Math 10 — Spring 2013
Project
Proposal Due April 26th, 2013
Paper Due May 24th

Data! Data! Data! I can't make bricks without clay! — Sherlock Holmes

---

The project is to write up a paper which analyzes a large data set of your choosing which is freely available online. Your paper should address 2–3 research questions about this data set which can be answered using a hypothesis test or confidence intervals. These questions should be questions that you are actually interested in answering (and don't already know!) All computational work should be done in R.

Proposal — due April 26th by 5 pm

Your proposal should be turned in to the box outside Kemeny 008 and should consist of the following sections. (It need not be written up in paper format.)

- **Research questions:** In a few sentences, what are the 2–3 questions you would like to answer?
- **Data Source:** Include the citation for your data, and a link to the source or description of where we can find it.
- **Data collection:** How was the data collected? Does this data come from an observational study or experiment?
- **Variables:** What are the variables (numerical or categorical) you will be studying?
- **Data clean-up:** In some cases you will need to do some work in R to clean up your data. If so, include a description of any clean-up necessary and the relevant commands you will need to use in R to do so.
- **Exploratory data analysis:** Perform relevant descriptive statistics (In R), including summary statistics and visualization of the data. Include the commands required to import your data into R, and the commands required to generate these summary statistics.
- **Data:** Print out 1 page of your data set and attach it to your proposal. Your data likely will not fit on one page, so only include as many rows as will fit, along with relevant column headings. Your print out should, however, contain all relevant columns.

Project — due May 24th by 5 pm

Your project submission will consist of a research paper and an R script containing the relevant commands necessary to import your data and reproduce the analysis and statistics/visuals used in your paper. Submit a hard copy of your paper in the boxes outside of Kemeny 008, and upload a copy of both your paper and R-script to the Blackboard website. Your paper should address the following topics:

- **Introduction**
  What is your research question? Why do you care? Why should others care?
- **Data** A description of your data, how it was obtained, what sort of study or experiment was involved, and who gathered the data.
- **Data analysis** Include relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.

- **Inference** Discuss whether or not relevant conditions are met to do inference related to your questions and why. Perform a theoretical inference (if possible) including hypothesis tests and confidence intervals. If a theoretical analysis is not justified, then use a simulation based inference.
- **Conclusions** A summary of your findings without repeating your statements from earlier. Also include a discussion of what you have learned about your research question and the data you collected. You may also want to include ideas for possible future research.

We aren't too concerned about the length of your paper, however if you're going much over seven pages, you're probably writing far too much. Don't fill it with useless information that isn't relevant; every sentence should add something to your argument!

Please feel free to ask us questions about the project. If you are unsure of the suitability of a particular data set, we're happy to help you.[1]

[1] Adapted from `http://stat.duke.edu/courses/Spring13/sta101.001/projects/project1.pdf`