The second phase of the project is the actual analysis and write-up.

You will answer the research question you have developed using the dataset you chose as part of the first peer assessment, use exploratory and inferential methods and techniques we learn in this class to answer your research question, and summarize your findings into a report. Your goal is to submit a completely reproducible project that conveys that you have mastered statistical inference techniques that we have learned in class and that help you answer your research question.

Your project should be a write-up of Parts 1 - 5 below in the form of a research paper. Your proposal should be written using R Markdown, so that all R code, output, and plots will be automatically included in your write up. The R Markdown template and instructions can be found on the project page.

Your write-up should be **at most seven pages** (including figures and R code/output, excluding references and the appendix). This is not very long, so you will need to be concise. Every sentence should add meaning to your paper.

Submit one file which contains your data analysis project. The write-up should include the answers to all of the questions listed below, but should be formatted as a cohesive write-up instead of as bullet point list of questions and answers. While there are boxes beneath each of the specific evaluation criteria, you should not enter your answers in these boxes, but instead submit your write-up in one document as an attachment.

To submit the assignment:

- Knit the RMarkdown file to HTML. This will create an HTML file in your working directory.
- Use the Attach a File button under the submission text box to upload this HTML file to Coursera.

Please make sure you have addressed all of the evaluation criteria listed below.

Please upload a single HTML file containing your project. Please address all of the evaluation criteria described below; review them carefully before submitting your project.

When you submit your project, you may receive an error message: "You are attempting to submit a partially-completed assignment". **Ignore** this message and submit a single HTML file with all of your responses.

Project file (HTML): Project (https://s3.amazonaws.com/coursera-uploads/user-5d0a93ca40b205ac1127e6ba/972358/asst-4/0ef67d20570711e4b052c132e27a545c.html)

Part 1: Introduction (5 points)
What is your research question? Why do you care? Why should others care?

Reminder: **please upload a single file containing all of your answers using the "Attach a file" button above this question.**

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Is there a well-defined and clearly stated research question?

▾

Did the author provide background on the research question as to why they care and why others should also care?

▾

Part 2: Data (10 points)

Write about the data from your proposal in text (not bullet-point) form. Address these points:

- *Data collection:* Describe how the data were collected.
- *Cases:* What are the cases? (Remember: case = units of observation or units of experiment)
- *Variables:* What are the two variables you will be studying? State the type of each variable.
- *Study:* What is the type of study? Is it an observational study or an experiment? Explain how you've arrived at your conclusion using information on the sampling and/or experimental design.
- *Scope of inference - generalizability:* Identify the population of interest, and whether the findings from this analysis can be generalized to that population, or, if not, a subsection of that population. Explain why or why not. Also discuss any potential sources of bias that might prevent generalizability.
- *Scope of inference - causality:* Can these data be used to establish causal links between the variables of interest? Explain why or why not.

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Data collection: Is the data collection explained clearly?

▾

Cases: Are the cases (the units of observation or experimental units) explained clearly?

▾

Variables: Are the variable types identified accurately?

▾

Study: Is the type of study identified correctly? Is the supporting information on the sampling and/or experimental design of the study satisfactory for making the decision on the type of study?

Scope of inference - generalizability: Did the writer correctly identify the population of interest? Did the writer correctly decide whether the findings from this analysis can be generalized to that population, or, if not, a subsection of that population? Is their explanation satisfactory to make this decision? Are potential sources of bias discussed, and if so, is the discussion satisfactory?

Scope of inference - causality: Did the writer identify correctly whether these data can be used to establish causal links between the variables of interest. Is the explanation satisfactory?

Part 3: Exploratory data analysis (10 points)
Calculate and discuss relevant descriptive statistics, including summary statistics and visualizations of the data. Also address what the exploratory data analysis suggests about your research question.

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Are appropriate summary statistics calculated, and are they explained/interpreted in context of the data and the research question?

Are appropriate visualizations included, and are they explained/interpreted in context of the data and the research question?

Did the writer address what the findings from the exploratory analysis suggests about the research question?

