Syllabus for STAT 504: Analysis of Discrete Data
Spring 2012, updated 01/09/2012

Class Schedule: T, 9:45-11:00, 216 Borland.
                    Th, 9:45-11:00, 007 Life Sci.

Instructor: Professor Qunhua Li
Email: qunhua.li@psu.edu
Office: 421A Thomas Building
Office Hours: Th 2pm-3:45pm or by appointment

Teaching Assistant: Wei Sun, wus122@psu.edu
Office: 330 Thomas Building
Office hours: W 2pm-4pm or by appointment

Textbook:
An Introduction to Categorical Data Analysis by Alan Agresti, 2nd edition (2007), Wiley.
or

Software: SAS, R, etc...
Information: Announcement, handouts & homework will be available on ANGEL

Course Objectives
- To develop a critical approach to the analysis of contingency table data
- To examine the basic ideas and methods of generalized linear models
- To link logit and log-liner methods and graphical model with generalized linear models
- To develop facility in the analysis of discrete data using SAS/R and other programs

Prerequisite
Stat 504 is intended primarily for graduate students outside of the Statistics department. It may also be appropriate for first- or second year graduate students in Statistics. Advanced graduate students in Statistics should consider taking Stat 544 instead.
- Stat 504 assumes knowledge of basic techniques of applied statistics, including normal theory, confidence intervals and hypothesis tests (i.e. one and two-sample t-tests, etc.), multiple linear regression and basic analysis of variance.
- A course in applied probability, or at least some familiarity with discrete probability and distributions, expectation, variance, etc. is important.
- Students are expected to have basic mathematical ability to deal with summations, square roots, logarithms, etc. and occasionally some simple matrix algebra.
Textbook

- Agresti, Alan (2002). *Categorical Data Analysis*, Second Edition, Wiley. If you want more mathematically rigorous version of the first book, this is a popular and highly cited reference book on categorical data. Some of the lectures will follow this book closely, and others will not. It may be possible to survive without purchasing either Agresti (2002) or Agresti(2007), but either book is worth owning.

Suggested Reading Materials:


Computing

We will use SAS (http://css.its.psu.edu/es/ilsd/sas.html), and R (http://www.cran.r-project.org/). Students who wish to use other packages (S-PLUS, SPSS, Stata, etc.) are welcome to do so. Users of these other packages, however, will be responsible for teaching themselves how to perform the analyses in these packages, and for ensuring that the results are consistent with what they would obtain from SAS/R. Sample analyses in SAS and/or R will be provided throughout the course. Students who use other statistical packages should probably re-work these examples to make sure that they obtain the same results. The amount of class time that we can devote to computer issues is limited. Students who encounter difficulties in computing will be expected to seek help outside of class time – in office hours with the instructor and grader or, preferably, through working together with other students in the class.

A book that maybe useful is Categorical Data Analysis Using the SAS System, Second Edition by Stokes, Davis, and Koch (2001, SAS Institute). This book is not required, but can be helpful for graduate students who anticipate doing a lot of categorical data analysis in SAS. This book was written by biostatisticians and has a strong biostatistical flavor. It focuses on the mechanics of performing analyses in SAS, rather than on the underlying statistical principles.

Class attendance and participation

This course will cover a broad range of topics, and will frequently go beyond material found in the textbooks. Students will be responsible for all material covered in class,
whether or not it is found in the textbooks. Hence it is essential for students to attend class on a regular basis, participate in class and take good notes. You are encouraged to ask questions whenever you have them. Your questions show me both what I have made clear and what needs to be clarified and, consequently, they help me to teach more effectively.

**Course Grading**
The course grade will be based on the following allocation:

- homework – 40%
- two exams – 40% (each exam: 20%)
- course project – 20%

**Homework**
Homework assignments will be given frequently throughout the semester, typically distributed on Thursday and due on the stated due date. It is your responsibility to download them from the course home page. The assignments will contain both data analysis exercises and conceptual/theoretical questions that challenge your understanding of the key ideas.

The homework SHOULD be typed, especially data analysis part. If hand written, it should be legible. Clear writing and presentation are important parts of the assignments. Applied statistical analyses are useless without clear explanations. Thus, you should NOT include raw computer output in your reports.

**Collaborative work**
You are encouraged to work together – for example, to help one another with computer issues, to share class notes or to discuss the materials, etc. On the homework assignments, a reasonable amount of collaboration is allowed. Each student, however, must turn in his or her own written work which reflects his or her own individual analysis and understanding of the material. Because this is a graduate course, the students will be assumed to have sufficient motivation and maturity to come to their own understanding of the material without exams or a strict working alone policy.

**Exams:** There are 2 exams, each worth 20% of your grade. Time: TBA.

**Course Project:** The group project is an opportunity for you to work on a problem of interest to you, by utilizing methods and tools for analysis of discrete data. You will work in groups of 3, and will need to submit a proposal to ensure that the project is acceptable for this course. More information will be provided later in the semester.

**Outline (TENTATIVE) of course**
The following outline is tentative, and may be modified as the semester progresses, according to the interests of students and the discretion of the instructor.

1. Quick review of discrete probability distributions: binomial, multinomial, Poisson.
   Introduction to the concept of likelihood.
3. Introduction to contingency tables. 2x2 and r x c tables, tests for independence and homogeneity of proportions, Fishers exact test, odds ratio and logit, other measures of association.
4. Introduction to 3-way tables, full independence and conditional independence, collapsing and Simpson’s paradox.
5. Introduction to generalized linear model.
6. Logistic regression for dichotomous response, including interpretation of coefficients, main effects and interactions, model selection, diagnostics, and assessing goodness of fit.
7. Poisson regression and Log-linear models.
8. Polytomous logit models for ordinal and nominal response.
9. Other topics as time permits (and due to the interests): repeated measures, generalized least squares, mixed models, latent-class models, missing data.

**Physically disabled and learning disabled students**

It is Penn State’s policy to not discriminate against qualified students with documented disabilities in its educational programs. If you have a disability related need for modifications in this course, contact your instructor and the Office for Disability Services (located in 116 Boucke Building) or the Disability Contact Liaison at your Penn State location. Instructors should be notified as early in the semester as possible. You may refer to the Nondiscrimination Policy in the Student Guide to University Policies and Rules 1997.

**Plagiarism**

Feel free to talk to me if you have any questions or comments about what constitutes plagiarism. All Penn State and Eberly College of Science policies regarding academic integrity apply to this course. See: http://www.science.psu.edu/academic/Integrity/index.html for details.