Info 271B
Fall 2013

Lab 2: Basic Statistical Tests

Introduction.

In this lab, you will work in groups of two to write your own R script and perform an analysis on data from the General Social Survey (GSS). Your group should submit two files:

1. A script file (.R) containing your program code
    a. Include a title section containing your full names, program description, approximate time it took to complete the lab, and whether you mind me using your results as a class example.
    b. Include comments before each logical section of the program, explaining what it does
    c. Label your variables clearly
    d. Name this file your-last-names-lab2.R

2. A commented log / output file containing your results
    a. Please use .doc or .pdf format
    b. Include the important output from running your script with comments, as well as your discussion, and graphs that you generate
    c. Name this file your-last-names-lab2.doc or .pdf

Place both files in a zip folder, named your-last-names-lab2.zip, and email them to agswigart at ischool.berkeley.edu. Please title your email "lab2 submission"

All files are due at 3:30pm on Thursday, Nov 14.

Data.

Every other year, the General Social Survey collects responses to thousands of questions, covering a wide variety of topics. You will be using a subset of data from 1993, including a small number of variables, which is available on the course website.

While some variables may be self-explanatory, others may not make sense until you look at the GSS codebook. An easy way to investigate a variable is to look it up in the GSS mnemonic index, located at:

http://www3.norc.org/GSS+Website/Browse+GSS+Variables/Mnemonic+Index/

Before you run a statistical test on a variable, you should always read its description in the codebook, in order to understand what different values mean. For example, the codebook

may explain if certain values stand for missing data. If this occurs, you should make sure those values are recorded as NA in R before proceeding.

Like any survey, GSS data creates additional concerns that would normally go into a statistical analysis. Surveys are usually weighted in order to compensate for over- or under-representation of subgroups. For this lab, however, you will be using unweighted data, which limits how well your findings generalize to the U.S. population.

Program Requirements.

Write a well-commented R script to perform each of the following tasks, following the best practices described in class. For each new variable, look for obvious errors and make sure that appropriate values are coded as NA. In your comments, please document any recoding you perform, and all the assumptions you must check before administering a test. Then explain how you select the correct test variety (e.g. parametric versus non-parametric)

1. Comparing means
    a. Create a binary variable based on a categorical variable with more than two categories.
    b. Select a metric variable (or an ordinal variable, but you may have to place it in the as.numeric function before conducting a test), and perform an analysis to determine if the mean of this variable is significantly different between the two groups you created in part a.
    c. In your write-up, describe your null and alternate hypotheses, and comment on the significance of your findings.
2. Measuring correlation
    a. Select two metric variables, and perform an analysis to see if they share a linear relationship.
    b. In your write-up, examine the scatterplot of the variables and relate it to your test results. Comment on the type of relationship you find. Finally, describe what you think might be going on (or not) between the variables.
3. Testing independence
    a. Select two categorical variables, and perform an analysis to see if they are independent of each other. Check to see which combinations are driving your result, and comment on their significance. Finally, comment on what you think might be going on (or not) between the variables.