

Clusters, Trees, and Types: Statistical Methods for Assessing Complex Data Patterns and Relationships

Trace Kershaw, PhD
Assistant Professor
Social Behavioral Sciences Program
Epidemiology and Public Health
Yale University

Limitations of Traditional Statistical Techniques

- Difficult to identify and assess complex data patterns
- Focus on linear associations
- Difficult to model higher order interactions

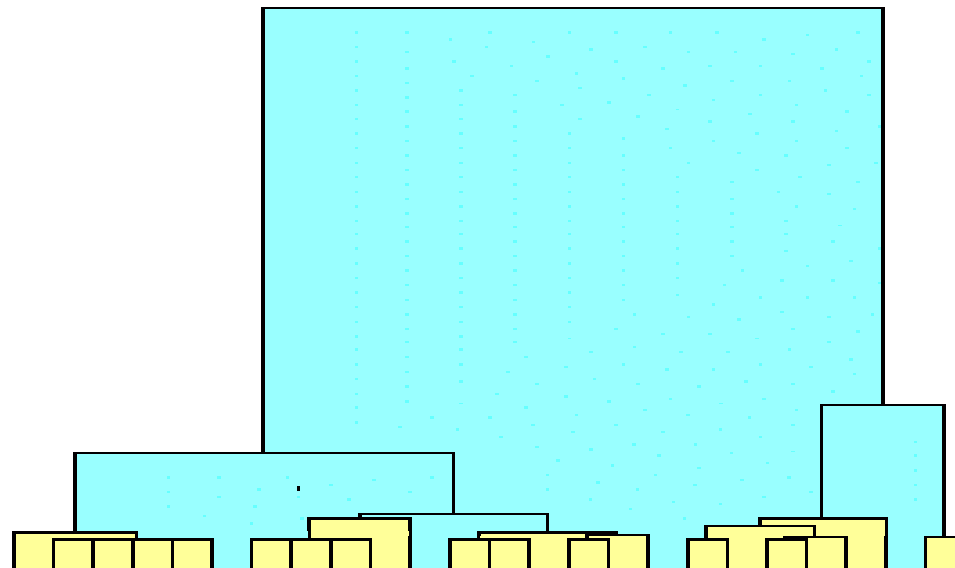
Statistical Methods to Identify Complex Data Patterns

- Cluster analysis
- Recursive partitioning (regression trees)
- Configural frequency analysis

Presentation Objectives

- For each of the analyses:
 - Describe
 - Compare to traditional approaches
 - Give examples of use in HIV prevention research
 - Outline strengths and weaknesses

Cluster Analysis



What is Cluster Analysis?

- A multivariate statistical technique that identifies homogeneous patterns of objects
- A data reduction technique
 - reduces a group of N subjects into a smaller number of g subgroups using a set of p variables

When Would I Use Cluster Analysis?

- Classify similar subjects on a set of variables
 - Traditional: Factor Analysis
- Find complex associations among predictor and outcome domains
 - Traditional: Regression; SEM
- Identify patterns of change
 - Traditional: Repeated Measures GLM
- Identify change in patterns across time
 - Traditional: Doubly Multivariate

Cluster Analysis: Characteristics

- Predictors, outcomes, or both
- Binary or continuous variables
- Needs similar scaling units
- Reduces to a single categorical variable
- Needs relatively large samples
- Exploratory and confirmatory

Types of Cluster Analyses

- Hierarchical Agglomerative Cluster Analysis
- K-means Cluster Analysis

Hierarchical Agglomerative

- Subjects start out alone
- Two closest subjects merge to form a cluster
- Next closest subject-subject or subject-cluster pair merge
- Continues until all subjects belong to the same cluster
- Decide on # clusters

How do we Conduct a Cluster Analysis?

- Convert data into an $n \times n$ matrix of distances
 - Distance measures: how far apart two subjects are on a set of variables
 - Squared euclidean distance

How do we Conduct a Cluster Analysis?

- Choose a clustering algorithm:
 - Nearest neighbor
 - Furthest neighbor
 - Average linkage
 - Ward's method

Determining the Number of Clusters

- Heuristic approaches
 - Dendrogram plot
 - Scree plot
 - Mojena's stopping rule
- Systematic approaches
 - Cubic clustering criterion
 - Split sample replication

K-Means Cluster Analysis

- Pre-specified number of clusters
- Pre-specified location of clusters
- Subjects placed in closest cluster
 - Centroid: variable means for members of that cluster

Cluster Analysis Strategy

(Morey, Blashfield, & Skinner, 1983)

- Derivation Phase
- Replication Phase
- External Validation Phase
- Cross-Validation Phase

Example 1: Student Profiles

- Research Question:
 - To identify student profiles in order to target possible interventions to prevent:
 - Pregnancy
 - Delinquency
- Sample: 1005 9th grade girls from 21 schools

Example: Student Profiles

- Variables:
 - GPA
 - Locus of control
 - External stress
 - Friends abuse of substances
 - # of extracurricular activities
 - Proportion of friends that are boys

Example: Student Profiles

- Derivation Sample
 - Hierarchical CA
 - Results suggested 7 clusters
- Replication Sample
 - K-means CA using cluster centers from derivation sample
 - Hierarchical CA

