

# (TENTATIVE) Course Description and Syllabus for STAT 504: Analysis of Discrete Data Spring 2005, updated 12/14/2004

<b>Class Schedule</b>	Lecture: Tue, 9:45-11:00 220 Thomas Lecture/Lab: Thr 9:45-11:00 069 Willard
<b>Professor</b>	Aleksandra (Seša) Slavković Office: 412 Thomas Phone: 863-4918 Email: sesa@stat.psu.edu
<b>Office Hours</b>	Tue 2-4, and by appointment
<b>TA</b>	TBA Office: TBA Thomas Email: TBA@stat.psu.edu
<b>Office Hours</b>	TBA
<b>Text</b>	<i>Categorical Data Analysis</i> by Alan Agresti 2nd edition (2002), Wiley, ISBN: 0471360937
<b>Software</b>	SAS, R, Splus, Minitab, etc...
<b>Information</b>	Announcement, handouts & homework at <a href="http://www.stat.psu.edu/~sesa/stat504/">http://www.stat.psu.edu/~sesa/stat504/</a>

## 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.

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 Agresti (2002), but the book is definitely worth owning. I've placed a request to put the book on reserve in the PAMS library.

## Alternative Texts and Suggested Reading Materials:

- Agresti, Alan (1996). *An Introduction to Categorical Data Analysis*, Wiley.
- Bishop, Y.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT Press.
- Edwards, D. (2000). *Introduction to Graphical Modeling*. Second Edition, Springer.
- Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. MIT Press.
- Wasserman, L. (2004) *All of Statistics: A Concise Course in Statistical Inference*. Springer. (<http://www.stat.cmu.edu/larry/all-of-statistics/index.html>)
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

## Computing

We will primarily 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, Minitab, 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. 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, but I will try to incorporate some hands-on experience during the lectures/labs on Thursdays.* 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 may be useful is *Categorical Data Analysis Using the SAS System*, Second Edition by Stokes, Davis, and Koch (2001, SAS Institute). This book is not required, however, it can be helpful for graduate students who anticipate doing a lot of categorical data analysis in SAS in their future research. 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. I've also placed a request to put the book on reserve in the PAMS library.

## **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 absolutely essential for students to attend class on a regular basis and to take good notes.

You are encouraged to participate in class discussion. This helps you become more comfortable with the material, and, at the same time, gives other members of the class the benefit of your ideas and perspective. Students will be asked questions regarding current course material during each class meeting. In particular, you should 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. Remember, we learn from our mistakes too!

## **Course Grading**

The course grade will be based on the following allocation:

- homeworks – 70%
- final take-home data analysis exam – 30%

## **Homework**

Homework assignments will be given frequently throughout the semester, typically distributed on Thursday and due the following Thursday. *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. 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 and discuss the material, 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.

### **Final exam**

The last assignment will be a take-home final; it will be more comprehensive and longer than the others, and it will be worth 30% of the final grade. It will focus on analysis of categorical data sets, but will also include at least one "theoretical" problem.

### **Outline 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. Tests for one-way tables using Pearsons  $X^2$  and likelihood-ratio  $G^2$  statistics.
2. Introduction to contingency tables.  $2 \times 2$  and  $r \times c$  tables, tests for independence and homogeneity of proportions, Fishers exact test, odds ratio and logit, other measures of association. Introduction to 3-way tables, full independence and conditional independence, collapsing and Simpsons paradox.
3. Introduction to generalized linear models. Poisson regression. Logistic regression for dichotomous response, including interpretation of coefficients, main effects and interactions, model selection, diagnostics, and assessing goodness of fit.
4. Polytomous logit models for ordinal and nominal response.
5. Loglinear models (and graphical models) for multi-way tables.
6. Other topics as time permits (and due to the interests) : causality, repeated measures, generalized least squares, mixed models, latent-class models, missing data, algebraic statistics approach.

### **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 come 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.