STAT 113: TERM PROJECT DESCRIPTION

COLIN REIMER DAWSON, FALL 2022

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) **Collect or 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** using RMarkdown, in the form of a research paper with an introduction section, data (methods) section, results section, and discussion section

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.

Project Sub-Tasks

Forming Groups. You may work with any **one or two other students** (for a total group size of 2 or 3) in the class.

If you prefer to be put into a group rather than picking your own, **let me know via Slack as soon as possible**, and I will create random groups out of those who notify me that they want this.

Date: Last Revised April 26, 2022.

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, in which case you are most likely going to be looking at correlations and/or regression models.

We do not have tools from this class for investigating questions involving a **quantitative explanatory variable and a categorical response variables**, so if your response variable is categorical, then you will be limited to a categorical explanatory variable. However, you can always convert a quantitative variable to a categorical one by binning.

Note: Although there is an option to use existing data rather than collecting your own, you may not duplicate an existing *analysis* (if using data from FiveThirtyEight, for instance, look at what they wrote about with that data so you can make sure you're doing somtehing different). There are plenty of different questions that can be asked using a given dataset.

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 (in person, on Zoom, or on Slack) by **Friday, 04/29** so that I can sign off on your plan. You should have your **question**, your **explanatory and response variables**, and your **hypotheses** identified, and should have your **data-collection plan** worked out (see details on data-collection in the next section).

Gathering Data.

Option 1: Collect Data Yourself. The "standard" option is to collect your own data by going out into the world in some fashion and recording something: from other people, from trees, grocery shelves, etc.

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 something **approximating a random sample** of this population? It is likely not possible to collect a perfectly random sample, but **think hard about what sources of bias** might be present in your sampling procedure, and try hard to **minimize** 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 if you wish, 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. You **may not ask for sensitive personal information** such as educational or medical history, grades, etc., as this would require additional review by a third party.

Option 2: Using Existing Data. If your travel plans and the shifted semester will make it logistically difficult for you to collect data with your group, you can instead 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) from another format.

If you opt for this approach, SLO E2 (which is about collecting data) will be omitted from the grading for the project, and instead the remaining "E" SLOs will be weighted more heavily. Essentially this means that the **description and analysis** of the data count for more, and because there is less to do, the technical standard will be slightly higher for that material.

You can get data from 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.

Data Collection: Pilot Phase. You should begin by collecting a small dataset using your planned methodology. This gives you a chance to revise and refine your methodology, and also to perform an initial analysis that may guide your final work and perhaps lead you to revise your hypotheses as well.

Once you have finished collecting your pilot data, your should write up an analysis **in the same format as the final written report** (see below). This "draft writeup" should look as much like the final project as possible, with the exception that it is based on a small initial dataset.

Content of the Pilot and Final Written Reports. Both the pilot writeup and the final writeup 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. You may want to take one of the lab documents as a template: delete the text paragraphs and the code chunks after the initial set-up chunk, but leave the header section and first setup chunk.

Formatting Note: Before Knitting the document to turn it in, you should set `echo=FALSE`, `message=FALSE`, `results='hide'`, and `warning=FALSE` in the chunk options in the initial setup chunk. This will hide the code, raw text output, and various messages from R from the Knitted report, while leaving plots and object definitions that you might want to refer to in text or tables.

The idea is to make the report **looks like a research paper**, without code and unformatted output of code displayed in the writeup). In other words, someone reading it should **not necessarily be able to tell by looking at it that you used RMarkdown** to write the report.

This also means you will need to describe your methodology in enough detail that it is clear what you did **without seeing the actual code**. I will ask you to turn in the `.Rmd` file alongside your writeup, but **the paper must stand on its own**: all necessary information must be in the text and figures so that someone can **read the Knitted document without seeing the code** and get all of the necessary quantitative and qualitative results.

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. If your data is being read in from a `.csv` file or similar, this file should be placed in the same folder as the `.Rmd`, and then read in using a `read_csv()` (or similar) command **within the Markdown document**.

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** (whether by you or someone else), 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

Turning In Your Work. Both the pilot and final writeup submissions should consist of **three files** (two if your data is read in directly from the web):

1. The `.Rmd` **source document**
2. **A Knitted 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. Avoid copying and pasting special characters into your `.Rmd` file: this can prevent it from Knitting. If you need mathematical notation, such as μ, use the dollar sign syntax that we've seen in various labs.
3. Your **dataset** as a plain text `.csv` file (unless it is read in directly from the web)

Copy these files to the turnin folder at `~/stat113/turnin/pilot-draft/` for the pilot phase and `~/stat113/turnin/final-project/` for the final writeup. I should be able to Knit the Markdown file from there and reproduce your report.