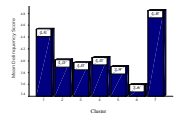
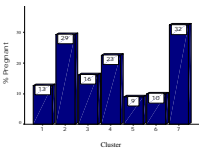
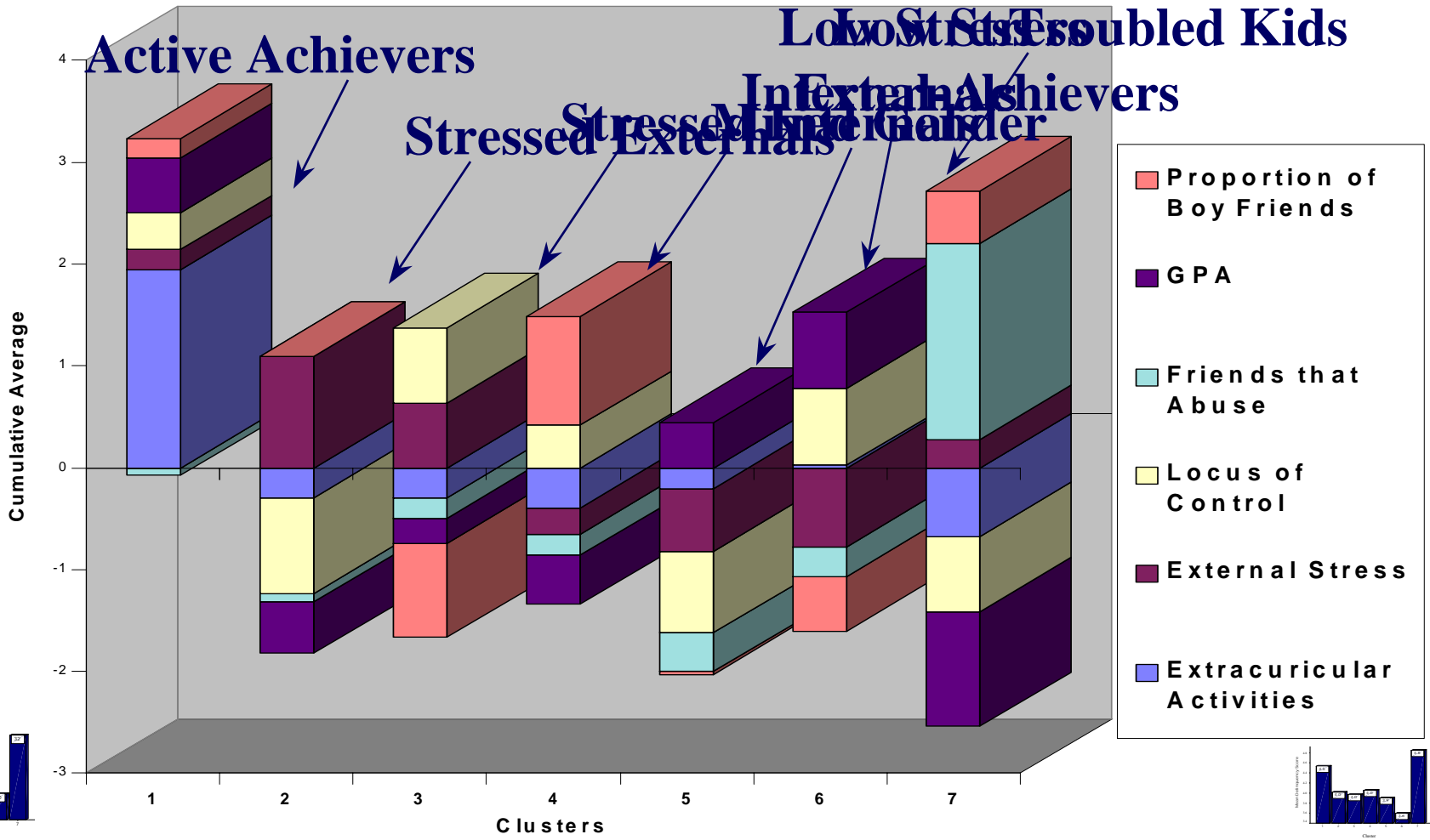
Replication Phase

- Conducted a k-means cluster analysis using the cluster centers from the derivation sample to classify subjects in the replication sample
- Conducted a hierarchical agglomerative cluster analysis on the replication sample

Replication Phase

- Good agreement between the two cluster solutions
 - $\kappa = .61, p < .001$
 - Overlap ranged from 39% to 82%
 - Only 1 cluster had less than 50% overlap

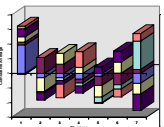
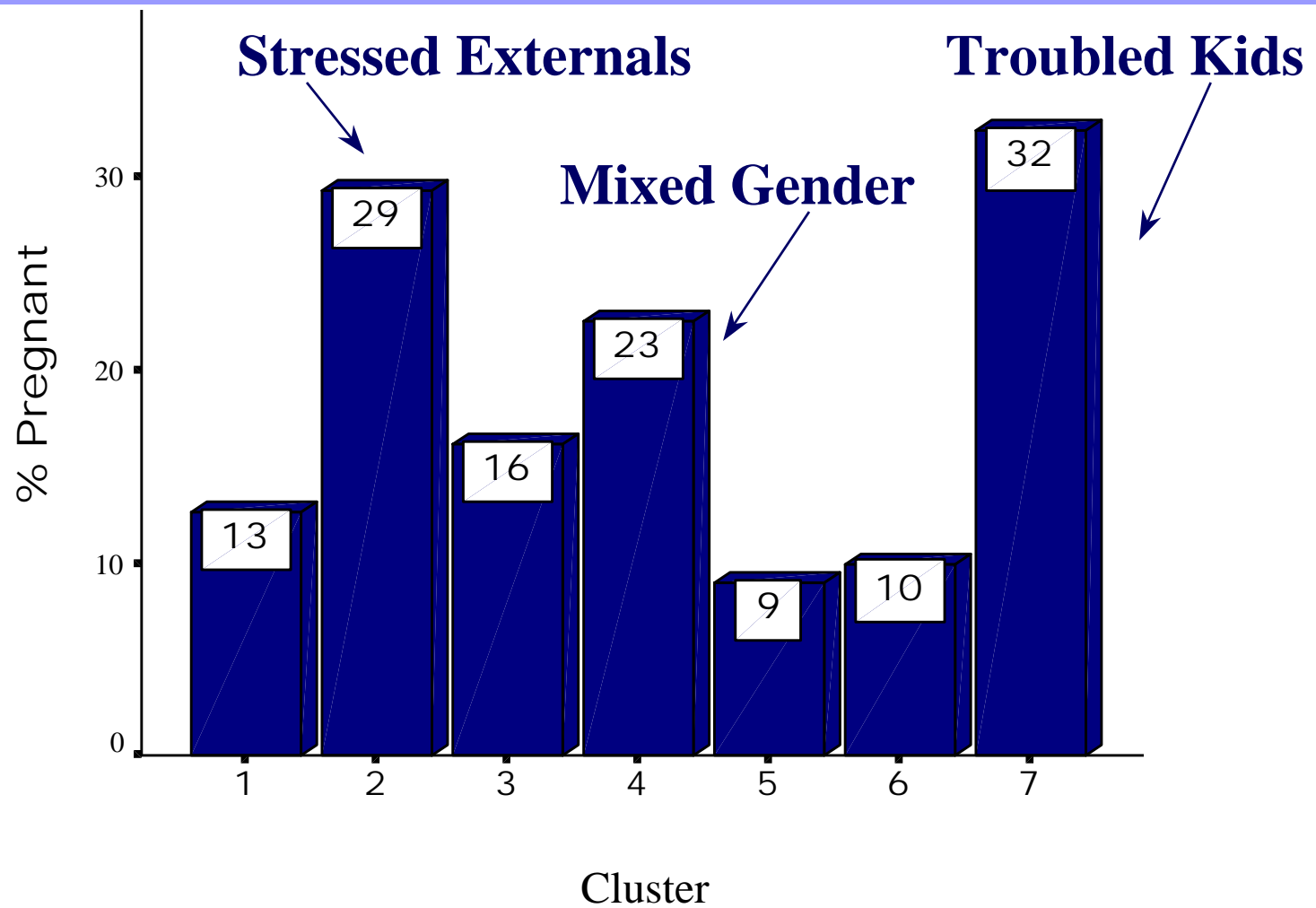
Results



