I271b Lab 2
Fall 2014

Overview:
This lab has two parts, a multiple choice question section and a series of tasks using R. You will provide just the letter to each multiple choice item for Part 1, and then answer the questions from Part 2. You should include any key output and/or graphics in the primary report file as well. In addition, include your fully commented R script from Part2 along with your submission (we should understand your answers to part 2 without having to read the R script).

Submission:
Please organize all of your answers into one primary "lab report" file using a standard format (.pdf, .doc, etc). Please name this primary document "lab2_YourLastName.doc". Similarly, your script should be named "lab2_YourLastName.R"

This lab is due on Thursday, October 30th by 10am. You should email your two files (the main lab report and your R script) to Toshiro (tnish@berkeley.edu). Please plan ahead, we do not accept late papers.

Part 1: Multiple Choice (36 points)
1. In a survey, a question asks how many pets you have, with four possible responses: 0, 1 to 2, 3 to 5, and 6 or more. What type of variable does this question produce?

a) Interval
b) Dichotomous
c) Normal
d) Ratio
e) Ordinal

2. Many scientists view the concept of gender as a wide spectrum based on biological and social factors. In light of this, the choice to measure the concept of gender using a strict male/female dichotomy is an example of:

a) measuring a potentially interval or ordinal variable as a binary variable
b) measuring a nominal variable as an interval variable
c) giving priority to psychological over social variables
d) the ecological fallacy
e) giving priority to conceptual over operational variables

3. Which of the following is a benefit of using standard deviation as a measure of dispersion?

a) Standard deviations are unaffected by outlying data points.
b) The chance that a single draw from a population falls within one standard deviation of the mean is always the same for any population.
c) Standard deviations are unaffected by multiplying a variable by a constant.
d) Standard deviations can be directly compared to the individual deviation of one data point away from the mean.

4. Your friend thinks that gender and place of residence influence perceptions about health care legislation. She obtains a list of voters in the state of North Carolina and divides the list into subpopulations of men who live in small towns, men who live in large towns, women who live in small towns, and women who live in large towns. She then randomly selects 500 people from each subpopulation. What kind of sampling procedure is your friend using?

a) Non-probability quota sampling
b) stratified random sampling
c) social network sampling
d) cluster sampling
e) nonrandom sampling
f) systematic random sampling

5. Suppose that weekly beer consumption among the San Francisco population is normally distributed, with a mean of 50oz. Which of the following is more likely to occur:

a) Choosing one San Franciscan at random and finding that they drink over 70oz of beer a week.
b) Choosing 100 San Franciscans at random and finding that they drink an average of over 52oz of beer a week.
c) a and b are equally likely.
d) It depends on the standard deviation of the population.

6. Why is the Central Limit Theorem Important to scientists?

a) For large samples, it guarantees that a sample mean approaches the true population mean.
b) For large samples, it suggests that the normal distribution is a good model for the distribution of the mean and other statistics.
c) For large samples, it suggests that the sampling distribution of a variable approaches the population distribution of that variable.
d) When a population distribution is normal, it tells us that the sampling distribution of the mean will also be normal.

7. Say you collected data from 15 fellow classmates. One of your measures included age. Surprisingly, all 15 of your fellow classmates are of the same age: 25 years old. Given this information, which of the following statements below must be false.

a) The mean, median, and mode for your age variable are equal.
b) The distribution of your age variable is unimodal.
c) The variance and standard deviation of your age variable are both exactly zero.
d) The distribution of your age variable is platykurtic.

8. An unexpected software bug affects all Mac laptops that were manufactured in April of 2012, permanently replacing the user's desktop with photos of cats. You want to take advantage of this (for scientific purposes!) by measure productivity levels among users that own these "buggy" laptops and compare them to users that own "non-buggy" laptops made in the months immediately after April 2012. This is an example of a

a. Pretest-postest experimental design
b. Natural experiment
c. Associational non-experiment
d. Confounding covariate
e. Solomon 4-group design

9. A researcher obtains a convenience (non-random) sample of 100 undergraduates to come to a lab. Each participant is given an anxiety questionnaire before being randomly assigned to only one of four conditions, "music by Rebecca Black," "music by Nickleback," "music by Yanni," and "quiet room." The participants are then given the anxiety questionnaire again after their treatment. This is an example of a

f. Pretest-postest experimental design
g. Quasi-experiment
h. Associational non-experiment
i. Confounding covariate
j. Solomon 4-group design

Part 2: Data Analysis and Short Answer (64 points)

Dataset:
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. This may be found in the file, GSS.Rdata. This is the same dataset that we have been using in class examples for the past two weeks.

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.

Write a well-commented R script to perform each of the following tasks, then answer the provided questions in your main report. **Include all important output and answers to each question in your main lab report. You can also copy any graphics into the main lab report to make it easier for you to provide context for your answers.** We should be able to understand what you did and what your answer is for each item in your main lab report without hunting for things in your R script.

1. Data Import and Error Checking: Using the GSS dataset.
    a. Examine the “agewed” variable (age when married).
        i. What are the value(s) of agewed, if any, that do not meaningfully correspond to ages?

    b. Recode any value(s) that do not correspond to age as NA.
        i. What is the mean of the agewed variable?

2. Checking assumptions
    a. Produce a QQ plot for the agewed variable.
        i. Using this plot information, is agewed normal and how precisely do you know?

    b. Perform a Shapiro-Wilk test on the agewed variable.
        i. What is the precise null and alternative hypothesis for your test?
        ii. What is your p-value, and what is your specific conclusion?

    c. What is the variance of agewed for men? What is the variance of agewed for women?

    d. Perform a Levene’s test for the agewed variable grouped by men and women.
        i. What is the precise null and alternative hypothesis for this test?
        ii. What is your p-value, and what is your specific conclusion?

3. More on Plots and Visualizing Data

    a. Pick one metric variable from the dataset (other than agewed) and any other variable of interest to you. Produce a plot or visualization that allows you to look at your two variables in an interesting or informative way. (Feel free to get creative and try some things with ggplot, but you are welcome to produce any type of plot or graphic that you want no matter how simple or complex)
        i. Write a short paragraph describing what your chosen plot or graphic shows about your variables (**be descriptive!**).