STA 6505 ANALYSIS OF CATEGORICAL DATA Spring 2008

Instructor: Alan Agresti

Office: Griffin-Floyd 204

Phone: 392-1941, ext. 234

E-mail: aa@stat.ufl.edu

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

Course Homepage: www.stat.ufl.edu/~aa/sta6505/index.html

Course Text: *Categorical Data Analysis*, second edition, by A. Agresti (Wiley, 2002). (This is on 2 hour reserve for this course at the Science library)

This course surveys methods for the analysis of categorical response variables. The main subject areas covered are descriptive and inferential statistics for two-way and three-way contingency tables, generalized linear models for discrete responses, binary regression models (emphasizing logistic regression), models for multi-category responses, loglinear models for contingency tables, matched pairs, and maximum likelihood inference for categorical response data.

1. Exam Schedule

Exam	Date
1	Tuesday, February 12 (100 pts)
2	Tuesday, March 25 (100 pts)
3 (Final)	Thursday, May 1, 8-10 pm (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.

2. Homework

Homework problems are listed on the outline of topics in Section 3 of this syllabus. Outlines of the solutions to most of the homework problems are available in a pdf file at

http://www.stat.ufl.edu/~aa/restricted/index-cda2.html.

You are encouraged to work together with other students on problems with which you have difficulties. Please keep a neat, organized file of solutions, including computer printouts, and hand them in with each exam. 20 points of the 100 total points on each exam will be based on the quality of this work; the maximum score of 20 will be given if detailed solutions are provided of at least 90% of the exercises. Some exam questions may be taken directly from the homework.

Some homework will require the use of software. I'll give class examples primarily using SAS. The website for the text

http://www.stat.ufl.edu/~aa/cda/cda.html

has a section with information about various software for categorical data analysis. That site has a link

https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf

to a detailed manual prepared by Dr. Laura Thompson showing how to use R and S-Plus to conduct all

the analyses in the text. I highly recommend this resource if you would like to use R for statistical analyses of categorical data.

2a. Optional Extra Credit Assignment

My lectures and the textbook uses the traditional, frequentist approach to statistical inference, with emphasis on maximum likelihood estimation and likelihood-ratio tests. Some statisticians prefer the Bayesian approach that combines a prior distribution for the parameters with the likelihood function to generate a posterior distribution for inference about the parameters. Following the second exam, students will be given an optional assignment in which they would use Bayesian methods to analyze a data set with a categorical response variable. The assignment is to prepare a report of at most 3 pages explaining the Bayesian analysis and showing how to implement it with software. The assigment will be due on the day of the final class and will be graded with a maximum possible of 10 points that would be added to your total score for the course in determining your course grade.

3. Outline of Topics and Homework Problems

Topics	Text Pages	Homework
1. Introduction: Distributions and Inference		
Discrete distributions	1-9	1, 3, 12
Inference for categorical data	10-26	7(a-d), 11, 18, 30, 33, 34
2. Describing Contingency Tables		
Probability structure	36-43	21
Comparing proportions	43-47	3, 4, 8, 10, 11, 23a, 24
Stratified tables	47-54	12, 16, 29
3. Inference for Contingency Tables		
Deriving large-sample normal distributions	70-78, 577-580	22, 24, 26, Ch 14: 5, 7
Chi-squared tests of independence	78-86	3, 4, 29, 31, 34, 35
Exact tests for small samples	91-100	13, 40, 42, 43
4. Introduction to Generalized Linear Models		
Generalized linear models	115-119	17, 30
GLMs for binary data	120-125	1, 2, 5, 19, 20, 29
Inference and fitting GLMs	143-145	22, 34
5. Logistic Regression		
Interpreting parameters	165-171	15, 28, 29, 30, 32, 33(a-b)
Inference for logistic regression	172-177	1, 4
Categorical and multiple predictors	177-192	8, 9, 12
Fitting logistic regression models	192-196	36, 37, 38
6. Building and Applying Logistic Regression Models		
Model selection	211-219	22, 23
Diagnostics	219-230	5, 6
Inference in stratified tables	230-236	7(a-d)
Power	236-245	26
Probit and complementary log-log link	245-250	14, 28, 29, 30, 32
7. Models for Multinomial Responses		
Baseline-category logit models	267-272	1, 26
Cumulative logit models	274-282	5, 7, 9, 29, 31
8. Loglinear Models		
Loglinear models for two-way tables	314-318	14, 15
Loglinear models for three-way tables	318-324	18, 19, 21
Inference for loglinear models	324-326, 333-343	1, 6, 29, 31, 32, 38b
Loglinear – logit connection	330-333	9
9. Extending Loglinear Models		
Association graphs and collapsibility	357-360	24, 25, 26
Poisson regression for rates	385-387	14, 18
Empty cells and sparseness	391-398	47
10. Models for Matched Pairs		
Comparing dependent proportions	409-413	1, 21, 22
Conditional logistic regression	414-417	23(a,b)

4. References

Some Relevant Texts:

Lloyd, C. J. (1999). *Statistical Analysis of Categorical Data*, Wiley.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall.

Santner, T., and Duffy, D. (1990) *The Statistical Analysis of Discrete Data*, Springer-Verlag.

Some Research Articles:

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," (with discussion) *Statistical Science*, 7, 131-177.

Birch, M.W. (1963), "Maximum Likelihood in Three-Way Contingency Tables," *J. Roy. Statist. Soc.*, B, 25, 220-233.

Breslow, N., and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *J. Amer. Stat. Assoc.*. 88, 9-25.

Cochran, W.G. (1954), "Some Methods of Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417-451.

Cox, D.R. (1958a). "The Regression Analysis of Binary Sequences," *J. Roy. Statist. Soc.*, B, 20, 215-242.

Goodman, L.A. (1970), "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications," *Journal of the American Statistical Association*, 65, 226-256.

Goodman, L.A. (1979), "Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories," *Journal of the American Statistical Association*, 74, 537-552.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969), "Analysis of Categorical Data by Linear Models," *Biometrics*, 25, 489-504.

Liang, K.-Y., and Zeger, S.L., and Qaqish, B. (1986). "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13-22.

McCullagh, P. (1980), "Regression Models for Ordinal Data" (with discussion), *Journal of the Royal Statistical Society*, B, 42, 109-142.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, A, 135, 370-384.

Neyman, J. (1949), "Contributions to the Theory of the χ^2 Test," *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, pp. 230-273, Berkeley: U. of California Press.