**Paper 2: STA 2183W**  
Initial Draft Due Thursday, October 29that 11:59PM.

In 2008, the American National Election Study (AMES) surveyed roughly 1800 individuals on various facets of their identity in order to excavate patterns in the voting behavior of the United States. The data is given in the Excel “Data-Paper2”. The variables are:

|  |  |  |
| --- | --- | --- |
| Variable | Label | Values |
| Id | Id Number |  |
| Region | Census Region | 1 = Northeast |
|  |  | 2 = Midwest |
|  |  | 3 = South |
|  |  | 4 = West |
| Ideology | Liberal/conservative self-placement | 0 = No Answer |
|  |  | 1 = Extremely Liberal |
|  |  | 2 = Liberal |
|  |  | 3 = Slightly Liberal |
|  |  | 4 = Moderate |
|  |  | 5 = Slightly Conservative |
|  |  | 6 = Conservative |
|  |  | 7 = Extremely Conservative |
|  |  | 8 = Don’t Know |
|  |  | 9 = Refused |
| Ethnic | Race/ethnicity of Respondent | 1 = Black |
|  |  | 2 = Asian |
|  |  | 3 = Native American |
|  |  | 4 = Hispanic |
|  |  | 5 = White |
|  |  | 6 = Other |
|  |  | 7 = Don’t Know |
|  |  | 8 = Refused |
|  |  | 9 = No Answer |
| Gender | Respondent gender | 1 = Man |
|  |  | 2 = Woman |

The political firm *PolitiTACT* wishes to utilize the results from the 2008 AMES survey.Their goal is to see what effect various categorical factors have on political ideology, with the intent on utilizing such patterns to revise public policy, one facet including whether specific claims widely assumed to be true have any substantive merit. For the scope of this report, the following categorizations will be used for analysis:

* An individual identifies as **politically** **moderate** if they either selected moderate, slightly liberal, or slightly conservative for the ***Ideology*** variable.
* An individual identifies as **politically** **liberal** if they either selected liberal or extremely liberal for the ***Ideology*** variable.
* An individual identifies as **politically** **conservative** if they either selected conservative or extremely conservative for the ***Ideology*** variable.

The firm has hired you as a statistical consultant in order to address the following claims and determine their merit from a statistical framework (at a 5% significance level):

**Claim 1:**

**Much of the population is composed of moderates, with a small portion on each side who are strongly identify as “liberal” or “conservative”.**

In doing so, *PolitiTACT* wants for you to test to see if the following proportions are true:

* 60% of the population identifies as **politically** **moderate.**
* 20% of the population identifies as **politically** **liberal.**
* 20% of the population identifies as **politically** **conservative.**

**Claim 2:**

**Southerners and Midwesterners have a different ideological makeup than Northeasterners and Westerners.**

**Claim 3:**

**Race and/or gender may be associated with political conservativism.**

To test this claim, test one of the following hypotheses:

* Men are more likely than women to identify with the conservativism.
* Individuals who identify as White are more likely than others to also identify with the liberalism.

Regardless of whether there is a statistical effect, interpret the corresponding estimated odds ratio associated with this claim. Report the confidence interval and determine whether the confidence interval matches the conclusion found in your hypothesis test.

Import the Data into R and for each hypothesis you should:

* Set the values of “No Answer”, “Don’t Know” , or “Refused” to missing when necessary.
* Check whether the assumptions are met in order to perform our procedure.
* Obtain a contingency table of the data, and perform the appropriate chi-square test.
* If a test results in rejection of independence, explain the dependence.
* Construct appropriate graphs (pie / bar charts), and be sure to separate based on group.
* Where necessary, create new variables and label them appropriately.

**Objective:**

We wish to conduct a formal report for the research firm *PolitiTact*, properly analyzing each of the three claims. You will do so by first determining which statistical procedure you should use for each of the three situations. Note that you may use the assumption that the variable is *categorical/multinomial* (but is up to you to demonstrate other assumptions).

It is understood that the firm has a fair understanding of introductory-level statistics, but nothing beyond the scope of the basics taught in an intro course. Thus, it will be up to you to provide a formal report that also describes the methodology in a way that is consumable to the client (so, some colloquial descriptions are acceptable, as long as formal definitions are provided in an appendix). Furthermore, the goal is to provide a substantive narrative for the beekeeper to follow and effectively establishes a narrative with various title headers: Introduction, Statistical Background & Terminology, Methodology, Results, Conclusion.

