Syllabus for STAT 232 - 01, Categorical Data Analysis

College of the Holy Cross, Spring 2025

Professor: Neranga Fernando

Office: Haberlin 306

E-mail: nfernand@holycross.edu

Office Phone: 508-793-2458

Office Hours: 3:30pm - 5pm on Wednesdays and Thursdays, or by appointment

Meeting Times and Location: 8am - 8:50am on MWF, TBD

Course Description: Categorical Data Analysis is the second half of a full-year sequence in statistical modeling. Whereas in the first half of the course (either STAT 231 or ECON 314), you learned how to model a quantitative response variable, here our focus will shift to modeling categorical response variables. The course will cover both the theoretical foundations as well as applications of statistical modeling with a categorical response variable, including maximum likelihood estimation, contingency table analysis, and logistic regression. You will develop analytical thinking skills, learn how to communicate technical information effectively, and develop a robust understanding of the role of statistical modeling with a categorical outcome in statistical inference. Ideally, you will learn how to think critically, write coherent solutions to problems, and ideally become an articulate and confident statistician. Since modern statistical analysis is done in a computing environment, this course will have a strong computational focus. We will be using the R statistical package, which is widely used in both industry and research. As an added bonus, R is free, open-source software, so obtaining this valuable skill will come at no extra cost to you. It is expected that you have some experience with statistical software (but not necessarily R itself) from a previous statistics course.

Course Objectives: By the end of the course, all students should be able to demonstrate competency in the following areas:

(1) performing the appropriate inference test for a proportion and understanding the differences between Wald, score, and likelihood ratio tests for a proportion
(2) computing expressions for maximum likelihood estimates for functions of parameters
(3) understanding the differences between sensitivity, specificity, relative risk, and odds ratio
(4) performing tests for independence/associations between categorical predictors and response
(5) using data to fit simple/multiple logistic regression model, interpreting the model parameters, computing confidence intervals, and performing tests on the coefficients for significance;
(6) understanding the correct interpretation of the logit and odds ratio in the context of simple and multiple logistic regression
(7) performing model checking using goodness of fit tests and tests based on residuals
(8) demonstrating a robust ability to code in R and R Markdown.

Textbook: An Introduction to Categorical Data Analysis by Agresti (second edition)
Link to the book is here: `https://mregresion.wordpress.com/wp-content/uploads/2012/08/agresti-introduction-to-categ`
`pdf`

Recommended Reading: *Applied Logistic Regression* by Hosmer, Lemeshow, and Sturdivant (third edition)
Link to the book is here: `https://github.com/Drxan/Study/blob/master/Books_Need2Read/David%20W.%20Hosmer%`
`20-%20Applied%20Logistic%20Regression%20-%203rd%20Edition.pdf`

Course Materials: All announcements, materials and grades will be posted on Canvas.

Quizzes: There will be seven in-class quizzes during the semester. The lowest quiz grade will be dropped. Here are the dates you will be taking a quiz:

Quiz 1 (Feb 7), Quiz 2 (Feb 14), Quiz 3 (Feb 21), Quiz 4 (Mar 14), Quiz 5 (Mar 28), Quiz 6 (Apr 11), Quiz 7 (April 25)

It is strongly advised that you do all of assigned homework since the quizzes will closely resemble the homework problems.

Homework: Each homework assignment will be posted on Canvas a week prior to its due date. You must submit your solutions as a hard copy. Ten points will be deducted from late homework. No homework assignment will be accepted after 5 days from the due date. No help from any Internet sources is allowed. Plagiarism will not be tolerated and will be treated as a violation of the Departmental Policy on Academic Integrity.

By doing statistics you learn statistics. You learn stat best when you approach the subject as something you enjoy. Learn to explain topics to your classmates. Statistics can be fun and rewarding when there are people around you who enjoy figuring out problems as much as you do. Take advantage of this opportunity and organize study groups. I will not consider working on homework problems with your classmates as a violation of the academic honesty policy in the department. However, you must prepare and submit your own solutions. Please follow these guidelines when you submit homework assignments:

- Put your name, the date, and the homework assignment number at the top of the first page.
- Write neatly and show all your work.
- On the last page of your assignment, please write the name(s) of your classmate(s) with whom you work on homework problems (with an asterisk).
- Make sure you attach the honor code.

No homework grade will be dropped. All homework assignments are due by 4pm on the following days:

Homework 1 (Jan 31), Homework 2 (Feb 7), Homework 3 (Feb 14), Homework 4 (Feb 21), Homework 5 (Mar 14), Homework 6 (Mar 21),Homework 7 (Mar 28), Homework 8 (Apr 4), Homework 9 (Apr 11), Homework 10 (Apr 25)

Exams: There will be two exams during the semester. The exams are 90-minute exams; they will be held from 6pm to 7.30pm on February 26 (Wednesday) and April 14 (Monday). Location of exams is to be determined. No exam grade will be dropped.

Final Exam: There will be a mandatory cumulative final exam in this course. Location and time of the final exam are to be determined. **Check for final exam schedule conflicts as soon as possible.**

Snow Days: If classes are cancelled due to snow, or for other official reasons, any scheduled quiz or test will occur during next class meeting.

