Final Project (Stage 2)

Stat140-04

- Final paper due on Gradescope by 6PM EST on Friday (12/11/20)
- Presentation due in class on Zoom on Thursday (12/10/20) and Friday (12/11/20)
- You can choose to either work on your own, or in pairs. If working in pairs, only one group member is supposed to submit the assignment, and tag the other student (do not all submit separately, or on the flip side forget to tag your teammates if you are the group's designated submitter). If you do not do this, you can submit a request and we will fix it, but we will also deduct 1 point.
- Link to grading rubrics and selected dataset from previous stage:

https://docs.google.com/spreadsheets/d/19lYI0NKS7iFpJAWN4hAkp2T3m5LZRF4llhv_101J_n0/edit?usp=sharing

1. Final paper

Choose a dataset from the list. Write a paper addressing one or two variables (variables can be categorical or quantitative) for which you want to do statistical inference. Generate a confidence interval and carry out a hypothesis test. Your paper will be comprised of the following parts:

PART 1: Data

Describe the data, the variable(s) being used, and define the parameter you want to estimate. Write out clearly your sample statistic.

If any of the variables are quantitative, please include the mean, standard deviation, and 5-number summary of the variable. If the variable(s) are all categorical, please summarize the count of data with a table.

PART 2: Confidence interval

Step 1: Bootstrap distribution

Generate a bootstrap distribution with at least 5000 simulations. You can use R or StatKey. To use StatKey, you can enter your own data by clicking on "Edit Data" and copying and pasting your data in from a spreadsheet. In your submission, please include a plot (a screenshot is fine) of the bootstrap distribution.

Step 2: Confidence intervals

Generate a 95% confidence interval

a. using the standard error from the bootstrap distribution.
b. using the percentile method, the middle 95% of bootstrap statistics. Give the interval, and include a plot (a screenshot is fine) of this.
c. using the CLT (normal or t-distribution) based methods and formulas (show your work and make sure you have checked all the conditions)

Step 3: Interpretation and discussion

Interpret the intervals you constructed in Step 2 in context. How do these intervals compare? (Hint: if you did everything right, and if the conditions are met for using the CLT based methods, the answers should be very close).

PART 3: Hypothesis testing

Step 1: Hypothesis

State the null and alternative hypotheses in words and mathematical terms.

Step 2: Testing

Do the hypothesis testing in R, and find the p-value.

Step 3: Make a decision and interpretation

Interpret the p-value in context. What can you conclude? Make sure you interpret the results with respect to: 1) significance; 2) rejection of null hypothesis; 3) real world conclusion about subject being studied

PART 4: Conclusion

Tie together the findings of your analysis in Part 2 and 3. Were there shortcomings in your data that prevented you from fully answering your research question. If so, how could future studies fix this problem? How might the results of your analysis be used to motivate or inform future research?

Instructions

You can write your final paper in any word processors (Rmarkdown, Microsoft Word, Apple Pages, LaTex), but you must submit your paper in .pdf format on gradescope. Your paper should include all relevant figures and R outputs. R code and the data should be uploaded separately as an appendix at the end of the paper. We should be able to simply run your R Script to reproduce your results. If we cannot easily reproduce your results from your script you will be penalized.

Write as if you are explaining your results to whoever would be interested in your research question, whether this is other scholars in your field or peers sharing your interest in the topic. Keep in mind this audience may or may not have taken statistics. You may assume familiarity with basic descriptive statistics and visualizations. You must be statistically accurate and use correct statistical terminology, but must also explain your conclusions in a way that anyone can understand.

2. Poster Presentation

You should put together a short PowerPoint-like presentation of your analysis to be presented in class. Keep you presentation under 5 minutes. If you are working with a partner, you will be presenting together for 10 minutes, and each of you should speak for roughly half of the time. You will be asked to stop once the time limit is reached. The presentation should include meaningful visualizations, text, video, and/or other relevant multimedia content. It is up to your discretion as to what kind of material you would like to put in the presentation, but the analytic process, findings, and conclusion should be clear. In general, the content in the presentation should be a condensed version of the written report.

Policy

You may discuss the project amongst yourselves and with me, but all work (e.g., coding, writing, statistical analysis) must be done entirely on your own. Plagiarism will not be permitted – do not copy other people's projects and code. Failure to abide by this policy will result in a 0 for everyone involved, and will be treated as a case of academic misconduct.

If you would like to something other than what is outlined here (e.g., logistic regression or multiple regression), this is completely OK but please talk to me first before you proceed.

For every additional day the final paper is late, 10% of the project grade will be deducted.