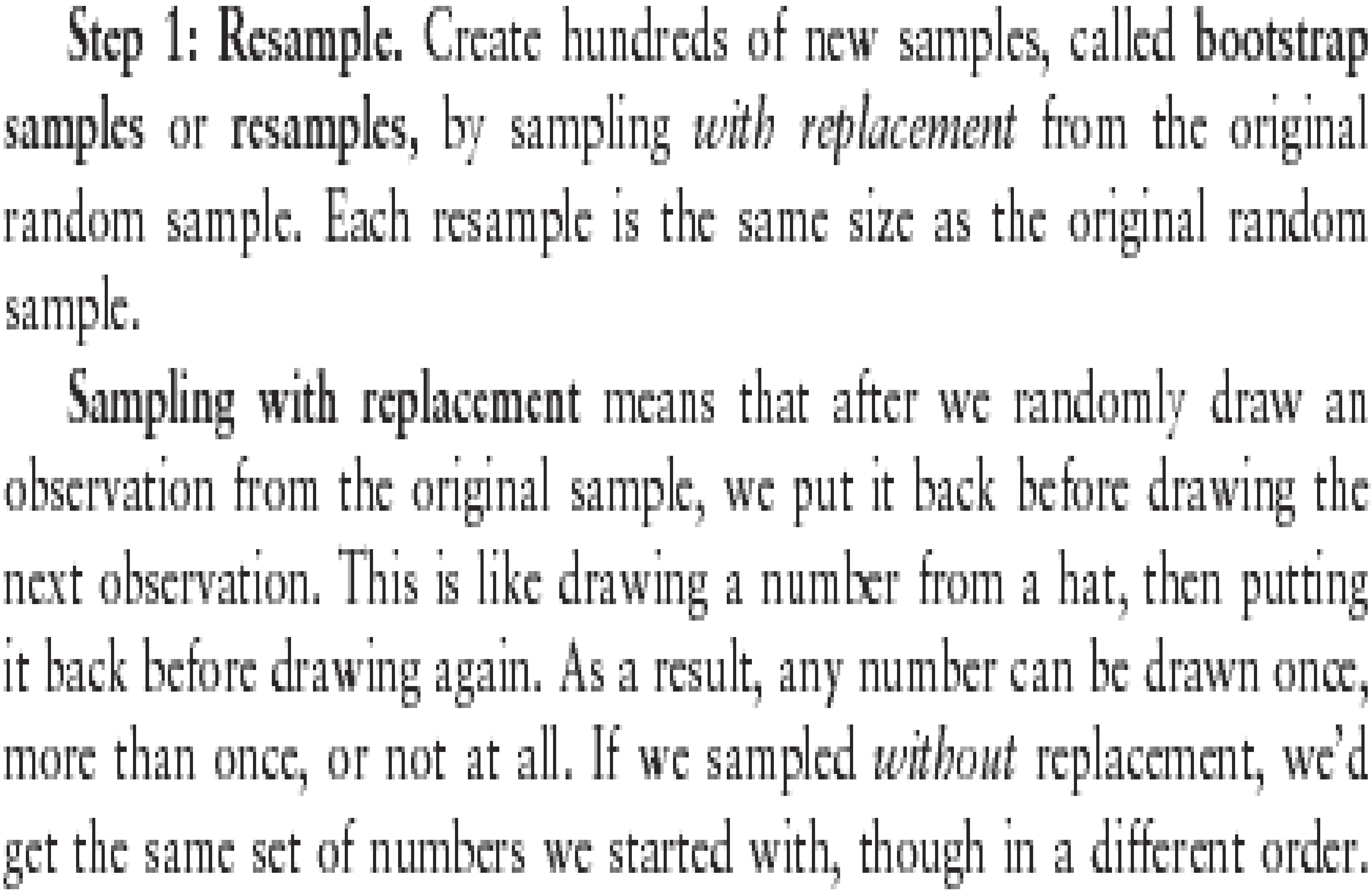
3rd validation set: 0%

…

Nth validation set: 100%

Average: 91%

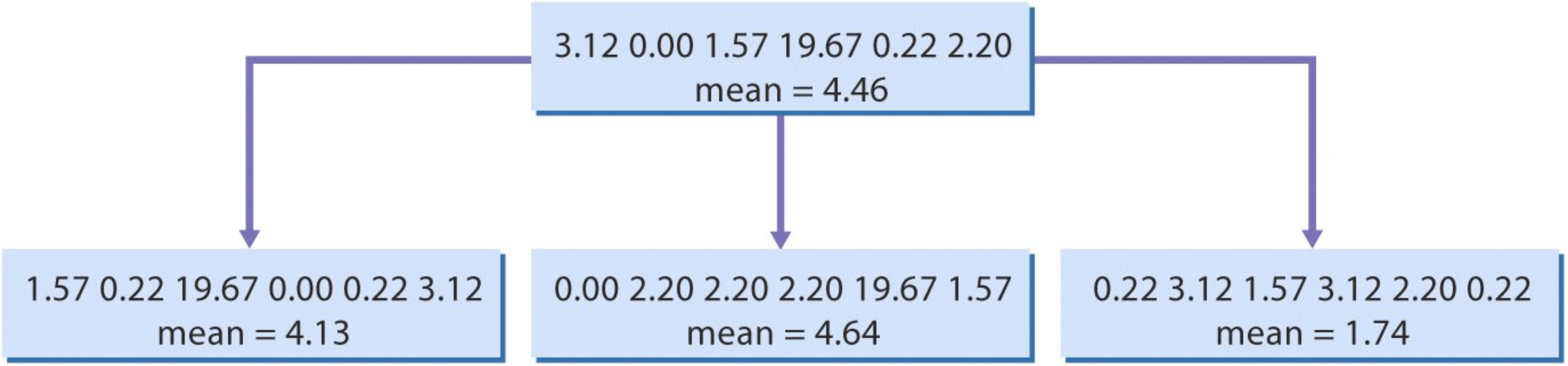
Procedure for Bootstrapping



Step 1: Resample. Create hundreds of new samples, called **bootstrap samples** or **resamples,** by sampling *with replacement* from the original random sample. Each resample is the same size as the original random sample.

Sampling with replacement means that after we randomly draw an observation