CHL5210H Categorical Data Analysis
Fall 2021

Course Instructor: Tony Panzarella, Dalla Lana School of Public Health, University of Toronto (tony.panzarella@utoronto.ca)

Lectures: Wednesdays, 2 – 5 pm; Online

TA: Sangook Kim
Tutorials: Fridays, 11 am – 12 noon; Online

Overall Objective:
This online course will provide students with an understanding of the statistical methods for categorical data analysis. These include traditional methods for two-way contingency tables. Most of the course will focus on regression models, with an emphasis on logistic regression models. Analysis of repeated categorical response data, namely marginal and random effects models will also be included. Inference using the method of maximum likelihood estimation will be emphasized. The course includes both application and theory.

Prerequisites:

Statistics at the graduate level or consent of instructor.
Working knowledge of SAS or equivalent is an asset

Enrollment:

This is a graduate level course for students of biostatistics, epidemiology, and statistics.

Teaching Format
The course will meet once per week for 3 hours on Wednesdays. Sessions will take the form of lectures.

Student Evaluation
Midterm Test: 20%
Two take-home assignments: 40% (20% each)
Final Examination: 40%

NB: Assignments are due by **9am** of the due date

Main Reference Text

Agresti, A. 2013. Categorical Data Analysis. 3rd Edition. New Jersey: John Wiley. (*)

Other recommended texts

Collett, D. 2003. **Modelling Binary Data 2nd ed.** Chapman and Hall.

Hosmer DW, Lemeshow S, Sturdivant RX. **Applied Logistic Regression, 3rd ed**. 2013, Wiley.

Allison, Paul D. 2012. **Logistic Regression Using the SAS System: Theory and Application.** Cary, NC: SAS Institute Inc.

Stokes ME, Davis CS, Koch GG. 2012. **Categorical Data Analysis Using the SAS System. 3rd Edition**. Cary, NC: SAS Institute.

(*) Chapters covered with specific sections per chapter are listed below

Software
SAS Version 9.4. If you do not yet have a SAS license, please email Yeonkyung Namkoong in the Biostatistics Division for assistance: biostat.dlsph@utoronto.ca

Late Policy
There will be a 5% penalty applied each day for late assignments.

Academic Integrity

Academic integrity is essential to the pursuit of learning and scholarship in a university, and to ensuring that a degree from the University of Toronto is a strong signal of each student's individual academic achievement. As a result, the University treats cases of cheating and plagiarism very seriously. The University of Toronto's Code of Behaviour on Academic Matters outlines the behaviours that constitute academic dishonesty and the processes for addressing academic offences: (http://www.governingcouncil.utoronto.ca/Assets/Governing+Council+Digital+Assets/Policies/PDF/ppjun011995.pdf)
University of Toronto's policy regarding plagiarism:
http://www.writing.utoronto.ca/advice/using-sources/how-not-to-plagiarize
Potential offences include, but are not limited to:
In papers and assignments:
•Using someone else's ideas or words without appropriate acknowledgement.
•Submitting your own work in more than one course without the permission of the instructor.
•Making up sources or facts.
•Obtaining or providing unauthorized assistance on any assignment.

On tests and exams:
•Using or possessing unauthorized aids.
•Looking at someone else's answers during an exam or test.
•Misrepresenting your identity.

Turnitin.com

Turnitin.com is a tool that will assist in detecting textual similarities between compared works. Normally, students will be required to submit their course essays to Turnitin.com for a review of textual similarity and detection of possible plagiarism. In doing so, students will allow their essays to be included as source documents in the Turnitin.com reference database, where they will be used solely for the purpose of detecting plagiarism. The terms that apply to the University's use of the Turnitin.com service are described on the Turnitin.com web site.