Grading: The course grade will be determined as follows:
Homework: 20% (2% each)
Quizzes: 15% (2.5% each)
Exams: 40% (20% each)
Final Exam: 25%

Final grades will be given according to the following percentage cutoffs. These cutoffs, although fairly strict, can be lowered (according to class performance), but not raised, no matter how well the class performs.

$A \geq 93,\ 90 \leq A- \leq 92,\ 87 \leq B+ \leq 89,\ 83 \leq B \leq 86,\ 80 \leq B- \leq 82,\ 77 \leq C+ \leq 79,\ 73 \leq C \leq 76,\ 70 \leq C- \leq 72,$

$67 \leq D+ \leq 69,\ 63 \leq D \leq 66,\ 60 \leq D- \leq 62,\ F \leq 59$

An incomplete grade is given if you have a good attendance record, have completed all the assignments with an overall grade of at least 70%, and have missed the final exam for a valid reason. An incomplete grade is given at the discretion of the instructor.

Calculators: You are allowed to use a scientific (not graphing) calculator on quizzes, mid-term exams and the final exam.

Issues with the Course/Instructor: If you have issues with this course and/or instructor which you are not comfortable discussing with your instructor, you should contact the Chair of the Department of Mathematics and Computer Science, Professor Eric Ruggieri, at eruggier@holycross.edu.

Academic Honesty: A necessary prerequisite to the attainment of the goals of the College is maintaining complete honesty in all academic work. Students are expected to present their own work in exams and in any material submitted for credit. Students may not assist others in presenting work that is not their own. Offenders are subject to disciplinary action. A violation of the Department Policy on Academic Integrity will result in a 0 for that quiz or exam, and a letter describing the occurrence of academic dishonesty will be sent to the Chair of the Department of Mathematics and Computer Science and your Class Dean. For more on Academic Integrity see:

`https://www.holycross.edu/academics/programs/mathematics-and-computer-science/node/211581/academic-integrity`

Diversity and Inclusion: It is my intent that students from all diverse backgrounds and perspectives be well-served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that students bring to this class be viewed as a resource, strength, and benefit. Any suggestions you have pertaining to diversity and inclusion are encouraged and appreciated.

Important:

(1) Any student with special needs is encouraged to meet with me during the first week of classes to discuss accommodations. The student must bring a current Memorandum of Accommodations from the Office of Accessibility Services. The following is the link to the Office of Accessibility Services:

`https://www.holycross.edu/health-wellness-and-access/office-accessibility-services`

(2) Please note that, consistent with applicable federal and state law, this course may be video/audio recorded as an accommodation only with permission from the Office of Accessibility Services. Students are not permitted to record the contents of this class under any other circumstances.

(3) If you are an athlete and have conflicts with an important class activity (homework, quiz, mid-term, or final), please let me know in advance.

(4) For College's Excused Absence Policy see:

`https://catalog.holycross.edu/requirements-policies/academic-policies/#coursepoliciestext`

(5) All electronic devices (mobile phones, laptops etc.) must be turned off during class time, quizzes, mid-term exams and final exam.

Syllabus: Syllabus is subject to change. It is your responsibility to be aware of any changes I may make to the syllabus as they are announced in class. Students are responsible for all information given when they are absent.

Some Additional Notes:

(1) I will hold an additional 2-hour final exam review session the day before (or two days before) the final exam. We will discuss and find a time that works for all of us. I will let you know the location before you go home for Easter Break.

(2) I will hand out worksheets in class. Since we do not have time to work on all the problems on problem sheets in class, I will post their solutions on Canvas. However, I encourage you all to work on the problems and bring questions to my office hours.

Important Dates:

March 3 – 7	Spring Break: no classes
April 17, 18 & 21	Easter Break: no classes
April 23	Academic Conference Day: no classes

Schedule of Topics

Introduction

Types of Data

Probability Distributions for Categorical Data

Statistical Inference for Proportions

Maximum Likelihood Estimation

The χ^2 Distribution, Wald, Score, and Maximum Likelihood Tests, Small Sample and Bayesian Inference

Contingency Tables

Probabilistic Structure of Contingency Tables

Comparing Proportions in 2×2 Contingency Tables; Relative Risk

Odds Ratios

χ^2 Test of Independence

Fisher Exact Test; Bayesian Inference

Association in Three-Way Tables; Simpson's Paradox

Paired Categorical Data – McNemar's Test

Logistic Regression

Components of a Generalized Linear Model

GLM for binary and count data

Introduction to Logistic Regression incl. Maximum Likelihood and Contingency Tables

Fitting and Interpreting a Logistic Regression Model

Inference on Logistic Regression Models

Types of Predictor Variables

Multiple Logistic Regression

Multiple Logistic Regression

Interpreting the Multiple Logistic Regression Model, Model Utility

Inference in Multiple Logistic Regression

Nested LR Tests

Summarizing Predictive Power: ROC Curves

Building Logistic Regression Models

Strategies in Model Selection; Stepwise Regression

AIC, BIC

Goodness of Fit; Model Comparison Using Deviance

Hosmer-Lemeshow Test

Final Exam Review

May 5, Monday, Last day of classes
May 10, Saturday – May 15, Thursday, Final Exams
Final Exam is based on all sections covered in class.

The mind is not a vessel to be filled but a fire to be kindled.

— Plutarch