from the original sample, we put it back before drawing the next observation. This is like drawing a number from a hat, then putting it back before drawing again. As a result, any number can be drawn once, more than once, or not at all. If we sampled *without* replacement, we'd get the same set of numbers we started with, though in a different order.

A simple example for Bootstrapping



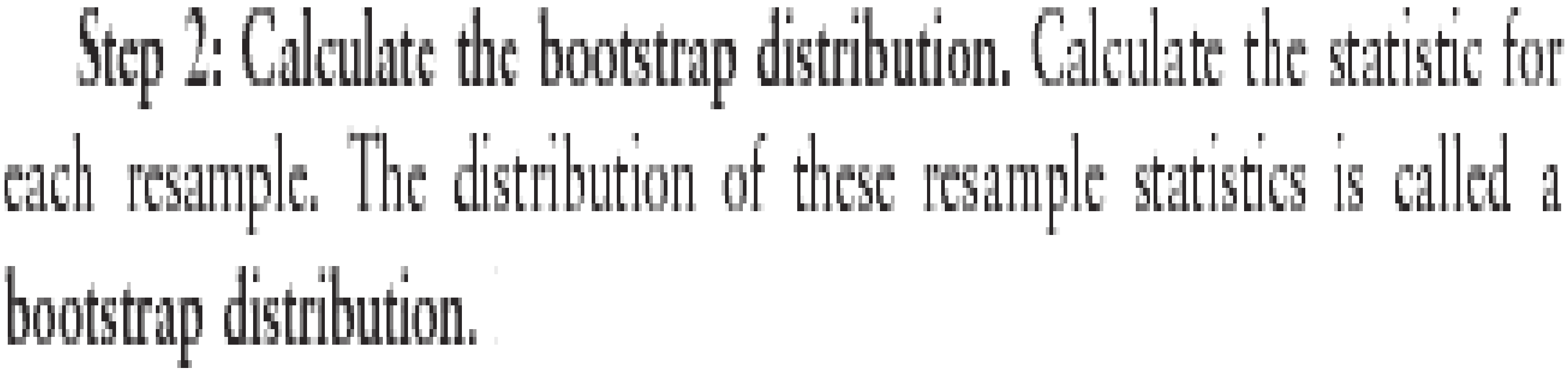
3.12 0.00 1.57 19.67 0.22 2.20
mean = 4.46

