STA 138 Exam I Project, due
Friday, Feb 8^{th} in lecture

Read the following instructions carefully:

- **You may work in a group of two, or by yourself.**
- **You are not allowed to discuss the questions with anyone other than the instructor or TA and your group mate.**
- **Any outside help beyond that from the instructor or TA is considered plagiarism. This including asking a tutor, your classmates (for example, comparing answers), posting the questions to homework help sites, etc. Should we believe you have sought outside help, you will be reported to the Student Judicial Affairs office.**
- **You are allowed to use or modify your previous functions, or the instructors functions that are posted online.**
- **Do not share answers, or specific values for calculations, particularly on Piazza**.
- You may ask clarifying questions about code and general approach on Piazza, but do not give away any numerical answers. If you are concerned you may be giving something away, email me or the TA's directly.

Problem 1: One Categorical Variable

Every group must pick one and only one of these three datasets.

For these questions, you are also given a hypothesis of interest that you should assess.

I. `btype.csv`. This gives the column `group`, with values of blood type (AB, A, B, O). These subjects were taken from a random sample of patients undergoing a specific surgical procedure.
The hypothesis is that the proportion of blood types should be: 44% O, 25% A, 20% B, and 11% AB.

II. `born.csv`. This gives the column `Season`, with values Summer (June, July, August), Spring (March, April, May), Winter (Dec, Jan, Feb), Fall (Sep, Oct, Nov), and measured what season subjects were born in.
The hypothesis is that a subject is equally likely to be born in any season.

III. `education.csv`. This gives the column `level`, with values E (elementary), S (Secondary), C (college credits), and CD (college degree).
The hypothesis is that the proportion of jurors matches the proportion in the county, which is 40% E, 40% S, 10% C, 10% CD.

For all datasets, consider the hypothesis of interest. You should consider using both confidence intervals and hypothesis tests. You should consider correcting multiple confidence intervals. You may assume all samples are random.

Problem 2: Two Categorical Variables

Every group must pick one and only one of these three datasets.

For this problem, explore the relationship between the two variables given. **Choose one of the following datasets**:

1. `compare.csv` -This is a dataset which compares injuries between soccer and martial arts players.

   Column 1: `sport`: What sport the subject played - `Martial`, `Soccer`

   Column 2: `injury`: The injury over the last year - `Muscular`, `BrokenBone`, `Concussion`,`None`

2. `students.csv` - A poll of undergraduate students asked if they believed it was likely that they would have a job when they graduated.

   Column 1: `year`: Was the student a freshman or senior - `Freshman`, `Sophomore`, `Junior`, `Senior`.

   Column 2: `job`: If they believed they would have a job on graduation - `Yes`, `No`

3. `horror.csv` - A sociologist was interested in how gender and death type of people in horror files were related. The columns are:

   Column 1: `gen`: The gender of the subject - `Female`, `Male`

   Column 2: `death`: The type of death of the subject - `Shot`, `Stabbed`, `BFT` (Blunt force trauma), `Other`

For all datasets examine dependence of the variables, and describe any dependence found. You should consider using both confidence intervals and hypothesis tests. You should consider correcting multiple confidence intervals. You may assume all samples are random.

Guidelines

For both problem 1 and 2, your results should be written in report form, with an appendix for your code at the end of the report. Please separate your reports for 1 and 2, and staple them together at the end. Include a cover page. For problem 1 and 2, a outline of your goals follows:

1. Introduction: Briefly summarize the goal of the analysis in your own words.
2. Summary: Summarize your data. These can be plots, or sample estimates. Interpret the plots and/or estimates.
3. Analysis: Choose your tests and/or confidence intervals, and report what type of test you used and the numerical results.
4. Interpretation: Interpret your tests and/or confidence intervals in terms of the problem.
5. Conclusion: Describe and interpret your findings. What conclusions, if any, can you come to?

Write in full paragraph form when appropriate. **Fully** explore any relationships / dependence you come across. Interpret the statistical analysis you choose to do.

Ideally your report would be as follows:
Cover page.
Introduction, summary, analysis, interpretation, and conclusion for problem 1.
Introduction, summary, analysis, interpretation, and conclusion for problem 2.
Code appendix.