Stat 152 Final Project

Spring 2020

Overview

For this project you are to work in **pairs**. You must download data from a large national survey and analyze it. This project consists generally of

1. Understand the underlying design of the experiment
2. Understand what is provided in the data regarding the design and how it differs from the design
3. Pose a question(s) of the data
4. Analyze the data to answer your question(s)

Stages of the Project

I will have a few stages where I ask you to report in that you have accomplished some tasks. This is to help keep you on track and make sure the groups are working together. These do not need to be formal, though I expect complete sentences. You will submit these 'progress reports' online via **bcourses**. They will be only superficially graded (more or less whether you are showing that you have met the goals).

Stage 0 (Choose your partner) – Due Wednesday April 22 You must have signed up for a group by this point and put the names of your group members in the google sheet linked to by the assignment.

Stage 1 (Initial Proposal) – Due Monday April 27 You need to tell me what survey you have decided to analyze, some possible questions you are considering, and describe the variables in the data set that you plan to make use of.

Stage 2 (Data Acquisition)– Due Sunday May 3 You need to have downloaded the data, read it into R, and done some basic checks of the data. This includes properly coding non-responses, recoding any variables into appropriate factors with informative names, and so forth. You should submit some summary statistics of the data, acquired through R:

- The number of rows of the data
- First 5 rows of the data (only relevant variables for the analysis)
- Standard summary of variables of interest.
- One graphical display of your data of interest (does not need to make use of survey weights or design)

For this stage you should briefly describe what steps it took to get the data to this point. R code (beyond that needed for the above items) is not necessary.

Please note that the data acquisition and processing may not be trivial, so **please start this as soon as possible.**

Stage 3: Final Report – Due Wednesday, May 13 at 11 pm on BCOURSES I will post more specific guidelines about the format of the final report on bcourses. The emphasis will be on understanding the design and trying to appropriately use the design elements to analyze a question. The structure will be

- Introduction
- Description of the survey (including its design components)
- Description of the publicly available data (including what has been done to the data that makes it differ from the raw data) and what processing you did to be able to analyze it
- Description of the question you are focusing on.
- Analysis of the data, which will include both exploratory data analysis (i.e. graphical tools) and estimation/inference
- Conclusion

Stage 4 (Self-Evaluation) – Due Friday, May 15, at 11 PM on Bcourses **Each** member of the group must *individually* submit a description (on bcourses) of how the work was distributed throughout the project. You should describe specifically the organization strategy you took in the project (see suggestions below). These do not need to be formal, but should be about 2-3 paragraphs.

If you have any qualms about the process and who contributed what, this is the time to tell me. If there was no problem that's great, but you still must turn in the description.

Suggestions for Organizing Your Project

- If you are having problems working with your partners, please come and talk to me as soon as possible. **I would prefer not to have people working alone. If you want to work alone, you will have to convince me of why you cannot work with someone. Being in a different time zone is not a sufficient reason.**
- Once a dataset has been chosen and some variables of interest decided upon, there are several large chunks for this project:
    1. Getting the data and making sense of the design elements and what is made publicly available
    2. Analyzing the data in R
    3. Writing and producing a final report.
- All group members should contribute to ALL of these portions. Be very careful to not give all of the responsibility of one of these portions to one person, as that will be reflected in the grading.
- At the beginning, assign responsibility for each of these portions.
- Perhaps assign one of the members to be a "project manager". This person will send out emails reminding everyone of duties, pressure everyone to get their part done, do a final check that the required writeup is not missing any requested components, etc.

- The actual analysis of the data is where we will be closely watching for the skills you have learned in the class, so you need to all make sure you are happy with your performance here. After the initial analysis has been done, you should make sure that there is time for everyone to read the initial analysis and then meet (in person!). During this meeting talk out whether anything is missing or could be done better. Once you have arrived on a final statistical analysis, you can then reasonably split up the writing of the final paper.

Data

You are welcome to use any complex survey. National statistical agencies are one of the best sources, and there are some examples below, but you can get data from other places as well. The following are national statistical agencies where we have found the data to be reasonably straightforward to obtain and download. Here are some examples of the surveys they contain just so you can compare, but there are generally many more which may be of greater interest to you. Please make sure you choose a survey for which *all* the material is in English so that I can read it and comprehend it. Please note that you need to have **actual data**, and *not* summarized tables. The data needs to be from a complex survey (using combinations of cluster sampling and stratified sampling, generally with weights) and must measure many variables (i.e. not a single question). The data has to provide enough information to be able to actually analyze the survey correctly i.e. accounting for the survey design. You can check by seeing if they provide variables giving the primary sampling unit and sampling weights for each observation, for example, as well as per-observation sampling weights. You also need to make sure that they provide detailed information about the design. For example, national government surveys usually provide a lengthy pdf document describing the design, what has been done to deal with non-response, how the sampling weights have been adjusted, what information has been changed, and so forth. A large portion of your report will be digesting this information and summarizing it.

- Bureau of Justice Statistics (`http://www.bjs.gov/index.cfm?ty=dca`). If you see a survey of interest, you can search the studies to find the data at `http://www.icpsr.umich.edu/icpsrweb/NACJD/studies`. You must create a free account, and agree to the terms of use. Examples:
  - Annual Survey of Jails `http://www.icpsr.umich.edu/icpsrweb/NACJD/series/7`
  - National Crime Victimization Survey `http://www.icpsr.umich.edu/icpsrweb/NACJD/series/95/studies/34650?archive=NACJD&sortBy=7`
  - Annual Probation Survey and Annual Parole Survey Series
- Bureau of Transportion Statistics
  - National Household Travel Survey `https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/subject_areas/national_household_travel_survey/index.html`
- National Center for Health Statistics (`http://www.cdc.gov/nchs/surveys.htm`)
  Note the data may be distributed as a SAS transport file (`http://www.cdc.gov/nchs/nhanes/sas_viewer.htm`) which you should be able to convert.
  - NHANES `http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm`
  - Series of surveys on health care available at: `https://www.cdc.gov/nchs/dhcs/dhcs_surveys.htm`
  - Firearm Injury Surveillance Study

There are other national statistical agencies, for example

- National Center for Education Statistics (`https://nces.ed.gov/`)
- Bureau of Labor Statistics (`www.bls.gov`)
- Census Bureau (`https://www.census.gov`)
  - American Community Survey `https://www.census.gov/programs-surveys/acs/`
-
- National Survey of Childrens Health `https://www.childhealthdata.org/learn-about-the-nsch/NSCH`
- Note that ICPSR (`https://www.icpsr.umich.edu/icpsrweb/` has a huge repository of data, so you can dig around there. Make sure that you find a complex survey to use.

You may have luck finding surveys there. Please note that you need to have **data**, and not summarized tables.