STA4504/5503  CATEGORICAL DATA ANALYSIS  SPRING 2009

Period 2-3 Tuesday, 3 Thursday

Instructor: Alan Agresti  (I am an emeritus faculty member at UF but am teaching part time during spring semesters 2009 and 2010.)

Office: Griffin-Floyd 204

Phone: 352-273-2981

E-mail: aa@stat.ufl.edu

Office hours: Tuesday and Thursday 1:45-3:15, and by appointment.

Teaching assistant: Ruitao Liu
office hours Monday and Wednesday, 1-3 pm
Griffin-Floyd 115A
e-mail rliu@stat.ufl.edu

Ruitao will handle questions about the homework exercises, including software questions, and in my office hours I will handle questions about the methods themselves.

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

Course description: 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, logit models for multiple response categories, loglinear models, inference for matched-pairs and correlated clustered data.

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 (including dummy variables for incorporating categorical predictors in a model) and should have had practice using statistical software.

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. There is information about software for categorical data analysis at www.stat.ufl.edu/~aa/cda/software.html.

**SAS:** SAS programs and data sets from the text are available at the website, http://www.stat.ufl.edu/~aa/intro-cda/appendix.html

See also the appendix of the text, starting on p. 332.

**R and S-plus:** At www.stat.ufl.edu/~aa/cda/software.html there is a link to a website of Dr. Chris Bilder, where the link to R has examples of the use of R 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).

**SPSS:** At www.stat.ufl.edu/~aa/cda/software.html I have summarized where to go on the ANALYZE menu to get access to 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 set up by the UCLA Statistical Computing Center.

Please take advantage of the TA, who is available to help you with the software that you decide to use.

**Exam dates:**

Exam 1  Tuesday, February 10  (100 pts.)
Exam 2  Tuesday, March 24  (100 pts.)
Exam 3  Tuesday, April 28, 10-12 am  (100 pts.)

The exams are not cumulative. Make-up exams will not be given except for medical or family emergencies, and must be approved before the time of the exam. The final exam will not be ready until shortly before the date of that exam and cannot be taken early.

**Homework:** Required and optional homework problems are listed in the outline of course topics on the next page. (You are not responsible for the “optional” problems in the list, but those students who want to extend their knowledge of the methods further and have practice with more difficult exercises are encouraged to try them.) To provide you with feedback about your solutions, brief outlines of the solutions to the homework problems are available in a pdf file at

http://www.stat.ufl.edu/~aa/restricted
Short answers for odd-numbered exercises are also available at the end of the textbook. You are encouraged to get help from the TA with homework problems that you are unable to do and/or to work together in teams to help each other in understanding the course material and completing the homework. Some exam questions will be similar if not identical to those on the required homework list. For each exam, you will be asked to bring to the exam your solution to one or two of the exercises (to be announced at the lecture right before the exam) that requires the use of software, and it will count toward 20 of the 100 points for that exam.
Topics 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, 6
6.2 Cumulative logit model for ordinal responses 179-189 5, 7, 12, 22ab

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 (Repeatead 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

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