Info 271B - Lab 2
Fall 2020

Overview:
This lab has two parts:

- Part 1: a series of tasks using R and short answer
- Part 2: multiple choice questions

In your lab report, for Part 1, please write up answers to the questions and include any key output and/or graphics. In your lab report, for Part 2 please just provide the letter answer to each multiple-choice item. In addition to your lab report, please turn in your fully commented Rmd script from Part 1 (note, we should understand your answers to Part 1 without having to read the Rmd script).

Working on the Lab:
You are welcome to talk about and work together to build out and test your R scripts for Part 1 in and outside of class. Importantly, the final script and answers that you turn in, including all comments, should be your own work.

Part 2 (multiple choice) should be done **completely independently.** Part 2 is an individual task, so do not work on or discuss this part with other students.

Submission:
Each student should submit the following:

- Lab report, named “lab2_LastName_FirstName.pdf”
- Rmd script, named “lab2_LastName_FirstName.Rmd”
- Html output from Rmd file, named “lab2_ LastName_FirstName.html”

This lab is due on **Thursday, Nov 5th by 11:59pm**. You should upload your three files (the lab report, Rmd and html files) to bCourses.

Please plan ahead; we do not accept late papers.

To create a new Rmd file, go to File -> New File -> R Markdown. In RStudio, your open Rmd file will have a “Knit” option in the toolbar. When ready to output, select knit -> Knit to Html.

Part 1: Data Analysis in R and Short Answer (70 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 2016, including a small number of variables. This may be found in the file, Lab2_GSS.Rdata (**bcourses -> Files -> Labs -> Lab 2 -> Lab2_GSS.Rdata**). This is a similar dataset to the one we’ve been using in class examples last week.

For any secondary data, like the GSS, we should look at the official codebook so that we know how variables are coded. **We are especially interested in knowing about 'missing' codes (those values such as not applicable, no answer, etc)!** Use this index to search for any variables you are interested in:
https://gssdataexplorer.norc.org/variables/vfilter

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 Rmd script to perform each of the following tasks, then answer the provided questions in your lab report. **Include all important output and answers to each question in your lab report. You can also copy any graphics into the 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 lab report without hunting for things in your Rmd script.

Before you start your work, we suggest exploring the variables you will be using by looking at them in R as we did in class examples. Also, look at the codebook. Try to identify missing values, or values that would otherwise not make sense in your analyses!

1. Data Import and Error Checking: Using the GSS dataset.
    a. Examine the *educ* variable.

        From the codebook, what are the value(s) of *educ*, if any, that do not meaningfully correspond to the highest year of school completed?

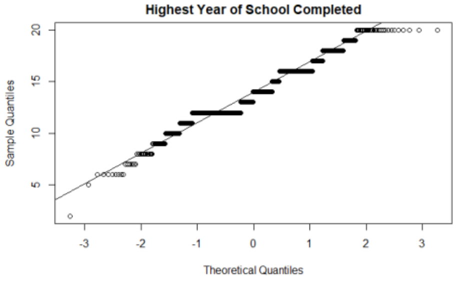
        - *Highest year of school completed.*
        - *99: No answers*

    b. Recode any value(s) that do not correspond to *educ* as NA.

        What is the mean of the *educ* variable when you properly account for missing values?

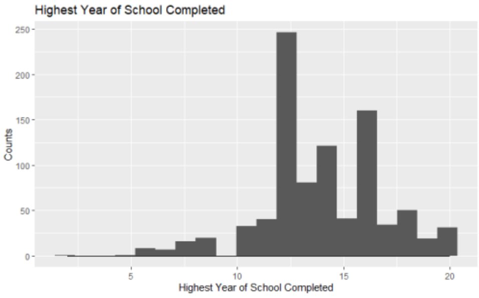
        - *Mean: year 13.7*

2. Checking assumptions
    a. Produce a QQ plot and a histogram for the *educ* variable.
        i. First, explain what a QQ plot is, and how it is interpreted.
            - *Compare the sample distribution over the normal distribution*
        ii. Using your plot information, is *educ* normal and how precisely do you know from the QQ plot?
            - *almost but not quite.*



iii. Using the histogram information, what can you say about the distribution of *educ*?

- *Skewed, almost normal distributed*



b. Perform a Shapiro-Wilk test on the *educ* 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?

- *2.51e-14, statistically significant, thus we say that these variables do not have a normal distribution relationship.*

c. What is the variance of *educ* for people whose family income in 2015 (*income16*) was greater than or equal to $90,000?

- *6.9*

d. What is the variance of *educ* for people whose family income in 2015 was less than $90,000?
(*Hint: you will need to create a recode from the "income16" variable, taking into account missing values.*)

- *7.83*

Commented [A1]: Needs attention on this, so far didn't take into account the nas

e. Perform a Levene's test for the *educ* variable grouped by whether their family income in 2015 (*income16*) was greater than or equal to $90,000, or less than $90,000. Remember you'll need to install and load the "car" package in order to use the leveneTest() function.

i. What is the precise null and alternative hypothesis for this test?

