**CSUMB 550**

**Lecture 1. Introduction to Multivariate Data Analysis**

**October 15, 2015**

**Lisa Schwarz**

1. What it is: A way of mathematically and statistically comparing or grouping variables that are usually similar to each other
2. Why we use it
	1. The world is messy and univariate statistics often does not cut it
	2. Reduce Type I error (randomly determining a relationship is significant when it is not)
	3. Summarize redundancies
	4. Detect true multivariate patterns in the data
3. Example: How do three different species of ground-dwelling rodents coexist in the same area?
	* 1. Capture 30 individuals of each species
		2. Collect 10 niche variables: predation, plants, water, substrate, hiding places, denning places **THESE VARIABLES ARE CORRELATED!!!**
		3. Cloud of 90 rows and 10 columns + a column for Indiv ID and Species type
4. More about what it is
	1. Descriptive:
		1. rules for combining the variables in an “optimal” way
		2. no assumptions about distributions
	2. Inferential:
		1. controls the error rate
		2. usual parametric assumptions: normality and homogeneity
	3. Not always independent
	4. Often used to describe secondary relationships (post hoc)
5. Types of data
	1. Categorical
		1. Dichotomous
		2. Polytomous (**po-*lit*-a-mus**)
			1. Nominal scale: hair color; Order of categories has no mathematical meaning
			2. Ordinal scale: satisfaction scale; Order matters but differences are not known
	2. Continuous
		1. Interval scale: Celsius, time; Differences or “interval” is known, but no true zero (no multiplication or division)
		2. Ratio scale: Mass, height; Have it all
	3. Count
		1. integer values only (discrete and indivisible units of the same size);
		2. often use the same statistics
6. $y=β\_{0}+β\_{1}x\_{1}+β\_{2}x\_{2}+β\_{3}x\_{3}+β\_{4}x\_{4}$
	1. x = independent, predictor, regressor, explanatory
	2. y = dependent, response, outcome
	3. Not applicable for some multivariate techniques = variables
7. Types of Techniques
	1. Ordination
		1. Principle Components Analysis, Factor Analysis, Multidimensional Scaling, Correspondence Analysis, Canonical Correspondence Analysis
		2. Quantify the relationships between interdependent (correlated) variables and explain them in a smaller set of underlying dimensions
		3. Condense the information into a smaller set of dimensions
		4. Data
			1. a single set of two or more continuous, categorical, and/or count variables
			2. no distinction between independent and dependent variables
	2. Cluster Analysis
		1. develop meaningful aggregations or groups of entities
		2. Data
			1. a single set of two or more continuous, categorical, and/or count variables
			2. no distinction between independent and dependent variables
	3. Discriminant Analysis – Logistic regression more common these days because of the assumptions
		1. Discriminant Analysis, Classification Rules, Multivariate Tests for Means (MANOVA), Classification and Regression Trees (CART)
		2. Steps
			1. Describe the differences among two or more well-defined groups
			2. Predict which group a new entity will fall in to given its discriminating characteristics
		3. Data
			1. one categorical grouping with at least two levels
			2. two or more continuous, categorical, and/or count variables
			3. independent and dependent variables assigned based on sampling design and the research situation (MANOVA: predict continuous variable given the categorical variable)
	4. Canonical Correlation Analysis – Multivariate (normal) regression becoming more common these days, but very computationally difficult
		1. An extension of multiple regression analysis
		2. Two or more sets of data where one set of data is the dependent variable (dependence can be arbitrary)
		3. Each set has two or more continuous, categorical, and/or count variables