STAT 113: TERM PROJECT

Description

For the term project you will work in a group of 2 or 3 to formulate a research question which is of interest to you, and collect and analyze a set of data that is informative for your question.

Forming a Question. You will need to first formulate a question (or set of related questions), and write down a set of hypotheses that you can collect data to provide evidence for (or against). The hypotheses should involve a relationship between two variables in a clearly identified population. These relationship can be in the form of a comparison between two or more groups or an association between quantitative variables. If you think of a question that does not fall into one of these categories, talk to me; it may work.

Examples of project questions that have been asked in the past: Is sodium content higher in canned than in frozen vegetables? Is there a relationship between fat content and protein content in fast food menu items? Does the weight of an acorn depend on whether it is from an isolated tree or a tree within a large cluster? Are more red/yellow cards given per game in World Cup soccer games than in other matches? Is there a relationship between outside temperature and students wearing brightly-colored clothing? You should come up with an *original* question. It may range from very simple (e.g., a comparison between two groups or a simple correlation analysis) to more complex (for example, how is the relationship between two variables influenced by one or more others?). When grading I will factor in the ambitiousness of your question and be more lenient on the technical details for more challenging projects.

Data-Collection. You will need to carefully plan how you will collect your data to ensure that the sample you collect is as representative as possible of the population that are interested in saying something about. Be clear up front about what that population is! How will you take a random sample of this population? And so on. It is likely not possible to collect a perfectly random sample, but think hard about

Date: Last Revised October 10, 2017.

what sources of bias might be present in your sampling procedure, and try hard to eliminate them.

You may study almost anything; however, **your project may not simply involve a survey of other Oberlin students**. You might study Obies, but not by asking them to fill out a survey (you may interview students in person, but you will need a clear plan for how you will recruit them, keeping in mind any potential sources of sampling bias that may be present).

Deliverables

Milestone 1: Project Proposal. The first step is to prepare a brief (no more than one or maybe two pages) plan in writing, in which you: describe your question; define your research hypotheses; describe how your data informs your hypotheses; identify what your intended population is, what your cases are, and what variables you will need to collect data for; describe your sampling procedure (how will you select cases from the population, and how many cases you expect to need to have a decent chance of having a statistically significant result?). Finally, you should include a hypothetical data table with a few rows of made up data, that have the same structure your actual dataset will have.

There is a folder in King 203 containing some examples of high quality completed projects from previous semesters of STAT 113 at Oberlin, which you should look at. The project specifications will not have been identical, but they should be close.

Deliverable: Proposal Writeup. Deadline Friday 10/27 by start of class

Milestone 2: Preliminary Data-Collection. Before you collect your full dataset, you will need to collect a small "pilot" dataset. This is a data set that contains all the variables you plan to collect for your full data, but fewer cases. The point of pilot data is to do a "sanity check" on your hypothesis, adjust your data collection procedure, and possibly suggest hypotheses that you didn't consider originally.

Milestone 3: Analysis of Pilot Data. After collecting your pilot data, you should describe it (qualitatively and quantitatively), visualize it (using graphs that illustrate the properties of the data that are most relevant to your question) and analyze it using the simulation-based estimation and testing methods (i.e., bootstrap confidence intervals and randomization tests) that you will have learned by that point. You should then identify any changes you plan to make to your data-collection procedure and any new hypotheses that may be suggested by the pilot data.

The writeup of the pilot analysis should have the same structure as the writeup of the final paper (see below). Some of your text (for example the introduction) and code (e.g., for descriptive statistics and visualizations) can be reused (with edits responding to feedback) in the final paper, though the inference methods and points of discussion will be different.

Deliverable: Pilot Analysis. Deadline Friday 11/10 by start of class

Milestone 4: Full Data Collection. Once you have made the necessary changes to your data collection procedure and hypotheses that were suggested in the pilot phase, you will need to collect a second, larger dataset that is equipped to address your revised question(s). You should *not* use data collected in the pilot phase in your final dataset; for one because your sampling procedure and variables collected are likely to change, but also because you will have already analyzed the pilot data and used it to formulate new hypotheses, and including it in the final analysis would inflate the risk of a "false positive" (Type I Error) finding. You should, however, use the pilot data to estimate the effect size of your central relationship, and use this effect size to determine the sample size needed in your final dataset. Your final sample size must also be large enough to satisfy the conditions for the analytic inference approximations that you will be using in your final analysis.

Milestone 5: Full Data Analysis. After collecting your full data set you will repeat the descriptive summary and visualizations that you did for your pilot analysis (perhaps with modifications), and carry out the estimation and testing process but this time using analytic distribution-based approximations instead of simulation-based methods.

Milestone 6: In-Class Presentation. The final presentations for the group projects will occur during finals week. Your final analyses should be finished, but this can serve as an opportunity to take feedback from your peers, and from me, into account and do some last minute revisions to your writeup. Presentations will be approximately 5 minutes each with 2 additional minutes for questions, and should contain the main elements of the paper. You should assemble slides with relevant visualizations (graphs, etc.) that support what you are saying. Each team member must deliver a part of the presentation.

Deliverable: In-Class Presentation. Friday 12/15 (scheduled final exam day)

Final Milestone: Finished Paper. The final product is a written report, structured like a research paper, and prepared as a reproducible report using RMarkdown. See below for details on how the paper should be structured. There is no formal length requirement, but I would expect most writeups to be around 5-7 pages, including figures but not including code.