1.57 0.22 19.67 0.00 0.22 3.12
mean = 4.13

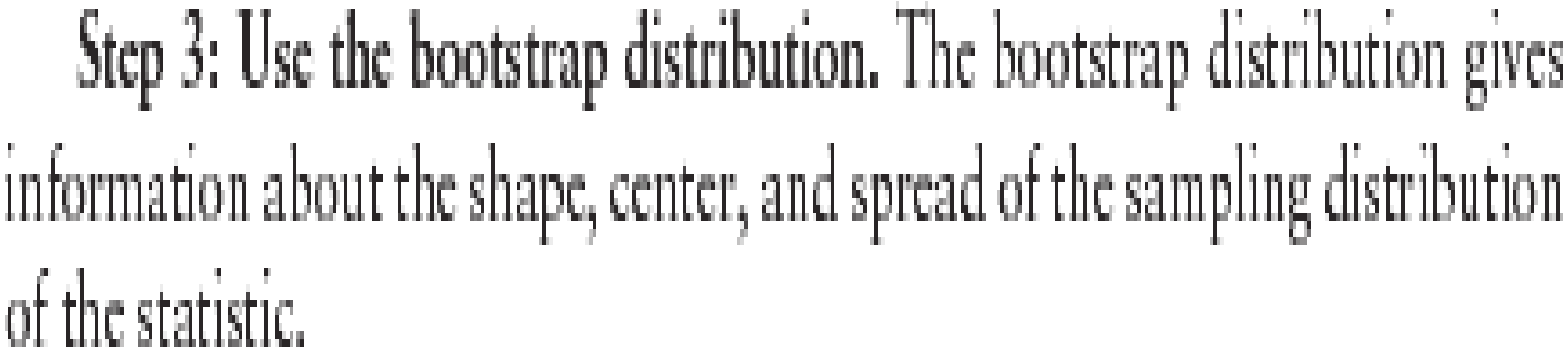
0.00 2.20 2.20 2.20 19.67 1.57
mean = 4.64

0.22 3.12 1.57 3.12 2.20 0.22
mean = 1.74

Procedure for Bootstrapping



Step 2: Calculate the bootstrap distribution. Calculate the statistic for each resample. The distribution of these resample statistics is called a bootstrap distribution.