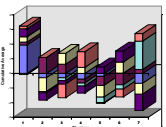
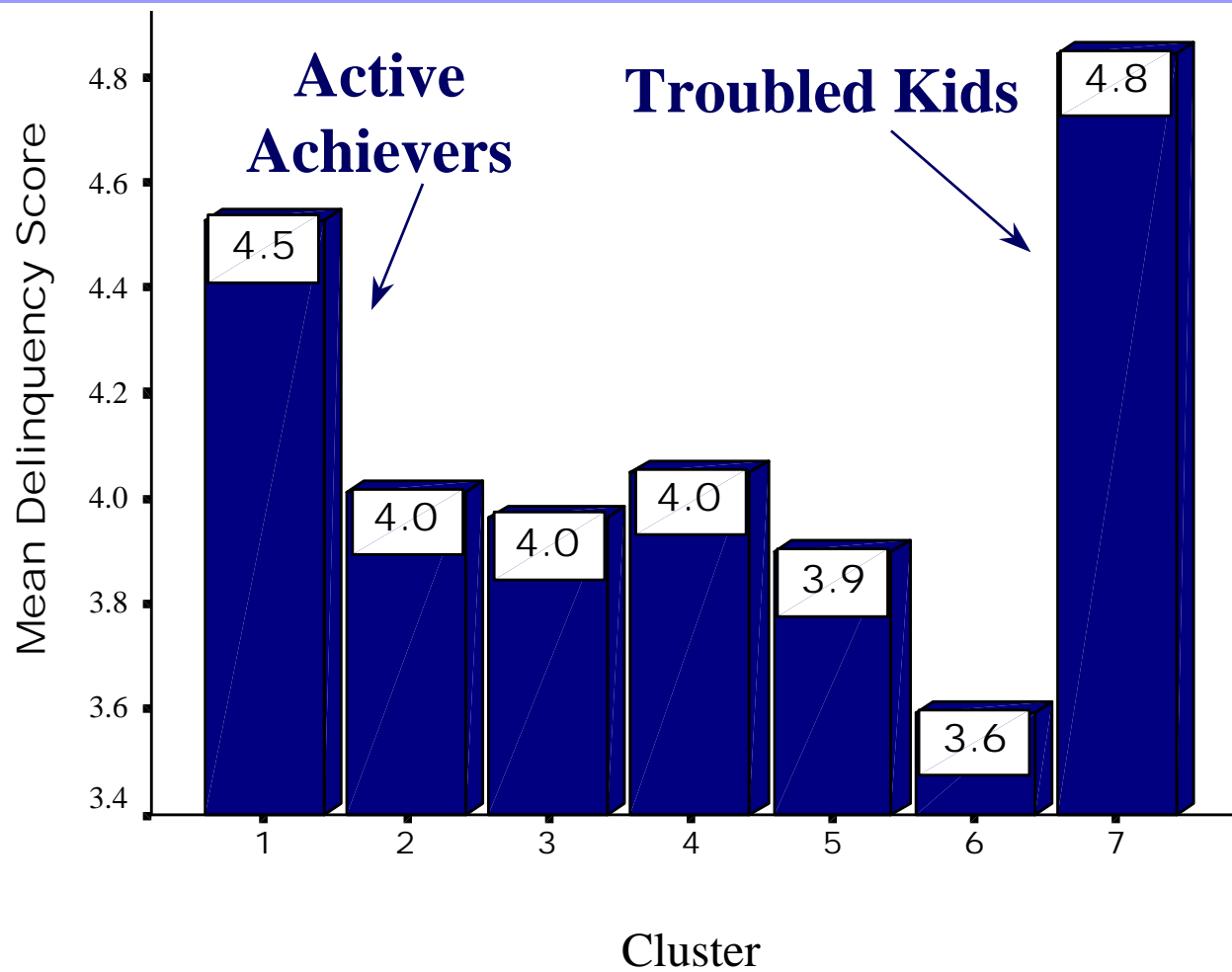
External Validation Phase

- Compare the clusters on meaningful outcomes:
 - Pregnancy
 - Chi-Square=50.49, $p < .001$ **
 - Delinquency
 - $F = 3.00$, $p < .01$ **

Percent Pregnant by Cluster



Delinquency by Cluster



Did Cluster Analysis Work?