Deliverable: Final Paper. Friday 12/15 (scheduled final exam day)

Honor Code

The central principle of the honor code as it applies here is that your work is original, and any outside sources that are used must be cited. Since this is a group project, the honor pledge also affirms that every member of the group contributed at every stage of the process (the proposal writeup, data collection, preparation of visualizations, and both sets of inferential analysis).

Content of the Final Written Report

Your analysis should be in the form of a report tailored to a reader who knows something about statistics, but is not a statistics expert; basically, your peers in this class. It should be in the form of an RMarkdown document, with all plots created from code embedded in code chunks. You should use the `echo=FALSE` option in code chunks so the code is *not* displayed in the final report (that is, so that the report looks like a research paper). You will need to describe your methodology in enough detail that it is clear what you did without seeing the actual code.

The Markdown document must be self-contained so that it can be run successfully, including the data import step, from a fresh RStudio session with nothing in the workspace. You should turn in three files: the `.Rmd` source, a PDF output file (i.e., follow the same procedure we use for labs that have required Markdown), and your dataset as a plain text `.csv` file (you may assemble the data in Excel or other spreadsheet software, but please export it to `.csv` before submitting).

Your pilot analysis and final report should each contain the following elements.

(1) Begin with an introduction: State what you studied (clearly define your response and explanatory variables), and why you studied it. State what the cases are (e.g., individual persons, dorms, plots of land) on which you have made your measurements.
(2) Write a methods section, in which you do the following:

- Identify the population from which your sample was drawn and to which you are drawing inferences. (E.g., complete the sentence "My population of interest is...")
- Discuss how you collected your data (if applicable), particularly how you protected against sources of bias. Describe any data collection problems and what you did about them.

(3) Write a results section, in which you describe and visualize your data, and report the inferential results.
- State all null and alternative hypotheses in both words and symbols.
- Investigate the hypotheses graphically and comment on what the graphs say about the hypotheses. For example, if you are comparing averages then you may want to provide side-by-side boxplots, histograms, dot plots, and/or density plots. If you are comparing proportions, provide a suitable bar chart and/or a table of conditional proportions. If you are fitting a regression model, provide a scatterplot with the best fit line depicted.
- Provide summary (descriptive) statistics that are informative about your research question.
- Discuss the conditions needed for your intended analytic inference procedures to be valid, and (for the final paper) provide evidence that those conditions are met. In the case of a regression model, this will mean that you provide one or more residual plots; in the case of a t-test, you will need to check for Normality; etc. If you decided a transformation of one or more variables is needed, describe how you came to that conclusion (with reference to your graphs and/or descriptive statistics), and how you chose what transformation to use (preferably with some conceptual justification from the context of the data).
- Provide confidence intervals (CIs) for the main population parameter(s) of interest, such as group population means, differences of population means, population correlation coefficients, regression slopes, etc. Interpret the CIs in context, by completing the sentence "We are 95% confident that..." (assuming you chose 95% as your confidence level). Again, for the pilot analysis, you will use simulation methods (bootstrapping); for the final writeup you will use analytic (e.g., Normal or t-distribution-based) methods.
- Conduct the formal hypothesis test that gives numerical support to the visual evidence from the graph(s), reporting a P-value. For the pilot analysis, you will use simulation methods (randomization); for the final writeup you will use analytic (e.g., Normal or t-distribution-based) methods.

- Interpret the P-value for the test as a probability, in context, by completing the sentence "If H_0 were true, the probability...."
- Calculate and interpret the effect size (e.g., Cohen's d or R^2) in the data and write a sentence that explains, in layperson terms, the statistical and practical significance, if any, that you have found. For a regression study, give the correlation coefficient and R^2 and comment on these.
- If there are more CIs or tests to be completed, then add them here. For example, if you conduct an analysis of variance then you will want to construct several CIs to compare means.
- If there are extreme outliers in your data, then discuss how you handled them. Do not discard data points lightly — if there appear to be outliers you should strongly prefer to use robust methods and/or data transformations that will mitigate the influence of outliers without having to ignore them completely. If you do remove any data points from your analysis, write one or more sentences explaining in detail why that point needed to be removed. Saying "it was an outlier" is not sufficient; you must explain why that point should be considered to not have come from the population that you are studying. In this case, you should report your results both with and without the outlier, and comment on any differences you find.

(4) Include a closing discussion section in which you describe the results in natural language, and interpret them in the context of the research question. You should discuss what you learned, how strong the evidence was for or against your research hypothesis, how the sampling went, and to what extent you feel comfortable drawing conclusions to your entire population based on your sample and your analysis of your data. Were there any surprises? If the conclusions were ambiguous, or you identified possible flaws in the data collection procedure, discuss what you might do differently if you were to conduct a second study on the same (or a similar) question. In the case of the pilot analysis, this is where you will discuss planned changes in data collection and/or hypotheses for the final product.

(5) Submit your raw data and your written report (as both .Rmd and .pdf) to Blackboard. Having downloaded your .Rmd and your data, I should be able to Knit the Markdown file and reproduce your report.

Summary of Deadlines

Deliverable	**Deadline**
1. Question and Plan	Friday 10/27
2. Pilot analysis	Friday 10/10
3. Presentations	Friday 12/15
4. Final writeup	Friday 12/15