- *H0: The group which it's family income in 2015 was greater than or equal to $90K* ***has the same variance*** *as the group which it's family income is less than $90K.*
- *HA: The group which it's family income in 2015 was greater than or equal to $90K* ***dose not have the same variance*** *as the group which it's family income is less than $90K.*

ii. What is your p-value, and what is your specific conclusion?

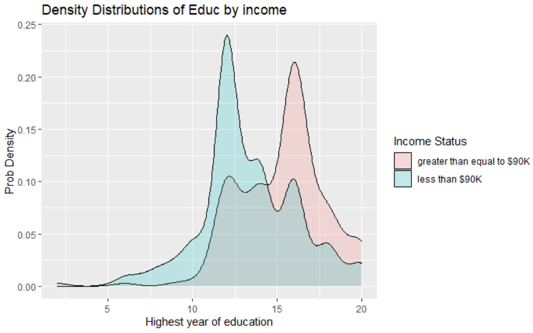
- *P value: 0.1182.*
  which is highly statistical significance level of 0.05.
  Thus, we reject the null hypothesis and conclude that the variance among the three groups is not equal.

3. Recoding variables and plotting data

a. Produce a single plot with two probability density distributions of *educ* by those whose family income in 2015 (*income16*) was greater than or equal to $90,000, or less than $90,000.

b. Write a line or two interpreting any differences in the two distributions. Make sure this plot is well formatted for a report (e.g. readable, labels, titles, no missing data showing, etc.)

   i. The probability density for those whose family income in 2015 income more than $90k have a higher education comparing to the other group, it has a mean around year16; while for those whose family income in 2015 less than $90K have a mean around year12.



4. Get Creative with Visualizing Data

From the GSS dataset, use *poleff11* (a Likert variable indicating level of (dis)agreement with “People like me don't have any say about what the government does.”) and a second variable (of any type) of interest to you. If you wish, you may recode either variable in an appropriate way. Produce a report-ready plot or visualization using ggplot2 that allows you to look at your two variables in an interesting or informative way.

a. Please use at least 2 of the following in your plot (**be creative!**)

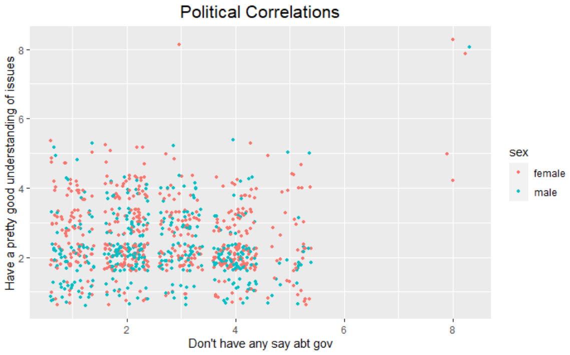
   i. faceting

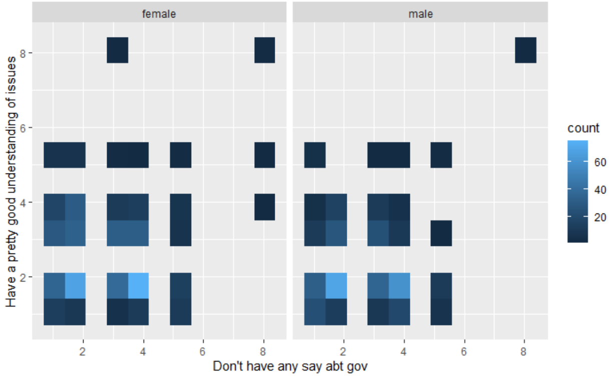
   ii. visually represent differences between two or more groups (e.g. differentiate by color, lines, shapes, transparency, etc.)

   iii. alter positioning (e.g. jitter, dodge, stack, etc.)

   iv. altering the theme (e.g., color customizations, etc).

b. Write a short paragraph describing what your chosen plot or graphic shows about your variables (**be descriptive!**)





i. I choose to explore the correlation between variable poleff11 – whether individual have a pretty good understanding of issues, and poleff13 – individual believe they do not have any say about the government decision, counts faceted by gender.

ii. The general trend in both gender groups is that people who knows better about the issue believe that they have a say in deciding government's decision. (since the count decrease vertically)

iii. Most of the population are scattered in the range of they think they don't have a say in what government do, and also do not have a good understanding of the issue.

Part 2: Multiple Choice (30 points)
Select one response for each question.

