STA4504/5503 CATEGORICAL DATA ANALYSIS SPRING 2010

Instructor: Alan Agresti (I am an emeritus faculty member at UF but I am teaching here part-time during spring semester 2010.)

Office: Griffin-Floyd 204

Phone: (352) 273-2981

E-mail: aa@stat.ufl.edu

Office hours: Monday and Wednesday 3-5, and by appointment.

Teaching assistant: Quan Tran, 117D Griffin-Floyd, quandtran@stat.ufl.edu,
office hours Monday, Wednesday, Friday 2-4 pm.
Tran will handle questions about the homework exercises, including software questions, and in my office hours I will handle questions about the methods themselves. Homework exercises will be graded by our second TA, Tezcan Ozrazgat, torazgat@stat.ufl.edu, meetings by appointment.

Course homepage: www.stat.ufl.edu/∼aa/sta4504

Course topics: Description and inference for binomial and multinomial variables using proportions and odds ratios, multi-way contingency tables, generalized linear models for discrete data, logistic regression for binary responses, multi-category logit models for nominal and ordinal responses, inference for matched-pairs and correlated clustered data, loglinear models.

Prerequisites: Familiarity with basic statistical methods as covered in courses such as STA 3024, STA 3032, STA 4210, STA 4322, STA 6127, STA 6167, or the consent of the instructor. Since much of this course deals with extensions of regression modeling to handle categorical response variables, students should be comfortable with multiple regression modeling, including the use of dummy variables for incorporating categorical predictors in a model, and should have had practice using statistical software such as SAS for regression and ANOVA.

Course text: *An Introduction to Categorical Data Analysis, 2nd edition*, by A. Agresti (2007), published by John Wiley & Sons. A copy is on reserve at the Science library. New and used copies are available for purchase at the UF bookstore or over the Internet. (My royalties from sales of new copies of the text for this course are donated to UF.)

Software: My lectures will illustrate computations using SAS statistical software. For the homework exercises that require software, you are welcome to use whatever software you prefer, but you will need to use some software. There is information about software for categorical data analysis at www.stat.ufl.edu/∼aa/cda/software.html. Please take advantage of the TA Quan Tran, who is available to help you with the software that you decide to use.

SAS: SAS programs and data sets from the text are available at the website,

http://www.stat.ufl.edu/∼aa/intro-cda/appendix.

See also Appendix A of the text, starting on p. 332, and the course website for other links including SAS help pages and manuals.

R and S-plus: At www.stat.ufl.edu/∼aa/cda/software.html (and the course home page) there is a link to a website of Dr. Chris Bilder, where the link to R has examples of its use for most chapters of the text. For more detailed information, there is also a link there to a comprehensive manual prepared by Dr. Laura Thompson showing how to use R and S-Plus to conduct all the types of analyses presented in this course (although the organization in her manual follows my more advanced text, *Categorical Data Analysis*, 2nd edition 2002). Another link is to a site set up by Dr. Brett Presnell when he taught this course at UF.

SPSS: At www.stat.ufl.edu/∼aa/cda/software.html I have summarized where to go on the ANALYZE menu to be able to use various methods discussed in the course.

Stata: At www.stat.ufl.edu/∼aa/cda/software.html there are some links. For examples of categorical data analyses for many data sets in the first edition of the textbook, see the useful site mentioned there that has been set up by the UCLA Statistical Computing Center.

Grading policy: Each exam will count toward 1/3 of the course grade, and the other 1/3 will be based on homework assignments.

Exam dates:

Exam 1	Tuesday, February 23
Exam 2	Wednesday, April 14

The exams are not cumulative. Although intended to be one-hour exams, they will be given in the evening so that students do not feel time pressure. Make-up exams will not be given except for medical or family emergencies, and *must be approved before the time of the exam*. Because of the extra evening periods for the exams, there will be no class on April 16 and one or two other dates to be announced.

