STAT 113: TERM PROJECT

COLIN REIMER DAWSON, FALL 2020

Description

For the term project you will work in a group of 2 or 3 to

(1) **Formulate a research question** which is of interest to you

(2) **Find a set of data** that is informative for your question

(3) **Summarize and visualize the data** using descriptive techniques from this class

(4) **Estimate parameters** of interest using confidence intervals

(5) **Test hypotheses** of interest

(6) **Write up your analysis** in the form of a research paper with an introduction section, data (methods) section, results section, and discussion section

Forming Groups

You may work with any **one or two other students** (for a total group size of 2 or 3) who are enrolled in either of the sections of STAT 113 that I am teaching (i.e., the 9am or 10am sections).

It is fine to have groups that contain a mix of people from the two sections, since both have identical requirements; however, since the two sections have different final exam schedules, if your group does contain a mix of people, **the due date will be the earlier of the two**.

If you prefer to be put into a group rather than picking your own, **let me know as soon as possible**, and I will create random groups.

Date: Last Revised November 12, 2020.

Finding Data

It is probably a good idea to peruse some code books of potential datasets before settling on a research question, so you can ensure that you will be able to address the question in a somewhat satisfying way.

You should either find a dataset which is available in a “data frame” format, or if there is data you want to use which is not in this form, you can create a data frame yourself (e.g., in Excel or a Google Sheet).

You can get your data wherever you want, but I suggest looking in the **FiveThirtyEight** data repository on GitHub (`http://github.com/fivethirtyeight/data`) for inspiration and potential datasets of high quality on a variety of topics (and also formatted in a way conducive to R, since most FiveThirtyEight articles are prepared at least in part using R). As you might expect there are a lot of political datasets there, but there are lots of other things too.

At the link you will see a long list of folders, each containing data associated with some article published on FiveThirtyEight at some point. You will need to click on a folder to see the code book associated with that dataset, which will usually also include a link to the article for which that data was used. You can also usually peek at the contents of a `.csv` file using GitHub’s data viewer by clicking on the file name.

Some datasets there that have a good mix of categorical and quantitative variables include `bechdel`, `comma-survey`, `san-andreas`, `thanksgiving-2015`, `weather-check`, but there are a lot of others as well.

Forming a Question

Your research question should involve a relationship (or potential relationship) between an **explanatory variable** and a **response variable**.

The response variable can be categorical or quantitative. The explanatory variable can be categorical, in which case you are making a comparison between groups or conditions, or quantitative. However, **we do not have tools** from this class for investigating questions involving a **quantitative explanatory variable and a categorical response variables**.

Note: You may not duplicate an analysis done on FiveThirtyEight; however, there are plenty of other questions that can be asked using the same data that was used for an article there.

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.

Clearing Your Topic With Me

You don't need to turn in a formal written project proposal, but at least one member of your group should have a **brief conversation** with me by **Monday, 11/23** at the latest so that I can sign off on your plan.

Content of the Final Written Report

Your analysis should be in the form of a report tailored to a reader who knows some basic things 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. I will provide a template.

You should set `echo=FALSE` and `results='hide'` in the initial setup chunk, which will hide the code and raw text output from the knitted 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.

What to Turn In. You should turn in three files:

1. The `.Rmd` source
2. A PDF output file (follow the same procedure we use for labs that have required Markdown, but select Knit to PDF instead of Knit to HTML)
3. Your dataset as a plain text `.csv` file (unless it is read in directly from the web)

Structure of the Writeup. Your writeup should contain the following elements:

1. **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, movies, plots of land).
2. **Data (Methods)** section:
   - Identify the **population/process/phenomenon** from which your dataset was drawn and to which you are drawing inferences.
   - Discuss **how the data was collected**, and what sorts of **sampling and measurement bias** might be present, and how they might impact the results
3. **Results** section. This should contain at least three subsections

   A. **Parameters and Hypotheses**

   - Define **parameters of interest**
   - State all **null and alternative hypotheses** in words and in terms of your parameters

   B. **Descriptive Analysis**

   - Investigate the hypotheses using **graphical summaries** appropriate to the type of explanatory and response variable
   - **Comment** on what the graphs suggest about the hypotheses
   - Provide **summary (descriptive) statistics** that are informative about your research question
   - 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 not to 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.

   C. **Inferential Analysis**

- Discuss the **conditions** needed for your inference procedures to be valid, and provide evidence that those conditions are met.
- Provide **confidence intervals** (CIs) for the main parameter(s) of interest
- **Interpret the CIs** in context, by completing the sentence "We are 95% confident that..."
- Conduct the formal **hypothesis test(s)**, reporting a P-value.
- **Interpret the P-value** for the test as a probability, in context, by completing the sentence "If there were no [difference, correlation, etc.; fill in based on your context], the probability that..."
- Calculate and interpret a suitable measure of **effect size** (such as Cohen's d or R^2) in the data and write a sentence that explains, in lay terms, the statistical and practical significance, if any, that you have found.
- 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.

4. **Discussion** section:
   - **Describe the results** in natural (non-technical) language, and **interpret them in the context** of the research question
   - Were there any **surprising results**?
   - Discuss **the main takeaways** from the analysis, and **what caveats** apply (for example, how the possibility of **confounding variables**, **sampling bias**, **measurement bias**, etc. complicate the interpretation of the results
   - Discuss **what additional data** you would like to have had and/or **what changes to the data-collection procedure** you would like to have made in order to gain additional insight into your research question and/or follow up with additional questions of interest

What to Turn In

Submit your **data** and your **written report** (as both `.Rmd` and Knitted `.pdf`) to the turnin folder at `~/stat113/turnin/project/`. I should be able to Knit the Markdown file and reproduce your report.