# *Syllabus for STA 559, Spring 2011*

***Instructor***: Robert Pruzek, Professor (rmpruzek@yahoo.com, fax & phone: 518-402-0391) Most communications with students to be within the **wiki** prepared for the course: sta559s11.pbworks.com (first page open for all to see; other pages require registration).

***Teaching Assistant:*** none

***Prerequisites:*** HSTA 558 or equivalent (see wiki).

***Schedule:*** Each week, classes meet on Tuesdays and Thursdays from 1:00 until 2:20 pm in the SPH Computer Classroom. (This facilitates computer-based interactions where each student has control of the software used for data analyses and graphics.) Assignments are spelled out in detail on website which references all assignments, projects, Quizzes and Exams, as well as their times; the main topics for the course are shown below. There is no general paper for this course.

***Textbook:*** No textbook is used. Rather, numerous documents, mostly pdfs, and various URL links, are used to support student learning. These range from wiki files (mostly) prepared by the instructor to cover the topics on which we focus, to encyclopedia, possibly Wikipedia, entries, as well as applets for demonstrations and commentaries. It is expected that students will regularly access and take advantage of these and other resources that you acquire (perhaps from the web).

***The course is Graded*: A-E**. Grades determined principally by student performances on two major exams as well as the final (adding all points). Participation, in the form of project work, submitted homework exercises (one per week), and discussion is also required. But while all homework is assessed, and detailed feedback is provided to each student, these assessments do not contribute to student grades. Interactions with students are in general (except in special situations) shared with all students online. Guidelines for homework are provided on the course wiki. (As a graduate class most grades are either A's or B's; consistent and effective participation usually ensures at least a B- in this class.)

***Catalog Description:*** *(latest version not available at this writing)*

***Course Overview:***

Statistics is sometimes thought of as a mathematical discipline. However it is more correctly thought of as a science, in particular, the science of data analysis. While data analysis and statistical methods often rely heavily upon mathematics, they also rely very much on computers, both hardware and software. Above all, effective practice of data analysis involves ingenuity or creativity within the context of applications. Because the most meaningful or useful data arise from well-formulated or carefully considered questions as well as careful planning, it follows that effective use of data analytic or statistical methods requires close attention to the questions that drove the collection of those data. When feasible, we shall discuss questions that drove initial data collection, and to some extent even the relevant design issues. But inevitably courses in statistics have to focus more on descriptions of and inferences from data than about background or substantive questions, with the (central) expectation that the student will try to keep research questions and contexts in mind. Initial purposes can sometimes be inferred, but this will often be difficult to do with assurance.

You are advised to be as focused as possible to maximize your opportunities to learn how statistical and data analytic methods and thinking can facilitate learning, not just of statistics, but of research issues and questions generally. The specific goals of this course will be spelled at the beginning of study of particular topics, or in the reviews of topics to be covered in the three major exams. The latter are provided below. Numerous handouts will be provided on the class wiki.

My fundamental concerns are that you begin to acquire sound working principles and concepts, that you learn to think carefully about statistics and graphics and that you learn how to execute effective data analyses, and corresponding graphics, as well as interpret results using sound language. A good deal of attention in any modern course on data analysis must focus on graphics and visualization of data, as well as comparisons of groups and studies of relationships in data; also on reasoning from such visual information. The latter will therefore be emphasized.

**The main topics on which we shall concentrate are the following:**

- Description of groups (for continuous variables), and related inferences (exploratory approaches & visualization to be emphasized)
- ANOVA methods, and related design issues; emphasis on graphics and interpretation
- Design of sample data collection and of selection of variables
- Bootstrapping methods and simulation (extending 558 coverage)
- Qualitative (categorical) & quantitative Data (especially displays, and relational issues)
- Graphical representations of categorical data (e.g. mosaic plots)
- Aids for improving interpretations, especially through visualization
- Analyses of categorical and partitioned data, numerical and graphical
- Methods for cluster analysis (of variables or of individuals)
- Hierarchical and non-hierarchical clustering (related to classification; graphics essential)
- Linear and non-linear methods for regression & prediction
- Missing data: issues, methods for dealing w/ MD, especially (multiple) imputation
- Classification and regression trees; especially as related to PSA (below)
- Matching and Propensity score analysis (PSA)
- Missing data: issues, methods for dealing w/ MD, especially (multiple) imputation
- Matrix operations (multiplying, factoring, eigen analyses, svd, spectral decomposition)
- Longitudinal data analysis, especially exploratory forms, and graphics (if time permits)
- Principal component analysis (PCA) and related algebra
- Dimension reduction for within and between set problems (related to several topics)
- Methods and concepts that account for errors in variables (e.g. factor analysis)
- Latent variable methods to study general structural relationships among variables (if time permits)
- Methods for transformation in multivariate contexts (*e.g.,* rotation of factors)
- Discriminant Methods (DFA), and MANOVA (if time permits)