Accessibility and Accommodation

The University provides academic accommodations for students with disabilities in accordance with the terms of the Ontario Human Rights Code. This occurs through a collaborative process that acknowledges a collective obligation to develop an accessible learning environment that both meets the needs of students and preserves the essential academic requirements of the University's courses and programs. For more information, or to register with Accessibility Services, please visit: http://studentlife.utoronto.ca/as

Acknowledgment of Territory

We would like to acknowledge the traditional territories of the Mississauga of the New Credit First Nation, Anishnawbe, Wendat, Huron, and Haudenosaunee Indigenous Peoples on which the Dalla Lana School of Public Health now stands. The territory was the subject of the Dish With One Spoon Wampum Belt Covenant, an agreement between the Iroquois Confederacy and Confederacy of the Ojibwe and allied nations to peaceably share and care for the resources around the Great Lakes. We would also like to pay our respects to all our ancestors and to our present Elders.

CHL5210H Categorical Data Analysis - Course Overview

Week	Date	Topic	Chapter(s) covered	SAS tutorial	Assignment	Assignment Due
1	Sep 8	Introduction; Distributions for Categorical Data; Statistical Inference	Ch. 1 (1.1-1.4)	SAS PROC FREQ; syntax and inference for one proportion		
2	Sep 15	Contingency Tables – description	Ch.2 (2.1-2.3)	SAS PROC FREQ; 2 x 2 and stratified 2 x 2 tables;		
3	Sep 22	Contingency Tables – inference	Ch. 3 (3.1-3.3, 3.5)	PROC FREQ; inference for 2-way and 3-way contingency tables	Assignment #1 posted	
4	Sep 29	Logistic regression I	Ch. 5	PROC LOGISTIC syntax and examples		
5	Oct 6	Logistic regression II – Building, Checking and Applying Logistic Regression Models	Ch. 6 (6.1-6.3)	PROC LOGISTIC syntax and examples		Assignment #1 due
6	Oct 13	**READING WEEK**				
7	Oct 20	**Mid-term exam**				
8	Oct 27	Introduction to GLM	Ch. 4 (4.1-4.6)	PROC GENMOD syntax and examples		
9	Nov 3	Multinomial logistic regression	Ch. 8 (8.1-8.3)	PROC LOGISTIC syntax and examples		
10	Nov 10	Log-linear Models	Ch. 9 (9.1-9.3)	PROC GENMOD syntax and examples	Assignment #2 posted	
11	Nov 17	Analysis of Matched Pairs	Ch. 11 (11.1-11.2)	PROC FREQ, PROC CATMOD, PROC LOGISTIC for conditional logistic regression		

12	Nov 24	Modeling Correlated, Clustered Responses	Ch. 12 (12.1-12.2)	PROC GENMOD syntax and examples		Assignment #2 due
13	Dec 1	Random effects: Generalized linear mixed models	Ch. 13 (13.1-13.2, 13.5)	PROC GLIMMIX syntax and examples		
14	Dec 8	**FINAL EXAM**				

CHL5210H Course Outline and Session Objectives

Week 1: Introduction to categorical data (Sep 8)

Suggested Reading: Agresti Chapter 1, sections 1.1 to 1.4

Objectives: Course Introductions, understand the nature of categorical data

- Definition of categorical response data
- Distributions for categorical data: Binomial, Multinomial, Poisson
- Likelihood function and maximum likelihood estimation
- Wald, Score and Likelihood Ratio tests
- Inference for Binomial parameter

Tutorial: Introduction to SAS

Week 2: Analysis of contingency tables: description (Sep 15)

Suggested Reading: Agresti Chapter 2, sections 2.1-2.3

Objectives: Description and inference for contingency tables

- Joint/Marginal/Conditional distributions
- Poisson, Binomial, and Multinomial sampling
- Relative risk, odds ratio
- Conditional association in stratified 2 x 2 tables

Tutorial: PROC FREQ – single proportions and 2x2 tables, take up exercises from Lecture 1

Week 3: Analysis of contingency tables: inference (Sep 22)