**Formatting of the Paper:**

* All papers must be roughly 4-6 double-spaced typed pages with 1-inch margins. Ideally, it is desired to be written in Times New Roman 12-point font. A *little* under or a little over is okay, but be cautious about not writing enough (for example, a solid conclusion with its corresponding discussion *should* amount to a page on its own).
  + Use headers to describe each of the following sections: *Introduction, Statistical Background & Terminology, Methodology, Results, Conclusion.*
  + Be sure to provide motivation and appropriate transitions between sections, as well as to establish a narrative that aligns with both the formality as well as the context of the researcher’s statistical understanding. Reminder – it is up to you to narrate the process of your procedure, as so she can fully understand and replicate it, if necessary.
* For the ***Introduction*** section:
  + You should provide a brief overview of scenario as well as the situation of data collection and values for which we are testing.
    - Avoid determining whether we are testing a *mean* or *median* until later when you test your assumptions.
* For the ***Statistical Background and Terminology*** section:
  + Discuss the importance and use of statistical terminology, and make it flow as to provide a narrative flow chart of what the researcher should know as she begin to analyze the results.
  + Be sure to use the following terms:
    - Qualitative Variable, Contingency Tables, hypothesis testing with multiple categories – the parametric tests we can use (along with assumptions),  
      odds ratio.
* For the ***Methodology*** section:
  + Determine which statistical procedure you should be using, and provide both appropriate graphs as well as contingency tables to determine which procedure you need to produce. Produce your hypotheses in narrative form.
  + Use this section to determine which test you will need to use for each of the three forms, but hold off on your results until the next section.
* For the ***Results*** section:
  + Now, with the given assumptions provided and knowledge of the procedure you will perform, provide a narrative form for each hypothesis test and the accompanying confidence intervals respective to the procedure needed.
* For the ***Conclusion*** section:
  + Briefly summarize your paper and provide an overall conclusion for the casual reader to read and obtain an understanding of your work.
  + A minimum of two paragraphs will be necessary to properly summarize your paper and results
* All papers must include a *title page* that includes a report title, your name, instructor’s name, course name, and date.
* Provide an *Appendix* with the following items:
  + The complete code for which you used to produce your results
  + A glossary with full technical definitions of the following terms and formulas:
    - Qualitative Variable, Contingency Tables, Chi Square Tests (Goodness of fit, homogeneity, independence), odds ratio.
* **IMPORTANT:** Aside from the glossary portion of the appendix, **no portion of the main paper should be in bulleted or list form**. As a reminder, this paper is to be a full report and to represent a complete narrative. Inclusion of bullets or lists instead of paragraph form will result in a reduction of the letter grade (see the Readability and Formatting section of the rubric).

**Grading**

Grading for Paper 2 will comprise of three components:

1. 5 Points for a complete version of the initial draft.
2. 5 points for attendance of one-on-one conferencing over the initial draft.
3. 50 points for the final draft.

This breakdown provides a total of **50 points**.

Sample Rubric for Paper 2, STA 2183W

Organization and Motivation: /10  
(Motivate, introduce and discuss all relevant statistical concepts (and terms) appropriately)

10 -> Full discussion of the motivation and conclusion behind each procedure.  
8 -> Notable errors that deter organization and understanding of procedures, but generally well-written.  
6 -> Difficulty in understanding the procedures and the reasons behind the benefits and/or flaws.  
4 -> Clear gaps in understanding the purpose behind the procedures and purpose of the report.   
2 -> Insufficient discussion.  
0 -> No discussion.

Readability and Formatting: /10  
(Write well. Make transitions smooth and natural. Formulate a narrative that clearly expresses the process.)

10 -> Clear writing with minimal errors.  
8 -> Notable errors that deter readability, but generally well-written.  
6 -> Difficulty in reading, notable gaps in logic and explanation.  
4 -> Bullets with lack of paragraphs, clear gaps in logic and flow.   
2 -> Insufficient writing.  
0 -> No writing.

Appendix and SAS/R Results: /15  
(Use SAS or R to do all calculations and copy those to your report. Put definition, interpretation and important properties of expectation of random variables in an appendix. This part should sufficiently help the researcher to understand the expectations you use to compare the three methods.)

15 -> Clear structuring of the Appendix & SAS/R results, aiding in the numerical comprehension.   
12 -> Minor errors in computation or definitions that do not align with the paper.  
9 -> Significant errors in the results and/or the information found in the Appendix.  
6 -> Significant and incomplete data with incorrect explanations of the terminology in both sections.  
3 -> Insufficient information.  
0 -> No writing.

Statistical Terminology and Results: /15  
(In particular, you must describe (or define) and explain Qualitative Variable, Contingency Tables, Chi Square Tests (Goodness-of-fit, homogeneity, independence), odds ratio within the context of the paper.)

15 -> Clear and correct explanations of the statistical terminology and results.   
12 -> Slight errors in explanation or interpretation, but a general understanding of each procedure.  
9 -> Significantly incorrect results, notable gaps in logic and explanation and interpretation of results.  
6 -> Significantly incorrect results with incorrect explanations of the terminology in the paper.   
3 -> Insufficient results.  
0 -> No writing or results.

Total: 50 Points