- Identified distinct groups of adolescent girls that differed on important outcome variables
 - Pregnancy
 - Delinquency
- Cluster analysis was more predictive than regression and logistic regression
- Cluster analysis had better cross-validation than regression and logistic regression

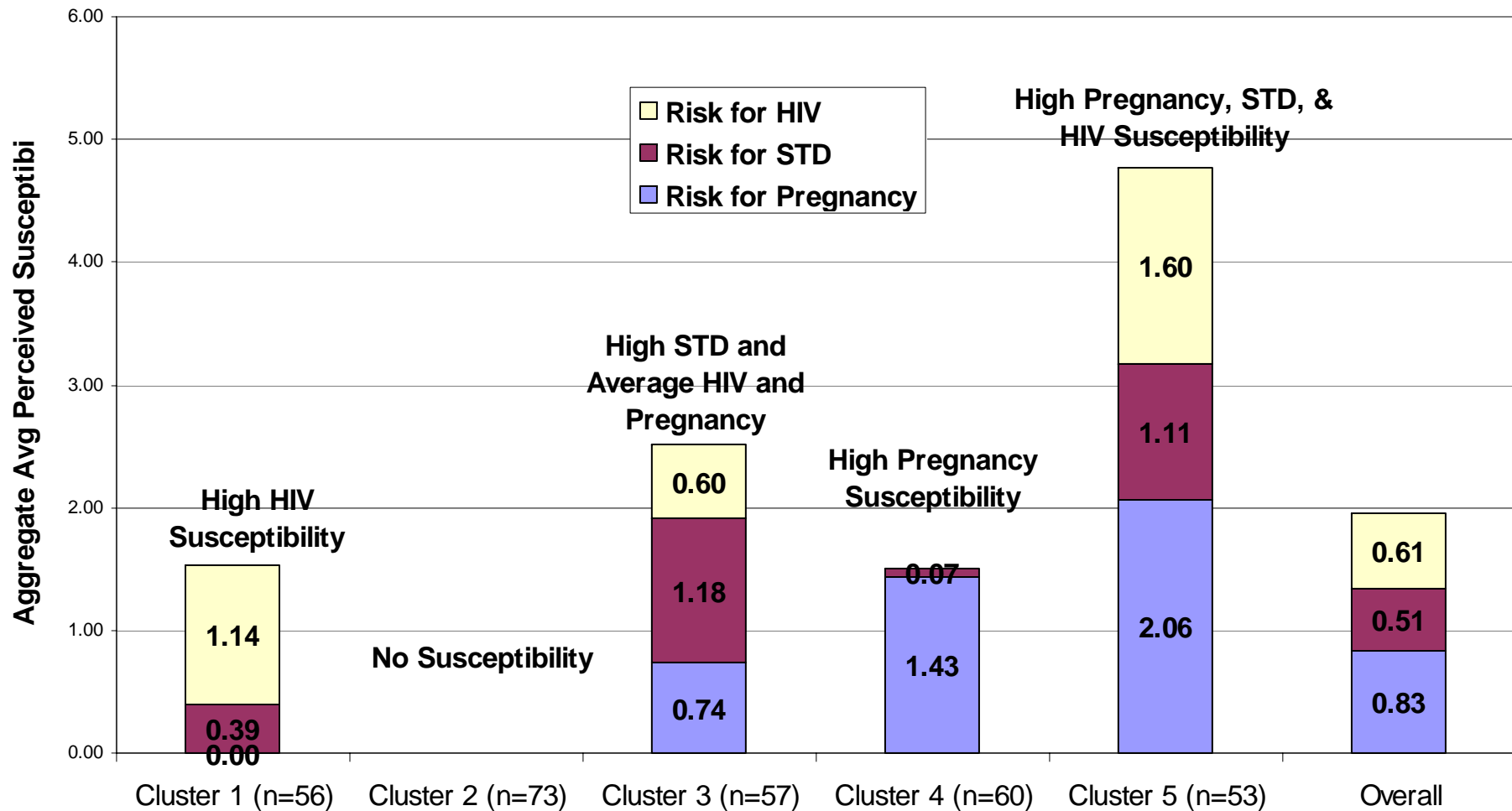
Example 2: Perceived Susceptibility

- Research Question:
- To identify subgroups of adolescent females based on their perceptions of susceptibility to pregnancy, STDs, and HIV
- Compare these subgroups on sexual risk behaviors
- Sample: 300 low-income adolescent females

Example: Perceived Susceptibility

- Hierarchical CA
 - Both scree plot and split-sample replication suggested 5 clusters