Suggested Reading: Agresti Chapter 3, sections 3.1—3.3, 3.5

Objectives: Inference for contingency tables

- Confidence interval for odds ratio
- Confidence interval for difference of two proportions and relative risk
- Deriving standard errors with the Delta Method
- Testing independence in two-way contingency tables using chi-squared tests
- Pearson and Standardized Residuals
- Partitioning chi-squared test
- Fisher`s exact test for 2 x 2 tables

Tutorial: PROC FREQ – CHISQ, MEASURES and RISDIFF options in TABLES statement; EXACT statement for tables with small counts; WEIGHT statement; ORDER=DATA option in PROC FREQ statement

Week 4: Logistic Regression I (Sep 29)

Suggested Reading: Agresti Chapter 5, all sections

Objective: Understand how to use logistic regression to analyze binary categorical data

- Interpreting parameters in logistic regression
- Inference for logistic regression
- Logistic regression with categorical predictors
- Multiple logistic regression
- Model fit and diagnostics

Tutorial: Fitting and interpreting a logistic regression model in SAS using PROC LOGISTIC; outline syntax and useful options e.g. PLOTS=EFFECT, LACKFIT

Week 5: Logistic Regression II (Oct 6).

Suggested Reading: Agresti Chapter 6, sections 6.1 – 6.3

Objective: Building, Checking and Applying logistic regression models

- Strategies for model selection
- Logistic regression diagnostics
- Summarizing predictive power

Tutorial: Example of model building using a logistic regression model in SAS

Week 6: READING WEEK (Oct 13)

Week 7: Mid-term exam (Oct 20)

Week 8: Introduction to Generalized Linear Models (GLM) (Oct 27)

Suggested Reading: Agresti Chapter 4.1 – 4.6

Objective: Introduce a family of GLMs

- Components of a GLM
- GLMs for binary data
- GLMs for counts and rates
- Moments and likelihood for GLMs
- Inference and model checking for GLMs
- Fitting GLMs

Tutorial: Introduction to PROC GENMOD syntax and examples

Week 9: Models for Multinomial Response (Nov 3).

Suggested Reading: Agresti Chapter 8, sections 8.1-8.3

Objective: Generalize logistic regression to handle multinomial response variables

- Nominal responses
- Ordinal responses - logit link
- Ordinal responses – alternative link functions

Tutorial: Using PROC LOGISTIC for multinomial response variables – review examples

Week 10: Loglinear Models (Nov 10).

Suggested Reading: Agresti Chapter 9, sections 9.1-9.3

Objective: Identify when to use loglinear models, provide inference details, and show connection between loglinear models and logistic models

- Loglinear models for two-way and three-way tables
- Inference for Loglinear models
- Connection between Loglinear models and logistic models

Tutorial: Running PROC GENMOD to fit models for log-linear models – review examples

Week 11: Models for Matched Pairs (Nov 17).

Suggested Reading: Agresti Chapter 11, sections 11.1-11.2

Objective: Introduce the analysis of dependent samples using matched pairs

- Comparing dependent proportions
- Conditional logistic regression for binary matched pairs

Tutorial: Running PROC FREQ with Agree option for McNemar test; PROC LOGISTIC for conditional logistic regression.

Week 12: Clustered Categorical Data I (Nov 25)

Reading: Agresti Chapter 12, sections 12.1-12.2

Objective: Introduce models for repeated response data

- Marginal models: MLE approach
- Marginal models: GEE approach

- Tutorial: Running PROC GENMOD to fit models for repeated response data

Week 13: Cluster Categorical Data II (Dec 1).

Suggested Reading: Agresti Chapter 13, sections 13.1-13.2, 13.5

Objective: Introduce models for repeated response data

- Generalized linear mixed models
- Logistic-Normal models
- Multilevel models

Tutorial: Running PROC GLIMMIX to fit models for repeated response data

Week 13: FINAL EXAM (Dec 8)

Time: TBD