Project 1

Your first project is writing a research paper addressing a research question that can be answered using a hypothesis test and/or confidence interval. You can choose to work with one numerical or one categorical variable, or you can choose to evaluate the relationship between a categorical and a numerical or two categorical variables (so the only combination that is out is two numerical variables – that is the focus of the second project).

Data

You can collect your own data, or find a dataset online or from another source. You may not use any dataset used in this class - from labs, textbook, etc. Avoid using population datasets (such as data from all countries, all counties, etc.) – the purpose of this project is to do statistical inference, which means you should be starting with sample data. If you collect your own data and your sampling method isn't perfect, or find a dataset where you are worried the data may not have been sampled randomly, you may still use the dataset with appropriate discussion of the shortcomings of the study, biases, etc. I would recommend that you choose to work with two variables and explore the relationshis between them, unless you have an interesting research question that involves only one variable.

Proposal - due on Friday, Oct 4, at 5pm

Your proposal should consist of the following sections.

- **Research question:** In one sentence, what is your research question? (Your research question should be something you are genuinely interested in!)
- **Data:** Include the citation for your data, and (if available) link to the source.
  - **Data source:** Include the citation for your data, and (if available) link to the source.
  - **Data collection:** How was the data collected?
  - **Cases:** What are the cases (units of observation or experiment)?
  - **Variables:** What are the one or two variables you will be studying?
  - **Type of study:** Observational study or experiment? Explain.
  - **Scope of inference:** Can these data be used to establish causal links and/or can findings be generalized to the population at large?
  - **Data clean-up:** (Optional) If you had to do any data clean up in R, you can include the code and a **very brief** description of your steps here.
- **Exploratory data analysis:** Perform relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.
- **Data:** Print out 1 page of your data set and attach it to your proposal. If your data fits in one page, great. If you have too many observations and it won't fit, that's ok too. I just want to get a sense of your data set, I do not need to see all rows. However your print out should contain all relevant columns (this shouldn't be an issue since you are working with one or two variables for your project).

Goal

Your goal is to submit a completely reproducible proposal (i.e. if I download your markdown file and run it on my computer I should get the same results and write-up) giving me enough detail about what you want to do in your project, so that I can give you feedback before you set out to complete it. You should also complete a thorough exploratory data analysis so that you familiarize yourself with your data and decide whether or not the data are appropriate for the project and whether you should be expecting surprising results.

Getting your data into RStudio

Instructions for getting data into RStudio are posted on the course FAQ page (See #3 under R related, "My dataset is in a .csv file on my computer. How can I get it into RStudio?".)

Format & length

Your proposal should be written using the markdown template, so that all R code, output, and plots will be automatically included in your write up. In order to download the template for the proposal type the following in RStudio:

```
download.file("http://stat.duke.edu/courses/Fall13/sta101/projects/prj1_proposal.Rmd",
    destfile = "prj1_proposal.Rmd")
```

Your proposal should be at most 2 pages. Keep it brief as the information on this document will eventually make it into your project. If you make it too long you'll end up having to cut it down later.

Submission

Late work policy applies. You will receive feedback on your proposal within a week.

- Hard copy: Proposal (including a 1 page print-out of your data) turned in at my office, 213 Old Chem. Slide under the door if the door is closed.
- Online at Sakai under Assignments: (These will be time stamped, and late penalty will be applied based on the time stamp.)
  1. Markdown file (.Rmd). (See #5 under R related "How can I export my .Rmd file so that I can submit it on Sakai?" on the course FAQ.)
  2. Data file (in .csv format).

Grading

You'll receive a check+, check, check-, or 0, and feedback. Note that proposals make up 10% of the Project 1 grade.

Project - due on Thursday, Nov 7, in class

Your project should be a write up of parts 1 - 5 below in the form of a research paper.