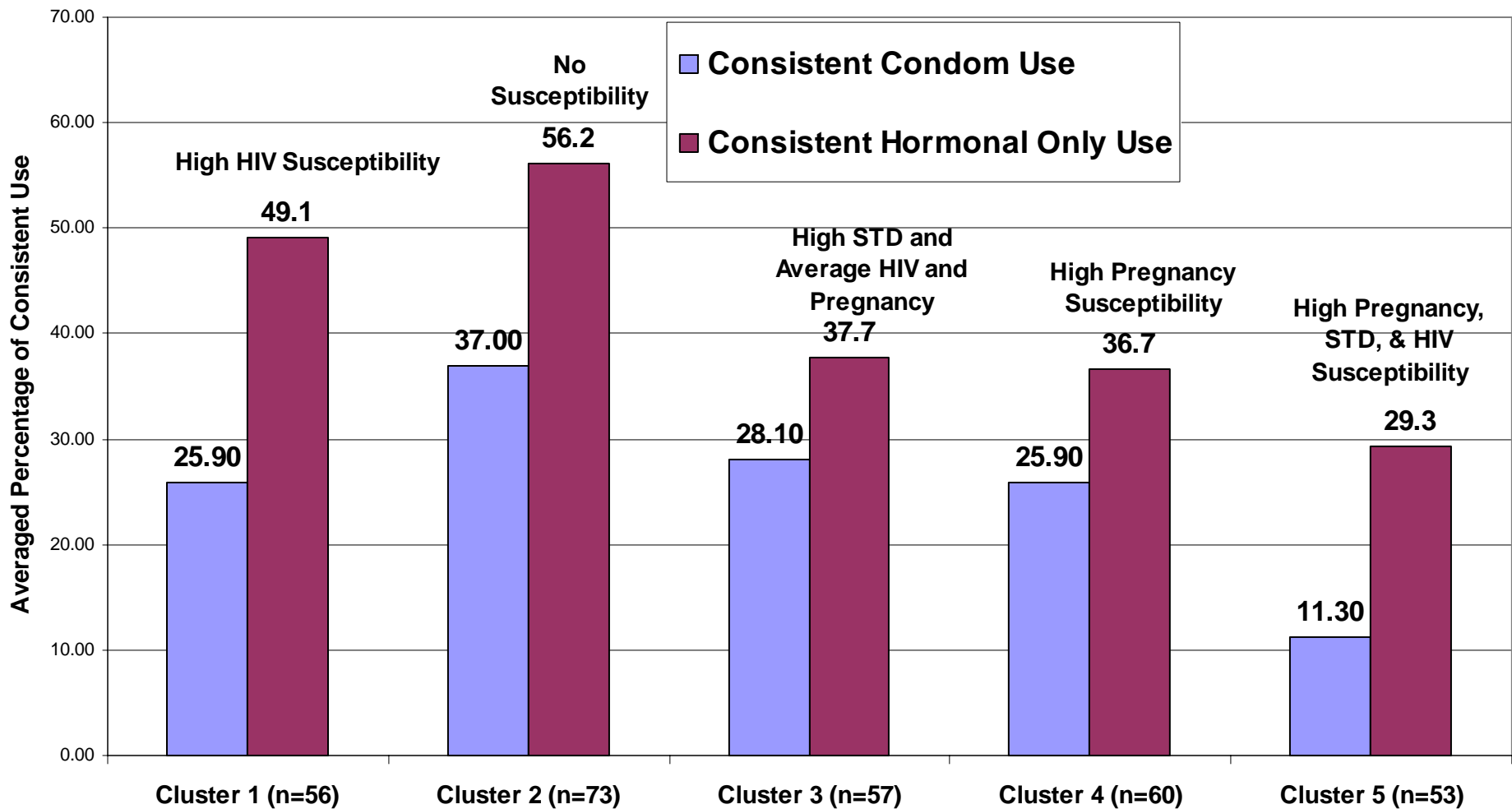
Perceived Susceptibility Clusters



Example: Perceived Susceptibility

- Compared clusters on sexual risk behaviors
 - Significant differences between clusters on:
 - consistent condom use
 - consistent hormonal contraception use
 - multiple partners
 - sexual frequency

Perceived Susceptibility Clusters



Did Cluster Analysis Work?

- Identified distinct groups of adolescent girls that differed on important outcome variables
- Cluster analysis was more predictive than regression and logistic regression

Cluster Analysis: Strengths and Weaknesses

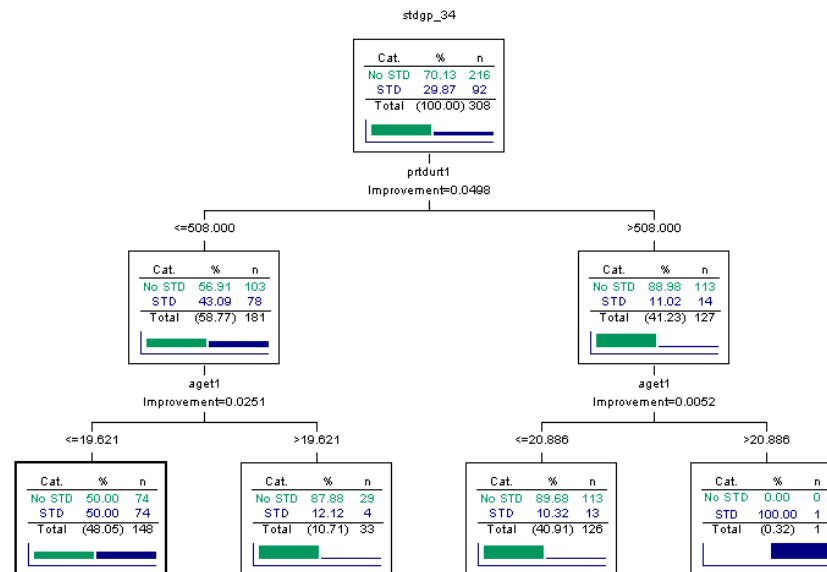
Strengths

- Multiple data levels
- Identify small sized but important subgroups
- Model complex variable patterns
- Exploratory and confirmatory
- Cross-sectional and longitudinal
- Available on popular software

Weaknesses

- Data levels need to be uniform
- Can not explicitly differentiate IVs and DVs
- Confirmatory methods for binary data more difficult
- Can not handle multinomial data
- Does not handle missing data

Regression Tree Analysis



What is Regression Tree Analysis ?

- A non-parametric technique that recursively partitions groups into smaller subgroups that maximally differ on a desired outcome
- Cross between stepwise regression and cluster analysis

When Would I Use Regression Trees?

- Predict occurrence of an outcome from a set of predictors (pathways to risk)
 - Traditional: Regression; Logistic Regression; Multinomial Regression
- Detect threshold effects
 - Traditional: Spline Curve Fitting
- Predicting censored data
 - Traditional: Proportion Hazard Regressions
- Identify patterns of change
 - Traditional: Repeated Measures GLM; Mixed Effects Modeling

Regression Trees: Characteristics

- Set of predictors and single outcome
- Categorical, ordinal, or continuous predictors
- Categorical, ordinal, or continuous outcome
- Allows mix of scaling units
- Need relatively large samples
- Exploratory and confirmatory

Types of Regression Trees

- CART- Classification and Regression Trees (Breiman et al., 1985)
- CHAID- Chi-Square Automatic Interaction Detection (Kass, 1980)
- QUEST- Quick, Unbiased, Efficient, Statistical Tree (Loh & Shih, 1997)

How do we Conduct a Regression Tree Analysis?

