Biost 536 / Epi 536
Categorical Data Analysis in Epidemiology

Syllabus
Fall, 2012

Instructor : Scott S. Emerson, M.D., Ph.D., Professor of Biostatistics
Office : HSB H655-J
Phone : 616-6678 (Biostatistics)
Email : semerson@uw.edu
Office hours : (see webpages for current hours or by appointment)

Assistants : David Benkeser (benkeser@uw.edu)
Office hours : HSLIC (see webpages for current hours)

Emily Mosites (emosites@uw.edu)
Office hours : HSLIC (see webpages for current hours)

Time and Place : Lectures : TuTh 1:30p - 3:20p HST 639
Disc AA Tu 12:30p - 1:20p HST 733
Disc AB Th 10:30a - 11:20a HST 639

Class Web Pages: `http://www.emersonstatistics.com/b536/`

The web page will be used to post copies of the PowerPoint slides presented in lecture, homework assignments, datasets, documentation, other information, etc. I urge you to check this site regularly.

Prerequisites : Biost 513 or Biost 515/518 and Epi 513 (or permission of the instructor)

Optional Text : Hosmer, Lemeshow, & Sturdivant: *Applied Logistic Regression*, 3rd ed., Wiley
Kleinbaum & Klein: *Logistic Regression: A Self-learning Text* , 3rd ed., Springer

Computing : Software : Stata (or, for the very brave, R)

Weekly homeworks will involve statistical analyses that will generally require access to statistical software. While students may most often use the statistical software of their choice, so long as the software is capable of performing the necessary statistical procedures, help with computing assumes the use of Stata or R. Stata is available on the computers in the HSLIC. Instructions for obtaining personal copies of Stata are available on the class website. R is freely available on the web, but is generally more difficult to learn and use. Computing commands covered in lecture will be Stata, though we are happy to help with R in office hours.

Attendance : Lectures : Occasionally required, always highly recommended
Discussions : Required

Assignments : Written problem sets approximately weekly
Weekly data analyses for discussion sections
Written quizzes and discussion during lecture approximately weekly
One midterm (in class, closed notes)

.

One written report of a data analysis
Final exam (in class, closed notes)

Homework problems requiring a written solution will be due approximately every 10 days. These assignments will consist of applications of statistical methods to real data analyses. Students are encouraged to seek help from the instructor, the TAs, or other students with the written homework problems. However, the work that is handed in should reflect only that student's work. That is, obtaining help from other students in order to learn the METHODS of solution is allowed, but copying another student's answer is NOT. Owing to the peer grading process (see below), **assignments handed in late will not be accepted unless pre-approved**. We reserve the right to grade only selected portions of the written homework.

Homeworks will be peer graded. Students will be expected to submit their homeworks electronically using an anonymized code. **It is the student's responsibility to ensure that no personally identifying information is included in their submitted homework or grading.** Students will then be assigned to grade another student's homework guided by a key and grading instructions provided by the instructor or TAs. A random sample of papers will also be graded by the TAs. Any student may also appeal to the TAs or instructor regarding any grade assigned by a peer.

Approximately weekly, the first portion of class will be devoted to a short quiz. Students will be asked to prepared to 1) answer brief written questions about that lecture at the beginning of class, and 2) discuss their answers to the questions during class.

The discussion sections will be used as a data analysis laboratory in which it is envisioned that the students will gain experience in the general approach to a data analysis and in the application of the statistical methods learned in lecture. Each week, a data analysis problem will be assigned. Students will be expected to analyze the data set to address the question of interest and to come to the discussion section prepared to answer questions about their methods and results. Because this is a learning situation, it is not expected that a student will necessarily have an error-free analysis. It is expected that a student will spend 1-2 hours each week thinking carefully about the problem and attempting to apply good statistical principles to its solution. A student's contribution to the discussion of the data analysis problem will be evaluated on the following scale: 0= clearly inadequate attention to the assignment, 8= analysis attempted but major problems with the approach, 9= analysis attempted and some thought given to the best approach, 10= exemplary insight into the problems posed by the data analysis.

On one occasion during the quarter, the instructor will designate a data analysis requiring a written report from the students. The length of the report should be approximately 10 - 15 pages, and it should be written to a statistically naive reader. This will be a group project, and reports will be "refereed" by other groups. The group will also orally "defend" their report to the instructor. Further details (and examples) will be distributed later in the quarter.

Grading :		
	Written homeworks	20%
	Quizzes and oral discussion	10%
	Midterm	25%
	Report	20%
	Final examination	25%