- **Part 1: Introduction**
  What is your research question? Why do you care? Why should others care?
- **Part 2: Data**
  This should be a **cohesive write-up** of the data section from your proposal, i.e. not in bullet point form.
- **Part 3: Exploratory data analysis**
  Perform relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.
- **Part 4: Inference**
  - Check conditions
  - Theoretical inference (if possible) - hypothesis test and confidence interval
  - Simulation based inference - hypothesis test and confidence interval
  - Brief description of methodology that reflects your conceptual understanding

If your data fails some conditions and you can't use a theoretical method, then you should use simulation. It is your responsibility to figure out the appropriate methodology.

Important note: If you're using a categorical variable with more than two levels, for the inference section you should either combine or drop some of the levels, so that you only have two levels to work with. This only applies to the inference section, there are no restrictions on the number of levels you can use for the exploratory data analysis.

- **Part 5: Conclusion**
  Write a brief 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.

Goal

Your goal is to submit a completely reproducible project (i.e. if I download your markdown file and run it on my computer I should get the same results and write-up) that conveys that you have mastered statistical inference techniques that we have learned in class and that help you answer your research question.

Format & length

Your paper should be written using the markdown template, so that all R code, output, and plots will be automatically included in your write up. In order to download the template for the project type the following in RStudio:

```
download.file("http://stat.duke.edu/courses/Fall13/sta101/projects/prj1_new.Rmd",
    destfile = "prj1_new.Rmd")
```

Your write up should be at most 5 pages (including figures and R code). This is not very long, you will need to be concise. Every sentence should add something to your paper.

Tone

Write as if you are explaining your results to whoever would be interested in your research question, whether this is other scholars in your field or peers sharing your interest in the topic. Keep in mind this audience may or may not have taken statistics. You must be statistically accurate and use correct statistical terminology, but must also explain your conclusions in a way that anyone can understand.

Submission

- Hard copy: Project write-up.
- Online at Sakai under Assignments: (These will be time stamped, and late penalty will be applied based on the time stamp.)
  1. Markdown file (.Rmd)
  2. Data file (in .csv format).

Late work policy

10% off for every day (24-hours) late - that's 1 point per day for the proposal, and 10 points per day for the project.

Support

As always, feel free to come to me or your TA with questions about your proposal. I will hold additional office hours (TBA) before the proposal and project deadlines.

Honor code

You may talk with your peers (feel free to ask questions of each other, share ideas, or discuss concepts), but all calculations, R code, and writing must be done individually. You may not share your paper or code with any classmates until after the deadline of Nov 9. Failure to abide by these policies will result in a 0 for everyone involved.

You must electronically sign the honor pledge associated with this assignment on Sakai.

Grading

A rubric will be provided as we get closer to the project deadline.

Stat 101 - Prof. Rundel Section:_____ Name:_____

Project 1 Grading Rubric

Part	Assigned	Earned
Part 1: Introduction What is your research question? Why do you care? Why should others care?	10 pts	
Part 2: Data Cohesive write-up of the data section from your proposal. Review instructions for proposal.	15 pts	
Part 3: Exploratory data analysis Perform relevant descriptive statistics, including summary statistics and visualization of the data. Also address what the exploratory data analysis suggests about your research question.	20 pts	
Part 4: Inference Use theoretical and/or simulation methods to perform a hypothesis test, and also construct a confidence interval. If your data fails some conditions and you can't use a theoretical method, then you should use simulation. It is your responsibility to figure out the appropriate methodology, and apply all methods that are appropriate.	30 pts	
Part 5: Conclusion Write a brief 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 should also acknowledge limitations of your study and include ideas for possible future research.	10 pts	
Overall writing quality Writing quality and clarity, including grammar, spelling and organization	5 pts	
Proposal Check+ = 10, check = 8, check- = 5	10 pts	
Penalties - Did not follow five part format - Difficult to follow, requiring to go back to the data and replicate the analysis - Each page over limit - Using population data instead of sample data	 -5 pts - 10 pts - 5 pts per page - 5 pts	
Total	100 pts	