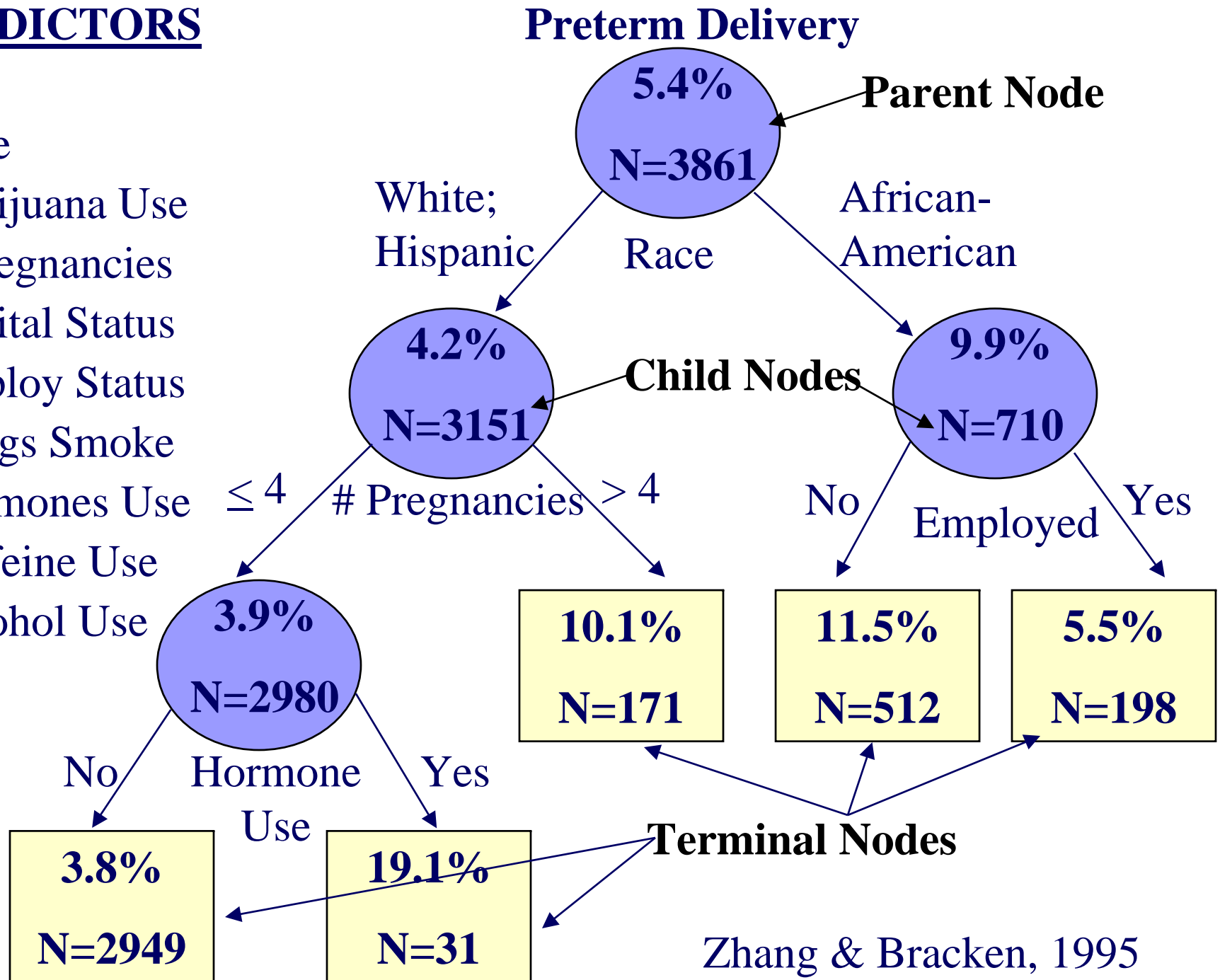
- Growing
- Stopping
- Pruning

Growing

- Start with all subjects in 1 group (parent node)
- Divide parent node into two “child nodes” based on best predictor
- Best predictor=lowest impurity
- Based on all possible variable splits
- Repeat process for each child node

PREDICTORS

- Age
- Race
- Marijuana Use
- # Pregnancies
- Marital Status
- Employ Status
- # Cigs Smoke
- Hormones Use
- Caffeine Use
- Alcohol Use



Stopping

- Stopping rules
 - Differences between resulting nodes is not significant
 - Tree depth
 - n of node (1% of N; 5)

Pruning

- Eliminate spurious branches
- Examines all subtrees of full tree
- Cross-validation
 - Iterative replication
 - Split sample
- Compares error rates of subtrees