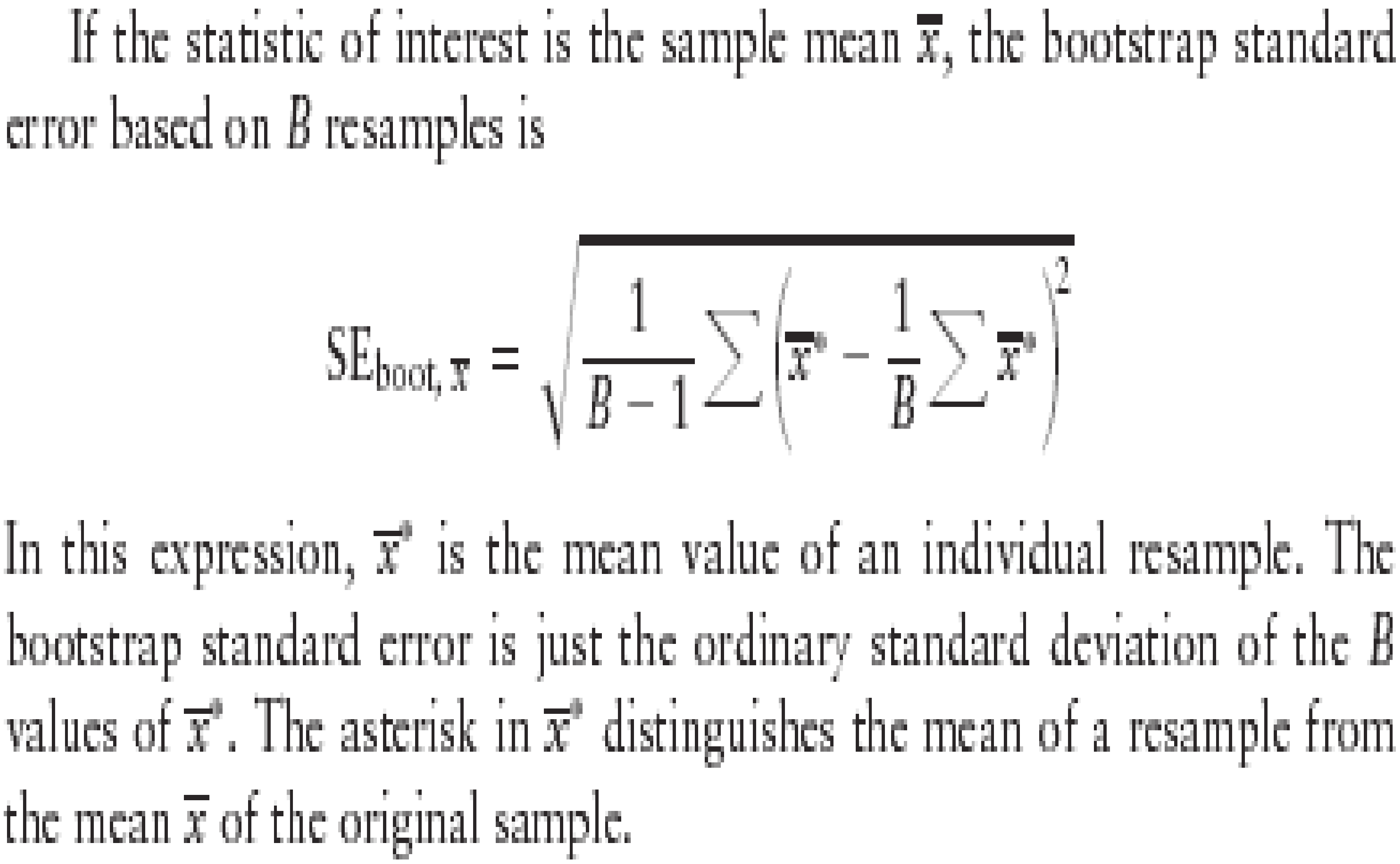


Step 3: Use the bootstrap distribution. The bootstrap distribution gives information about the shape, center, and spread of the sampling distribution of the statistic.

THE BOOTSTRAP IDEA

The original sample represents the population from which it was drawn. So resamples from this sample represent what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, represents the sampling distribution of the statistic, based on many samples.

Bootstrap Standard Error



If the statistic of interest is the sample mean $\bar{x}$, the bootstrap standard error based on B resamples is

$$SE_{\text{boot},\bar{x}} = \sqrt{\frac{1}{B-1}\sum\left(\bar{x}^* - \frac{1}{B}\sum\bar{x}^*\right)^2}$$

In this expression, $\bar{x}^*$ is the mean value of an individual resample. The bootstrap standard error is just the ordinary standard deviation of the B values of $\bar{x}^*$. The asterisk in $\bar{x}^*$ distinguishes the mean of a resample from the mean $\bar{x}$ of the original sample.

Misleading Results to Watch Out For

- When we examine the communalities and factor loadings, we are matching up overall patterns, not exact results: the communalities should all be greater than 0.50 and the pattern of the factor loadings should be the same.
- Sometimes the variables will switch their components (variables loading on the first component now load on the second and vice versa), but this does not invalidate our findings.
- Sometimes, all of the signs of the factor loadings will reverse themselves (the plus's become minus's and the minus's become plus's), but this does not invalidate our findings because we interpret the size, not the sign of the loadings.

When validation fails

- If the validation fails, we are warned that the solution found in the analysis of the full data set is not generalizable and should not be reported as valid findings.

- We do have some options when validation fails:
  - If the problem is limited to one or two variables, we can remove those variables and redo the analysis.
  - Randomly selected samples are not always representative. We might try some different random number seeds and see if our negative finding was a fluke. If we choose this option, we should do a large number of validations to establish a clear pattern, at least 5 to 10. Getting one or two validations to negate the failed validation and support our findings is not sufficient.

Outliers

- SPSS calculates factor scores as standard scores.
- SPSS suggests that one way to identify outliers is to compute the factors scores and identify those have a value greater than ±3.0 as outliers.
- If we find outliers in our analysis, we redo the analysis, omitting the cases that were outliers.
- If there is no change in communality or factor structure in the solution, it implies that there outliers do not have an impact. If our factor solution changes, we will have to study the outlier cases to determine whether or not we should exclude them.
- After testing outliers, restore full data set before any further calculations

Description of HBAT Primary Database Variables

Variable Description		Variable Type
Data Warehouse Classification Variables		
X1	Customer Type	nonmetric
X2	Industry Type	nonmetric
X3	Firm Size	nonmetric
X4	Region	nonmetric
X5	Distribution System	nonmetric
Performance Perceptions Variables		
X6	Product Quality	metric
X7	E-Commerce Activities/Website	metric
X8	Technical Support	metric
X9	Complaint Resolution	metric
X10	Advertising	metric
X11	Product Line	metric
X12	Salesforce Image	metric
X13	Competitive Pricing	metric
X14	Warranty & Claims	metric
X15	New Products	metric
X16	Ordering & Billing	metric
X17	Price Flexibility	metric
X18	Delivery Speed	metric
Outcome/Relationship Measures		
X19	Satisfaction	metric
X20	Likelihood of Recommendation	metric
X21	Likelihood of Future Purchase	metric
X22	Current Purchase/Usage Level	metric
X23	Consider Strategic Alliance/Partnership in Future	nonmetric

Multivariate Analysis Learning Checkpoint

1. What is multivariate analysis?
2. Why use multivariate analysis?
3. Why is knowledge of measurement scales important in using multivariate analysis?
4. What basic issues need to be examined when using multivariate analysis?
5. Describe the process for applying multivariate analysis.

The end, thank you.