Part 4: Inference (20 points)

- State hypotheses
- Check conditions
- State the method(s) to be used and why and how
- Perform inference
- Interpret results
- If applicable, state whether results from various methods agree

It is your responsibility to figure out the appropriate methodology. What techniques you use to conduct inference will depend on the type of data you're using, and your sample size. All of you should conduct at least a hypothesis test, and report the associated p-value and the conclusion. Those of you comparing two means, two medians, or two proportions should also calculate a confidence interval for the parameter of interest. Those of you working with categorical variables with more than two levels will need to use methods like ANOVA and chi-square testing for which there is no associated confidence interval, and that's ok. If your data fails some conditions and you can't use a theoretical method, then you should use an appropriate simulation based method.

- If you **can** use both theoretical and simulation based methods, then choose one and stick with it. You don't have to do both. However if you **can't** use both, then you need to decide which is appropriate.
- If you **can** do both a hypothesis test and a confidence interval, do both, and comment on agreement of the results from the two methods. However if your variables do not lend themselves to a confidence interval, that's ok.
- It's essential to make sure the method you're using is appropriate for the dataset and the research question you're working with.

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Are the hypothesis is stated clearly and matches the research question?

Are the conditions checked in context of the data (not just a generic bullet point list of the conditions, but reasoning through them for the given dataset)?

Are the appropriate method(s) the writer will be using stated? Did the author provide a discussion of why they chose these methods, and described how they work? Note that in this part the author should display a thorough and conceptual understanding of how the methodology works, however the write-up does not need to be as detailed as if they were teaching the method to someone with no background in statistics.

Was the correct code used and output provided for all required techniques?

See below for which situation requires which technique:
- One numerical and one categorical variable (with only 2 levels): hypothesis test + confidence interval

- parameter of interest = difference between two means (theoretical or simulation)
- parameter of interest = difference between two medians (simulation only)

- One numerical and one categorical variable (with more than 2 levels): hypothesis test only

- compare means across several groups
- no defined parameter of interest, ANOVA and pairwise tests (theoretical only)

- Two categorical variables (each with only 2 levels): hypothesis test + confidence interval

- parameter of interest = difference between two proportions (theoretical if success-failure condition met, simulation if not)

- Two categorical variables (either one or both with more than 2 levels): hypothesis test only

- compare proportions across several groups
- no defined parameter of interest, Chi-square test only (theoretical if expected sample size condition met, simulation if not)

Are correct interpretations and conclusions for all output provided? Note that this portion should be evaluated based on criteria stated earlier about which technique is required when). This includes some or all of

- conclusions of hypothesis tests,
- interpretations of p-values as conditional probabilities, and
- interpretations of confidence intervals

depending on the methods used. All interpretations must be in context of the data and the research question.

Is whether or not results from hypothesis test and confidence interval agree stated? Or, if doing ANOVA or chi-square testing, did the author state that no other methods were applicable and hence there's nothing to compare?

Part 5: Conclusion (5 points)
Write a brief summary of your findings without repeating your statements from earlier. Include a discussion of what you have learned about your research question and the data you collected, and include ideas for possible future research.

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Is there a brief summary of findings that does not repeat previous statements?

Is a discussion of what was learned about the research question provided?

Are ideas for possible future research and/or discussion of additional synthesis or possible shortcomings of study provided?

References (1 point)
Include a citation for your data, and if your data set is online, provide a link to the source. Also list other references (if any).

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Are references included?

Appendix: (1 point)
Take 1 page of your data set and attach it to your proposal. If your data fits in one page, great. If you have too many observations and it won't fit, that's ok too, just crop it at one page. However your print out should contain all relevant columns (this shouldn't be an issue since you are working with one or two variables for your project).

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase.

Is there an appendix containing one page of data, and does it display the variables used in the project?

Overall evaluation/feedback

Note: this section can only be filled out during the evaluation phase.

Is the writing clear, with few or no grammar, spelling or organization mistakes?

Please make any general constructive comments on this project that the student would find helpful.

You've written 0 words

Analysis was completed using R (and not other software/programming language).

Write-up was created using RMarkdown (and not other word processing software).

▼