Final Data Analysis Project
STAT 152
Due: 7pm Friday 15 May

Overview

For this project you are to work in groups of three. You must download data from a complex national survey and analyze it. In this project you will:

1. Describe the underlying design of the survey, including identifying the sampling unit, sampling frame at each stage of the survey, and the observation unit and target population.
2. Describe how non-response is addressed and other potential sources of bias
3. Analyze the survey data to answer a question, including calculating standard errors
4. Provide visualizations for your analysis
5. Prepare a poster of your work and present it during our class' exam period

Stages of the Project

There are several stages in this project, and you will be asked to give informal reports at those times. This is to help keep everyone on track and make sure the groups are working together. You will submit these progress reports online via bcourses and to report during in-person appointments during class. These reports will be graded based on completion.

Report 0 (Choose your group) – Due April 2nd Sign up with a group during class on Thursday.

Report 1 (Initial Proposal) – Due April 14th 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.

If you are having any problems pulling together as a group, you should tell me AT THIS POINT. I may reshuffle groups, depending on the problem.

Report 2 (Data Acquisition)– Due Apr 21st You need to have downloaded the data, read it into R, and done some basic checks of the data. This includes:

- The number of rows of the data
- First 5 rows of the data (only relevant variables for the analysis)

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 may not be trivial so please start this as soon as possible.

Report 3 (Initial Analysis)– Due Apr 28th You need to have prepared your data for analysis. This includes properly coding for non-response, recoding variables into factors with informative names, and so forth. You should submit some summary statistics of the data, acquired through R:

- Standard summary of variables of interest.

- Two graphical displays of your data of interest (does not yet 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. Submit the R code.

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

Report 4: Final Poster and Code – Due Friday, May 15th More specific guidelines about the format of the poster will be available later. The emphasis will be on understanding the design and trying to appropriately use the design elements to analyze a question. The structure of the poster will be

- Abstract which includes brief summary of your findings
- Description of the survey (including its design components)
- 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

Organizing Your Project

Students in the past have found GoogleDocs a good tool for people to able to work on a document jointly, minimizing the need to pass around a document by email.

Once a dataset has been chosen and some variables of interest decided upon, there are three large chunks for this project:

1. Data acquisition and making sense of the design elements
2. Analyzing the data in R
3. Preparing the poster

For Report #1, assign one person for each of these portions to be the organizer. However, EVERYONE should contribute to ALL of these portions. Note that along with the poster, you will submit the R code used to prepare and analyze your data. You will also regularly submit the Participant Form that specifies who was the person in charge of each section and which tasks within each section were performed by whom, and you will submit the Task Log recording when group members meet and what they worked on. The Participant Form and Task Log are live documents in that they should be kept up to date and submitted weekly.

The actual analysis of the data is where I 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 review the initial analysis and then meet (in person!). During this meeting talk out whether anything is missing or could be done better.

Data

The following are national statistical agencies where I have found the data to be reasonably straightforward to obtain and download. I give 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.

- 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/`. You must create a free account, and agree to the terms of use. Examples:
  - Annual Survey of Jails `https://www.icpsr.umich.edu/icpsrweb/ICPSR/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 Transportation 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 some of the data is distributed as a SAS transport file (`http://www.cdc.gov/nchs/nhanes/sas_viewer.htm`) which you should be able to convert. And, some offer R code for reading in the data.
  - NHANES `http://www.cdc.gov/nchs/nhanes/nhanes_questionnaires.htm`
  - National Immunization Survey `http://www.cdc.gov/nchs/nis/data_files.htm`
  - National Survey of Family Growth `http://www.cdc.gov/nchs/nsfg/nsfg_2011_2013_puf.htm`
- The Substance Abuse and Mental Health Data Archive `http://www.icpsr.umich.edu/icpsrweb/content/SAMHDA/index.html`
  - National Survey on Drug Use and Health `https://nsduhweb.rti.org/respweb/homepage.cfm`

The Scottish National Centre for Research Methods provides some data and exemplars at `http://www.restore.ac.uk/PEAS/surveys2.php`. These date back to 2009, but other more recent surveys may be available.

The United Nations Economic Commission for Europe has data, such as the Fertility and Family Survey at `http://www.unece.org/pau/ffs/ffsdata.html`. They require a signed application form so you may need to have a back up plan if not allowed or if the process takes too long.

There are other US national statistical agencies that may have survey data, for example

- National Center for Education Statistics (`https://nces.ed.gov/`)
- Bureau of Labor Statistics (`www.bls.gov`)
- Census Bureau (`https://www.census.gov`)

You may have luck finding surveys there.

Please note that you need to have data, and not summarized tables. Also, not that you need a complex survey design.