Cross-Validation

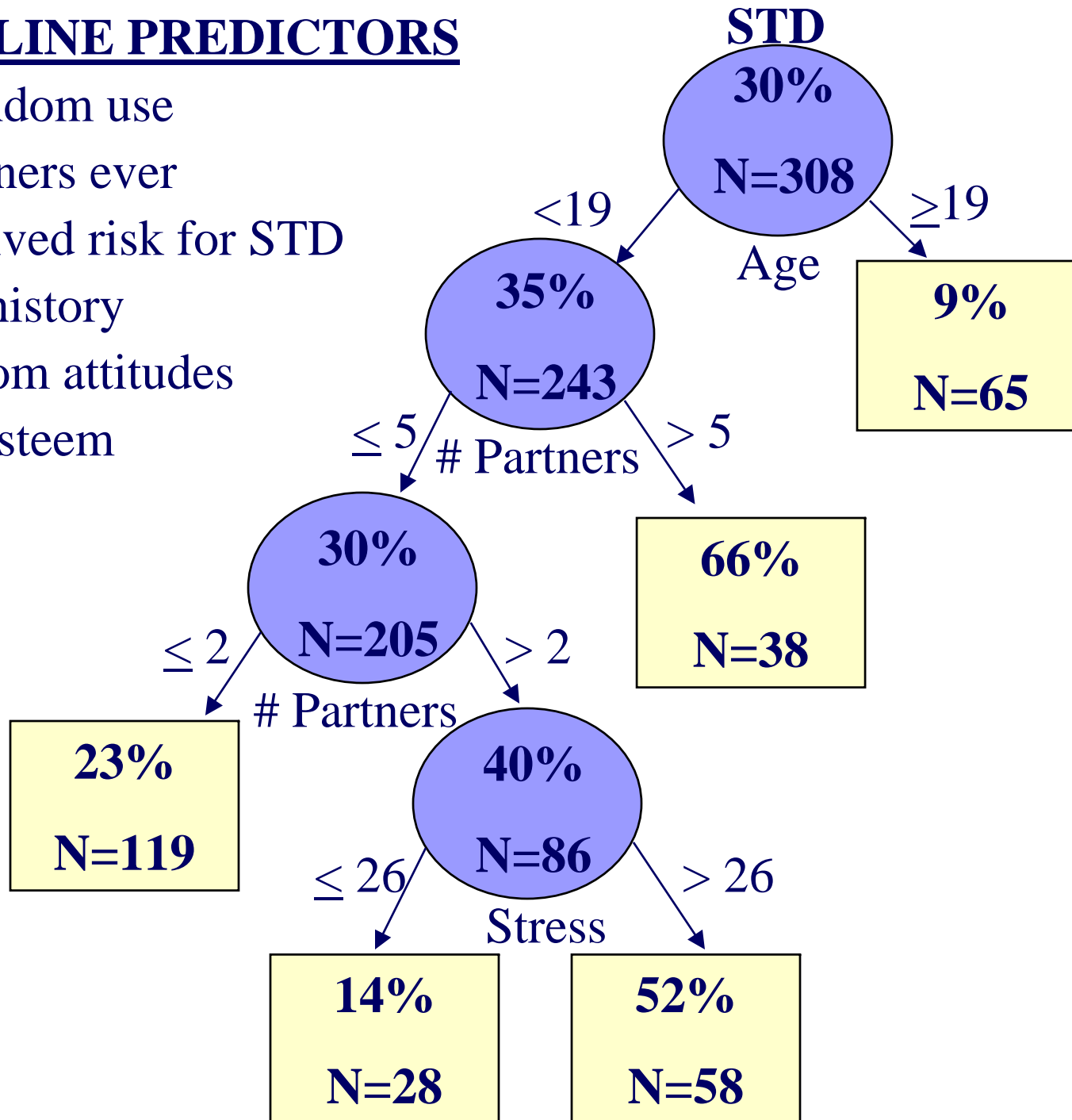
- Split sample cross-validation
- Confirmatory Regression Trees

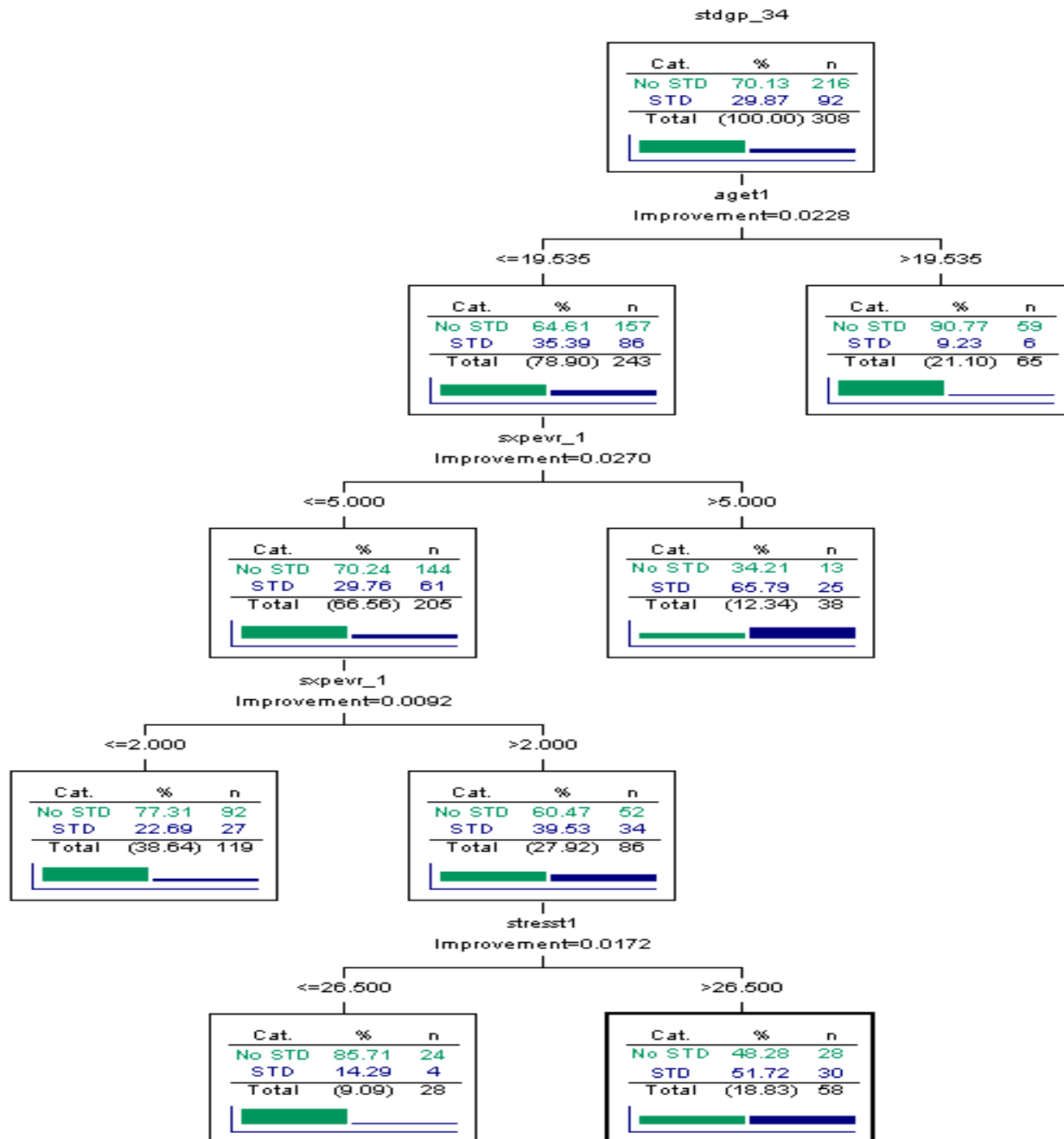
Example 1: Predicting STDs

- Predict subsequent STD acquisition from baseline risk factors for 308 adolescent females

