BIOSTAT702 Fall 2025: Final Project

DUE DATE: Saturday December 13, 2025 5 PM

No late work will be accepted for this project and no extensions will be given. Plan your time wisely.

This final project is worth 20% of your overall grade in the course.

By submitting an project, you are formally agreeing to the terms below and acknowledging that you have nethier given nor received unauthorized aid in the completion of the project.

Terms:

For this final project, you may use any technology or human resources available to you, including notes, exercises, Internet searches, generative AI tools, friends, instructor, etc. However, please be aware that you must submit answers to the questions written in your own words. This means that you should not quote phrases from other sources, including AI tools, even with proper attribution.

AI Usage Disclosure Policy

You must clearly disclose any use of Artificial Intelligence (AI) tools in completing this project. Your submission should include:

- Type of AI Tool Used (e.g., ChatGPT, image generator, code assistant).
- Brief Description of How It Was Used (one to two sentences explaining its role in your work).

Failure to provide this disclosure may result in a grade penalty.

Instructions:

For this assignment, you will choose a dataset and a research question given the requirements below. Your goal is to answer the research question using the techniques we have learned in class and turn in 2 deliverables.

1. Code: This is where you will run the statistical analysis. You should turn in a well-documented code file (.R or .Rmd if using R) with inline comments and section headers. The code should run start to finish without errors. All outputs in your analysis should be generated here and labeled to match those referenced in the final report.
2. Analytic Report: The analytic report should contain a brief introduction, methods section, results section, and brief discussion. See the template for more information.

Notes:

- I should see *no* code in the Analytic Report, only in the code file (If you use R-Markdown, hide the code when knitting the report, and turn in the .Rmd file for the code).

- I should not see raw code output *anywhere* in your deliverables.
- The target audience for the final report is a collaborative investigator, but it should have methods detailed enough for a statistician to thoroughly understand what you did. Figures and tables must be publication quality with readable labels.
- The target audience for the code document is a statistical supervisor or fellow statistician.
- While you don't need to use the Analytic Report template exactly, your report should contain all pieces specified in the template.
- I am *not* going to give you all the steps to answer this research question, and there is *not* one correct answer.
- You will be graded on the comprehensiveness of your report, how well you explain what you did, why you did it, and what you find.

The Dataset:

For this project, you will be analyzing data that was published by the Pew Research Center using their nationally representative American Trends Panel. Use the following to find a dataset:

- Go to this website
- Click on a Wave that might be interesting to you and download the dataset
  - You may need to create a free account. If you don't want to do this, tell me what wave you want, and I can download the data
- Click on a short read / report from that dataset that might be interesting to you
  - Pick a graphic / statistic from the report to inspire your research questions, using the parameters required below.

The Research Questions:

The goal of this project is to use logistic regression and chi-square tests to determine the association between two **binary** variables, potentially controlling for / interacting with a third. Therefore, you must pick a piece of your published Pew Research report that compares two binary variables. For your third variable, you can pick any categorical variable in the dataset you have.

To answer your research questions, you must:

- Run a chi-square test on your 2x2 table to test for independence/homogeneity/goodness of fit
- Run a logistic regression to quantify the odds ratio and compare the results to the chi-square test
- Run a CMH Test on your 2x2xK table to test whether or not the relationship exists across any of the K subgroups
- Run a Breslow Day Test to test whether or not it makes sense to report a common odds ratio
  - If it does make sense, report the common odds ratio
  - If not, report the subgroup specific odds ratios

For example, let's say I am looking at this article. I might choose age (adult vs teen) as my binary predictor and 'knows someone who is transgender' (yes vs no) as my binary outcome. Then, I might choose community type (urban vs suburban vs rural) as my third adjustment variable.

My research questions would be as follows: Is age (adult vs teen) independent of whether or not a person knows someone who is transgender? Does this relationship still hold after controlling for community type? Does this relationship differ by community type?