Additional Resources

1. The following materials will be posted on the webpages:
   a. Copies of the PowerPoint slides used in lectures. The dates for each lecture are approximate, and a given lecture period may cover material from more than one handout.
   b. Supplemental notes that will not be covered in lecture, but may be of use in preparing for the data analysis laboratory.
   c. Supplemental notes on material that should be a review for most students, but which some students may need to study in detail. This material will not be covered in class.
   d. Homeworks, exams, and keys from previous quarters that I taught this class.
   e. Homework assignments (typically posted on Wednesdays and due the following Wednesday).
   f. Keys to homeworks and exams from this quarter (most times and only after the due date).
   g. Streaming audio/video of lectures (and perhaps sometimes discussion sections).
2. Electronic mail (email) will be used for communication of errata and other announcements that are of interest to the general class. I will use the email address supplied by the university course registration list. It is the student's responsibility to ensure that they are receiving emails at their desired email address. Throughout the quarter, students may submit questions regarding the course material via email. Answers to questions that I feel are of general interest will be broadcast to the entire class (the identity of the source of the question will be protected). Questions that are likely to be of interest only to a single student will usually be answered individually. I try for reasonably quick turnaround on email questions, but backlogs do occur. It may happen that I think I have answered your question in a general message broadcast to the class, but you are still unsure of the answer. Do not hesitate to send your question again, and I will try to address it further.
3. Additional texts that might prove of value include These include
   a. Agresti, *An Introduction to Categorical Data Analysis, 2rd ed.*, Wiley (2007)
   b. Agresti, *Categorical Data Analysis, 3rd ed.*, Wiley (2012)
   c. Agresti, *Analysis of Ordinal Categorical Data, 2nd ed.* Wiley (2010)
   d. Bishop, Fienberg, & Holland, *Discrete Multivariate Analysis: Theory and Practice* MIT Press (1975)
   e. Fleiss, Levin & Paik *Statistical Methods for Rates and Proportions, 3rd ed.* Wiley (2003)

Course Objectives

This course provides advanced methods related to the statistical analysis of categorical data in epidemiology. Emphasis is placed on the analysis of data to answer scientific questions. Thus the major objectives of this course are

1. To explore the ways in which statistical methods can be used to address scientific questions,
2. To present alternative categorical data analysis methods, and
3. To teach a general approach to a data analysis problem.

To those ends, this course will stress the general abstraction of descriptive and inferential statistics to address a scientific question. We will primarily address methods for the setting of a categorical response variable modeling exposure variables, confounders, precision variables, and effect modifiers. Topics covered will include univariate analyses, two sample problems, stratified analyses, logistic regression, Poisson regresson, and conditional logistic regression.

At the end of Biost 536 / Epi 536, a student should be able to:

1. Demonstrate an organized approach to the analysis of data gathered to address a scientific question in which a response variable is categorical, including illustration of the importance of separating hypothesis generation from confirmatory studies.
2. Develop an appropriate statistical model to analyze such data to address a scientific question, including
   a. refinement of vaguely stated scientific hypothesis into a statistical framework,
   b. identification of the dependent (response) variable, including a reasonable probability model for that response and a summary measure to be estimated and/or tested,
   c. identification of the independent (predictor) variables denoting any groups to be compared and the form in which those variable might be included in a regression model, allowing for any modeling of effect modification that might be indicated by the scientific question,
   d. identification of any covariates to be used to reduce confounding, along with specification of appropriate forms for inclusion of those covariates in the statistical analysis model,
   e. identification of any covariates to be used to increase precision, specification of appropriate forms for inclusion of those covariates in the statistical analysis model, and explaining the impact of such adjustment in terms of collapsibility,
   f. describe assumptions of mechanisms for any missing data and describe potential methods for analyses that might best account for those mechanisms.
3. Perform suitable descriptive analyses of the data, including descriptions relevant to the sampling scheme and any patterns of missing data.
4. Compute estimates and/or test statistics using standard statistical software including full interpretation of all parameters in th model..
5. Make statistical inference about the generalizability of the analysis results to a larger population.
6. State any statistical assumptions that are the basis for the conclusions of your analysis.
7. Perform analyses to determine whether the assumptions are sensible both on sample-wide and individual case bases.
8. Present the results of your analysis to a statistically naive reader, including a full interpretation of all parameter estimates.

Biost 536 / Epi 536 Course Outline
Fall 2013

The following is a tentative outline of the topics to be covered during the quarter. We reserve the right to modify this outline as conditions require. ("H" and K"" refer to relevant sections in the textbooks by Hosmer, et al. and Kleinbaum & Klein, respectively.)

	Date	Day	Topic	Readings	Hand In
1.	26 Sep	Thu	Course organization, Confirmatory studies		
2.	1 Oct	Tue	Summarization of categorical data		
3.	3 Oct	Thu	Role of adjustment for covariates		Submit HW #1
4.	8 Oct	Tue	Matching / stratification		Grade HW #1
5.	10 Oct	Thu	Regression models	H-1; K-1	Submit HW #2
6.	15 Oct	Tue	Reparameterization	K-2,3	Grade HW #2
7.	17 Oct	Thu	Adjusting for confounding	K-7	
8.	22 Oct	Tue	Precision / collapsibility		Submit HW #3
9.	24 Oct	Thu	Modeling dose response	H-3	Grade HW #3
10.	29 Oct	Tue	Interactions	K-7	
11.	31 Oct	Thu	Model specification		Submit HW #4
12.	5 Nov	Tue	Model building	H-4	Grade HW #4
13.	7 Nov	Thu	MIDTERM (in class, closed book)		MIDTERM
14.	12 Nov	Tue	Diagnostics	H-5; K-9	Submit HW #5
15.	14 Nov	Thu	Conditional logistic regression	H-7; K-11	Grade HW #5
16.	19 Nov	Tue	Marginal vs conditional causal inference		
17.	21 Nov	Thu	Marginal structural models, standardization		Submit HW #6
18.	26 Nov	Tue	Propensity scores	H-10.2	Grade HW #6
	28 Nov	Thu	HOLIDAY (no class)		
19.	3 Dec	Tue	Prediction; ROC curves	K-10	Submit HW #7
20.	5 Dec	Thu	Polytomous regression; proportional odds	K-12	Grade HW #7
	13 Dec	Fri	FINAL EXAM 2:30 pm - 4:20 pm		Final Exam