In our coverage of the preceding topics you should be thinking about questions you want to be able to answer about the respective methods. To help you do this, I list the following as *typical questions you should be able to answer in class and on Exams*:

- How can one visualize simple data sets, such as for comparing 2 groups; also, many groups, possibly structured?
- What are independent samples (in experiments, observational studies and surveys)?
- Illustrate mean differences, effect sizes, t-statistics, confidence intervals; and how are they related to one another?
- What are dependent samples, and what advantages might ds's have over is's in practice?
- Be able to distinguish four main types (1a, 1b, 2a & 2b) in Jour. of Stat. Educ. article you read; also, be able to interpret granova.ds graphics, and related numerics.
- Describe the role of dependency, for all major 'paradigms', including related details.
- Why is randomization often recommended? What is lost without randomization?
- What is meant by the term 'choice of metric' when comparing groups?
- Describe planned comparisons [PCs] and illustrate how they can be used to advantage.

- Be able to assess whether two contrast vectors are mutually orthogonal, or not.
- Two-way and higher order ANOVA; blocking and factorial designs
- Describe roles of blocking? Elaborate (be able to give good examples). Efficiency issues
- What are row and column effects? How are they generally assessed?
- What are main effects? What are interactions? Be able to interpret, given examples.
- How can metric choice influence results? When might you want to transform the response variable? Explain.
- How does bootstrapping work? Be able to describe its basic rationale and why it can be useful, such as for scalar and vector statistics; and how can it be used for C.I. generation?
- What are Categorical variables & how do they differ from continuous ones?
- Be able to illustrate both description of relationships between two and three categorical variables, and also how to test hypotheses in such cases.
- What is an indicator matrix? What are its main characteristics?
- What are 'expected values' in the context of a 'cross-tabs' table in the context of examining relationships between two categorical variables? How is chi square statistic computed and used?
- What is correspondence analysis? How does it work? When can it be most useful?
- What is Logistic Regression (LR)? Be able to describe, also interpret, major features of a logistic regression, e.g., what is meant by a logit?
- What is the main use of LR in propensity score analysis (PSA)?
- What is propensity score analysis? What is a propensity score? What is needed to estimate the PS? What is the logic that underpins use of p-scores?
- What statistical questions are especially central to PSA applications? How does the quality and comprehensiveness of a set of covariates generally affect PSA interpretations?
- Be able to interpret PSA graphics for both Phase I and Phase II analyses.
- If you aimed to be 'critical' of a propensity score analysis (an applied study), what are some of the key issues on which you would be most likely to focus?
- What are some of the main methods and purposes for cluster analysis?
- What is meant by the terms MAR and MCAR in modern missing data analysis. Be able to give examples of each.
- Identify two key challenges that often arise in dealing with missing data. Elaborate.
- Describe one key graphic used in correspondence analysis (CA); be able to interpret categorical data results presented in a graphic for CA.
- What is meant by inner and outer products of vectors? Matrices?
- What is meant by the term 'rank' of a matrix? Be able to give an example of a matrix of a given rank (like 1 or 2 or k).
- Be able to describe two different ways to compute or to interpret the product of two matrices.
- What can you say about a covariance or correlation matrix based on knowledge of its eigenvalues?
- What is principal component analysis? Be able to describe using matrix manipulations.
- Be able to describe or interpret various results of a principal component or common factor analysis: especially, eigenvectors, eigenvalues, singular values, common factor loadings, communalities?
- And what role(s) can be served by factor transformation? Elaborate, based on what you find in the two articles on f.a. posted on the wiki.
- What are the main goals and methods of longitudinal data analysis, including two basic approaches to (or modes for) smoothing? How effectively to plot longitudinal data?

**Reasonable Accommodations:** Reasonable accommodations are provided for students with documented physical, sensory, systemic, cognitive, learning and psychiatric disabilities. If you believe you have a disability requiring accommodation in this class, please notify the Director of

Disabled Student Services (Campus Center 137, 442-5490). That office will provide the course instructor with verification of your disability, and will recommend appropriate accommodations. For more information, visit the website of the UAlbany Office for Disabled Student Services: http://www.albany.edu/studentlife/DSS/guidelines/accommodation.html

**Academic Integrity (A comparable statement from Universities' Admissions & Graduate Requirements appears on the moodle website, and is referenced at Exam times)**

Whatever you produce for this course should be your own work and created specifically for this course. You cannot present work produced by others, nor offer any work that you presented or will present to another course. If you borrow text or media from another source or paraphrase substantial ideas from someone else, you must provide a reference to your source.

The university policy on academic dishonesty is clearly outlined in the Student Bulletin, and includes, but is not limited to plagiarism, cheating on examinations, multiple submissions, forgery, unauthorized collaboration, and falsification. These are serious infractions of University regulations and could result in a failing grade for the work in question, a failing grade in the course, or dismissal from the University.