BASELINE PREDICTORS

- % condom use
- # partners ever
- Perceived risk for STD
- STD history
- Condom attitudes
- Self-esteem
- Stress
- Age
- Race





Regression Trees: Strengths and Weaknesses

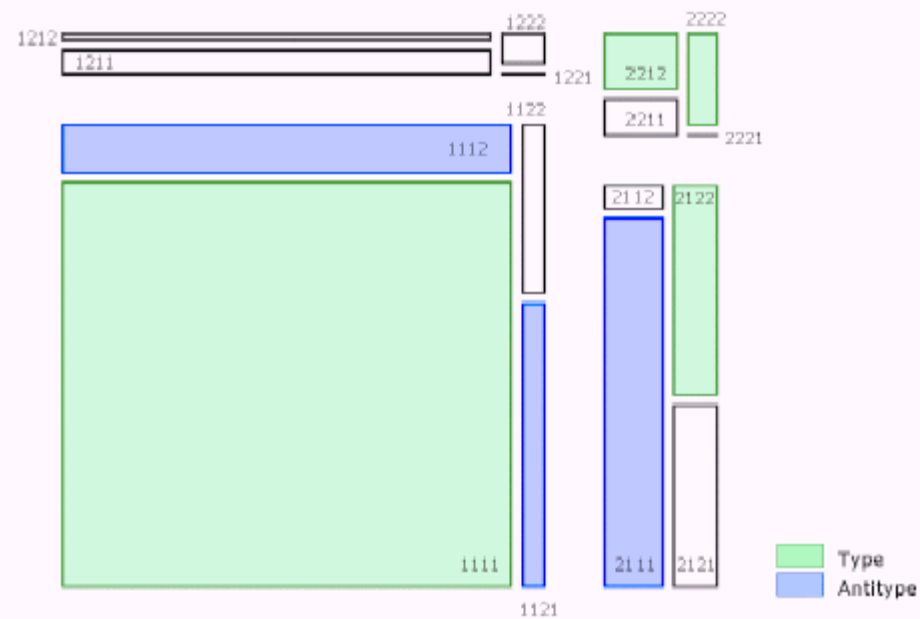
Strengths

- Explicitly model IV and DVs
- Forms groups that maximize difference on DV
- Easy to implement in applied settings
- Exploratory and confirmatory
- Cross-sectional and longitudinal
- Model missing data
- Not influenced by non-normal data

Weaknesses

- Poor fit if strong linear relationships
- Need to cross-validate
- Can have several good first splits
- Can have several “good fitting” trees
- Only allows 1 outcome

Configural Frequency Analysis



What is Configural Frequency Analysis ?

- A multivariate technique used to identify types and antitypes in cross-classification of categorical data
- Cross between cluster analysis and log-linear analysis

When Would I Use Configural Frequency Analysis?

- Classify similar subjects on a set of variables
 - Traditional: Factor Analysis
- Find complex associations among predictor and outcome domains
 - Traditional: Regression or SEM
- Identify patterns of change
 - Traditional: Repeated Measures GLM

Configural Frequency Analysis: Characteristics

- Set of predictors and outcomes
- Categorical variables only
- Needs relatively large samples
- Exploratory and confirmatory
- Requires special statistical programs

How Do We Conduct a Configural Frequency Analysis

- Specify chance model
- Estimate expected cell frequencies
- Test for types and antitypes
- Interpret results

Types of CFA

- One Sample CFA
- Two Sample CFA
- Second Order CFA

Example: Changes in Sexual Risk Categories by Accuracy of Risk Perceptions

- Looking at whether accurately perceiving sexual risk influences changes in categories of sexual risk
- Sample 300 low income adolescent females

Example: Changes in Sexual Risk Categories by Accuracy of Risk Perceptions

- Based on 4 risk factors (consistent condom use, number of partners, recent STD, partner risk) categorized women into three groups :
 - No risk
 - Moderate risk
 - High risk

Example: Changes in Sexual Risk Categories by Accuracy of Risk Perceptions

- Classified participants as accurate and inaccurate risk perceivers
- Conducted two-sample CFA

Example: Changes in Sexual Risk Categories by Accuracy of Risk Perceptions

Table of results

Configuration	f	statistic	p	pi*	Type?
11	54.				
12	8.	49.812	.000000	.466	Discrimination Type
21	20.				
22	6.	9.305	.002285	.396	Discrimination Type
31	15.				
32	1.	13.238	.000274	.504	Discrimination Type
41	6.				
42	15.	2.191	.138781	.251	
51	12.				
52	67.	40.587	.000000	.369	Discrimination Type
61	8.				
62	9.	.041	.840079	.012	
71	2.				
72	15.	7.319	.006823	.394	Discrimination Type
81	7.				
82	16.	1.930	.164810	.231	
91	15.				
92	23.	.568	.450928	.116	

Example: Changes in Sexual Risk Categories by Accuracy of Risk Perceptions

- Found 5 cells that were discrimination types:
 - Low risk women were more likely to remain low risk if accurately perceived risk
 - Low risk women were more likely to move to higher risk group if they overestimated their risk
 - Moderate risk women more likely to remain moderate risk if underestimated risk
 - High risk women more likely to move to low risk if underestimated risk

Configural Frequency Analysis: Strengths and Weaknesses

Strengths

- Can test very specific hypotheses
- Explicitly model IV and DV combinations
- Identify small sized but important subgroups
- Model complex variable patterns
- Exploratory and confirmatory
- Cross-sectional and longitudinal
- Based on strong statistical reasoning

Weaknesses

- Data level needs to be categorical
- Difficulty handling missing data
- Can not test overall model goodness of fit
- Limited in number of variables can easily model
- Statistical software not widely available