Stat 313: Project 2

Write-Up Guidelines

You will be summarizing your results in a (guided) written report following the **Project Writing Guidelines** posted on Canvas. The results **must** be written in the RMarkdown template provided. You can include all pertinent plots inline in the typed document, rather than in the appendix. Your group will be submitting both the RMarkdown file used to generate your report and the knitted HTML file to Canvas by **Sunday, November 22 at 11:59pm**.

Begin by downloading the RMarkdown and data files from Canvas and saving them on your computer in the **same** folder. Then open the RMarkdown file in RStudio and edit the first chunk of code reading the data in. Once you've modified the code to read in the data, click "Knit." If you encounter an error, make sure that the data are located in the same folder as the RMarkdown file!

Data Background

Tager, Weiss, Rosner, and Speizer (1979), and Tager, Weiss, Muno, Rosner, and Speizer (1983) reported analyses of a study aimed at assessing children's pulmonary function in the absence or presence of smoking cigarettes, as well as exposure to passive smoke from at least one parent. These papers represent some of the earliest attempts at systematic documentation regarding obvious signs of reduced pulmonary function from smoking and from exposure to second-hand smoke.

Rosner (1999) presented another investigation of subjects' respiratory function, and its relationship with smoking. The data are "cross sectional," measuring participants at one point in time, rather than following them throughout their adolescence. During the study, researchers measured participant's height (in cm) and their forced expiratory volume (FEV; in liters), which is, essentially, the amount of air an individual can exhale in the first second of a forceful breath. Following these measurements, participants completed a survey about:

- their age (coded as the stage of development)
- their sex (male, female)
- whether they smoked (non-smoker, smoker)

Our primary question of interest is, controlling for age, whether the pulmonary function of smokers differs from non-smokers.

Introduction (5 pts)

- (2 pts) Give a brief background of the research problem and how the data were collected.
- (3 pts) Clearly outline the question(s) of interest you will address with the statistical analysis. The more specific you define the question of interest here, the easier the rest of the analysis and report will be. The research questions should be as specific as possible! Your *Summary of Statistical Findings* should directly answer the question you pose here.

Statistical Methods (20 pts)

This section should lay out the steps, decisions, and logic leading to the statistical model you will use to answer the research question of interest.

- (5 pt) Describe the response and explanatory variables.
  - Provide a table summarizing the number of observations for every combination of smoking and sex (i.e. non-smoker & female, non-smoker & male, smoker & female, smoker & male).
  - Provide a summary of the ages for male and female participants.

- (3 pts) Provide a data visualization of the relationship between FEV and smoking status.
  - Discuss the relationship seen in the plot, making direct reference to the plot.
  - Additionally, discuss whether this relationship is expected **and** variables that might be confounding this relationship.

- (3 pts) Provide an additional data visualization of the relationship between FEV and smoking status, accounting for the confounding variable you outlined above.
  - **Note:** Keep in mind the different methods for adding a second categorical explanatory to your plot (colors or facets)!

- (5 pts) Describe the appropriate statistical method you will use to answer the question of interest that you stated previously.
  - Be specific about why the method being used are appropriate for the investigation at hand (e.g. types of variables).
  - Keep in mind that you should include the confounding variable you outlined above in your statistical model!
  - Also keep in mind that, like multiple linear regression, there are two different ways we can include two categorical variables in our model.

- (4 pts) Check all model conditions of the statistical method you used.
  - Describe what each condition is **in the context of these data**.
  - Reference and include appropriate plots necessary for checking the model conditions.
  - Justify your conclusions regarding the conditions!

Summary of Statistical Findings (10 pts)

In this section you will write up your findings for the question of interest.

- (5 pts) What is your conclusion for the questions of interest? Namely, “After controlling for age, does the pulmonary function of smokers differ from non-smokers?”.
  - Base your conclusion on the visualizations you created and the two-way ANOVA model you fit to these data.
  - When communicating the results of the statistical test make certain you include the value of the test statistic and the distribution it follows under the null hypothesis.
  - Arrive on a conclusion based on the strength of evidence provided by the p-value.
  - If you make a rejection decision, you **must** specify an α !
  - Do not use the word “significant”!

Example for an ANOVA: There is strong evidence that at least one credit rating has a different mean income (p-value = 0.027, F-stat = 3.067 on an F(3, 1616) distribution).

- (5 pts) Interpret **in the context of the data** the p-value associated with `smoke` in your ANOVA table.
  - Keep in mind what the p-value is *conditional* on!

Survey Design (5 pts)

The survey associated with these data specifically asked the study participants if they were smokers.

- What issues do you see with this type of survey question?
- How would these issues be reflected in the data from this study?

Scope of Inference (10 pts)

Write a brief Scope of Inference statement. Specifically, answer these two questions and comment on their implications:

- (2 pts) Were the observations randomly selected from some larger population? Based on the sampling method used, what larger population can you infer the results to?

- (2 pts) Was the explanatory variable randomly assigned to observations? Based on the study design, are cause-and-effect statements justified?

Make sure you write the scope of inference **in the context of these data**, not just generic statements!

Project Presentation (3 pts)

- (1 pts) Your report should not have any spelling errors! To check for spelling errors in RStudio, click the green check mark button next to the “Knit” button.

- (2 pts) Your report should look as neat and professional as possible. Make sure that your figures don’t end up in the middle of your paragraphs, and that your sections have headings.

  - Your plots should be included in the section in which they are discussed!
  - The output from your statistical model should look as nice as possible. Output from `summary()` looks terrible in a statistical report! I would highly recommend using the `tidy()` function from the **broom** package to tidy up your ANOVA table.

Group Evaluation (3 pts)

Each member of your group will fill out a group evaluation form detailing each member's contributions, cooperation, communication, and participation.

- It is not expected that every group member is an expert on these topics.
- Rather, it is expected that every group member articulates what they are and are not comfortable contributing.

- Every member of the group can (and should) contribute to proof reading your final report!