Students (especially graduate students taking STA 5503) have the option of substituting a project for the second exam. This consists of a written report, maximum length 5 pages typed double-spaced, which can be prepared as a team with another student in the class. This report should present a statistical analysis based on modeling a data set containing a categorical response variable. The report should include sections directed toward (1) description of data and statement of questions to be addressed, (2) specification of relevant models, (3) model-fitting and checking (e.g., residual analysis or checking whether the fit improves with interaction terms), (4) interpretations of results of model fitting, (5) summary and conclusions. Be careful not to overemphasize significance testing at the expense of inference about the size of effects (e.g., confidence intervals) and not to include too many predictors for the amount of data you have. Include, in a separate appendix (not included in the page limit), edited copies of relevant parts of your computer printouts and (if possible) the data. If you choose this option, you must submit to me a one-paragraph description (hard-copy) of what you plan by March 26. The project itself is due on April 14, the date of the second exam. The two best projects will receive a bonus of 10 points (of the 100 maximum) and be asked to make a 20-minute presentation to the class about the project on the final class day, April 21.

Homework: Homework exercises are listed in the outline of course topics on the next page. ("Optional" exercises are also listed, for students who want to extend their knowledge of the methods further and have practice with more difficult exercises, but these should not be handed in.) Due dates for the exercises for each chapter will be announced in class, and a sample of the exercises will be graded. *No homeworks will be accepted after the due date*, but we will drop your lowest homework grade before calculating your homework average. To provide you with feedback about your solutions, brief outlines of the solutions to many of the homework problems are available in a pdf file at

http://www.stat.ufl.edu/∼aa/restricted/solutions-icda-hw.pdf

Short answers for odd-numbered exercises are also available at the end of the textbook. You are permitted and encouraged to work together with other students to help each other in understanding the course material and completing the homework, but you should turn in your own solutions. Those solutions should be much more detailed than the brief solutions provided, and *you must turn in software printouts for exercises that require the use of software in order to receive credit.*

Topics (with text section number)	**Text Pages**	**Homework**	**Optional**
1. Introduction			
1.1-1.3 Statistical inference for a proportion	1-10	1-4, 8, 12	15, 16
2. Contingency Tables			
2.1 Table structure	21-25	2	
2.2 Comparing proportions	25-28	3	
2.3 Odds ratio	28-34	5-8, 12	
2.4 Chi-squared tests	34-40	17-19	21, 24-26
2.6 Exact tests for small samples	45-48	29	
2.7 Association in three-way tables	49-54	33-36, 39	37, 38
3. Generalized Linear Models			
3.1 Components of generalized linear model	65-68	1, 22ab	
3.2 GLMs for binary data	68-73	2, 5	6
3.3 GLMs for count data	74-84	11-12, 16	17-18, 20-21
3.4 Inference and model checking	84-87	9, 13	14
3.5 Fitting generalized linear models	88-90		
4. Logistic Regression			
4.1 Interpreting logistic regression	99-106	1, 4	35, 36
4.2 Inference for logistic regression	106-110	2, 8	
4.3 Categorical predictors	110-115	11, 16-17, 37	
4.4 Multiple logistic regression	115-120	19, 21, 23, 24	
4.5 Summarizing effects	120-121	28	27
5. Building and Applying Logistic Regression Models			
5.1 Strategies in model selection	137-144		
5.2 Model checking	144-150	4, 15, 19, 30	20
5.3 Effects of sparse data	152-156	22	
6. Multicategory Logit Models			
6.1 Logit models for nominal responses	173-179	1, 3, 6	
6.2 Cumulative logit model for ordinal responses	179-189	5, 7, 12, 22abd	
8. Models for Matched Pairs			
8.1 Comparing dependent proportions	244-247	2, 4	7, 8, 10
8.5.5 Measuring agreement	264	20ac	
9. Modeling Clustered Responses (Repeated Measures)			
9.1 Marginal models vs. conditional models	276-279		
9.2 Marginal modeling: The GEE approach	279-284	3-4, 7, 18	
9.3 GEE for multinomial responses	285-287		
9.4 Transitional models	288-290		
10. Random Effects: Generalized Linear Mixed Models	297-309		
7. Loglinear Models			
7.1 Loglinear models for 2-way and 3-way tables	204-212	27	
7.2 Inference for loglinear models	212-223	5-7	8