1. Which of these options best describes standard error?

a) The z-score of errors in a model
b) The amount of variance in a random sample
c) The standard deviation of bias in a random sample
d) The standard deviation of the population distribution
e) A measure of spread in the sampling distribution
f) None of the above.

2. In a survey, a question asks how many pets you have, measured as 1-2, 3-4, 5-6, 7 or more. What type of variable does this question produce?

a) Normal
b) Nominal
c) Linear
d) Dichotomous
e) Ratio
f) Ordinal

3. Scientists speculate that lower heart rates in humans may be associated with certain sounds. A researcher obtains a convenience sample of 50 (all freshman males) from a university to come to a lab. Each participant has a baseline heart rate measure before being randomly assigned to listen to only one of the following sounds for 5 minutes: "sounds of rain", "sounds of people falling from very large objects", "sounds of rabbits eating vegetables", and "quiet room/no sound." The participants have their heart rates taken again after their treatment. This is an example of a:

a) Associational non-experiment
b) Quasi-experiment
c) Natural experiment
d) Solomon 4-group design
e) Pretest-post test experimental design
f) Stratified random sampling design
g) This is not a valid experiment because it does not have random sampling.

4. Suppose that some scientists argue that the concept of "sexual preference" is best measured as a scaled spectrum of preference rather than as nominal categories. On an existing national survey, these scientists are forced to use a measure of "sexual preference" that uses nominal responses. For these scientists, the existing survey measure is an example of:

a) Prioritizing ecological validity over internal validity
b) measuring a categorical variable as an interval or ordinal variable
c) normalizing a non-normal distribution of a concept
d) multidimensional scaling
e) measuring an interval or ordinal variable as a categorical variable
f) giving priority to theory instead of operationalization

5. Suppose that weekly Netflix consumption is normally distributed among the population of all Berkeley students, with a population mean of 8 hours per week. Which of the following is less likely to occur? (Note: remember the logic and equation of a z-test as you think this one through!)

a) Choosing 400 Berkeley students at random and finding that they watch an average of 8.5 hours of Netflix per week.
b) Choosing one Berkeley student at random and finding that she watches an average of 18 hours of Netflix per week.
c) a and b are equally likely.
d) We cannot determine without knowing the standard deviation of the population.

6. Which of the following could serve as a specific null hypothesis for a statistical hypothesis test?

a) The standard deviation of age among Stanford students is 2 years.
b) The mean age of Stanford students is not 20.
c) The distribution of ages among Stanford students is not normal.
d) The mean age of Stanford students is different than the mean age of Berkeley students.

7. Which of these responses best explains why the Central Limit Theorem is important?

a) For samples with n > 30, it guarantees that a given sample mean is equivalent to the true population mean.
b) For samples with n > 30, it suggests that the population distribution of a given variable will be normally distributed.
c) When a population distribution is non-normal, it tells us that the sampling distribution of the mean will also be non-normal.
d) It tells us that the normal distribution is a good model for the distribution of the mean and other statistics when we have a large random sample.
e) If we take infinite samples from a population, our sampling distribution of the mean will eventually look exactly like our population distribution.
f) Central Limit Theorem is not actually important to statistics. Fake news.

8. Suppose you are interested in measuring the distribution of "life happiness" on a 10-point scale among students at a large university. You collect a sample of 100 students. 87% of the students in your sample get an exact score of 8, and the remaining 13% have scores of 3 or less. Given this information, which of the following statements must be true (pick one)?

a) The distribution of "life happiness" in the entire population is unimodal.
b) The mean of "life happiness" in your sample is lower than the median and mode.
c) The distribution of "life happiness" in your sample is negatively skewed.
d) The distribution of "life happiness" in your sample is positively skewed.
e) a and b.
f) b and c.
g) b and d.
h) a, b, and c.
i) a, b, and d.

9. Your friend says that he can recite the entire lyrics to “Mahna Mahna” from The Muppet Show in less than 17.3 seconds. What would be an appropriate set of hypotheses to test this claim? (For song reference, please see: https://www.youtube.com/watch?v=QTXyXuqfBLA)

a) H0: $\mu < \mu0$; Ha: $\mu = \mu0$
b) H0: $\mu \neq \mu0$; Ha: $\mu < \mu0$
c) H0: $\mu = \mu0$; Ha: $\mu \neq \mu0$
d) H0: $\mu = \mu0$; Ha: $\mu < \mu0$
e) None of the above.

10. An unexpected software glitch randomly affects 41% of all Berkeley online live course sessions, changing all virtual backgrounds in the affected sessions to animated pictures of rabbits. You want to take advantage of this (for scientific purposes!) by measuring productivity levels among students in the affected classes and comparing them to student in the unaffected classes. This is an example of:

a) An Interaction effect
b) A Natural experiment
c) A Quasi experiment
d) A Solomon 4-group design
e) An Associational non-experiment
b) A Pretest-post test experimental design