Week 6 HW
DATA 210

Data: You will need to load several different data sets to complete this problem set. All of the data are available on Canvas.Don't forget that all data we provide you for this class can only be used for class purposes.

1. If you've ever taken a probability class you may have heard of the 'birthday problem'.[1] The problem demonstrates (perhaps counter-intuitively) that in a group of only 23 people the probability of at least two people sharing a birthday exceeds 50

   In this question, we're going to explore a variation of the birthday problem. Ultimately we want you to answer the following question: in a large group of people, what is the probability that you pick a random day of the year and nobody in the group has that birthday? For these questions you can assume that all years have 365 days. Thoughout this question, we'll only worry about birth month and day (i.e. you can ignore birth year)

   a) We'll begin with the basics. Write a line or two of code that randomly draws birthdays for 500 people and checks whether nobody has a birthday on the first day of the year (i.e. January 1). The `sample()` function should be helpful here. Hint: use the numbers 1 through 365 to represent birthdays, and sample from those numbers.

   b) The previous question allowed us to calculate whether or not somebody had a January 1 birthday in the sample we drew. That's not especially interesting, since our answer could change if we drew a new sample. It would be more interesting to know (before we draw the sample) what the *probability* is that nobody in our group had a birthday on January 1.

      We can use for-loops to estimate these types of probabilities. The idea is that if we repeat what we did in part A a large number of times (each time drawing a new random sample), the proportion of the samples in which nobody had a January 1 birthday equals the probability that nobody has a January 1 birthday (for a group this size).

      Write a for-loop that repeats the calculation you did in part A 1,000 times. Each time, you should store the result (i.e. whether or not there was anybody with a January 1 birthday) in a vector so that you can see all 1,000 results after the loop is complete.

   c) Find the mean of the vector of results you created in part B. What is the probability that, in a group of 500 people, nobody in the group has a birthday on January 1? Hint: Your answer should be somewhere between 0 and 1 (i.e. 0% and 100%). If your answer equals exactly 0 or 1, you've done something wrong.

   d) Now let's use what we learned about writing our own functions to make our code a little bit more general. Write a function where you specify the size of the group (rather than fixing it at 500), and the function tells you the probability that nobody in a group that size has a January 1 birthday. Use the function to estimate that probability for a group of 750 people. What is that probability?

   e) Bonus question: Use your function and another for-loop to calculate the probability for every group size between 500 and 1500. Use the plot() function to make a simple scatter plot of your results. What do you notice about the results?

[1] You can read more about the birthday problem here: https://en.wikipedia.org/wiki/Birthday_problem. There's also a recent episode of This American Life that discusses what the birthday problem can teach us about voter fraud (in the section called 'Fraud Complex'):https://www.thisamericanlife.org/630/things-i-mean-to-know

2. Typically when we're working with survey data, our goal is to use the responses of the survey to understand the larger group of people that the survey respondents represent (usually called the *sample frame*). If we're interested in understanding the political attitudes of everybody who lives in the United States, then we could use a survey of Americans to further our understanding. It's almost impossible, however, that a sample of a few thousand (or even a hundred thousand) people would have demographic attributes that are exactly identical to the US population as a whole. Because of this fact, we use survey weights to slightly adjust the composition of our sample to match the sampling frame.

   In an ideal world, every respondents' survey weight would equal exactly 1. In practice, this is never really possible in any real-world scenario. The best we can do is to draw a survey that's as representative of the sampling frame as possible, and then use weights to make small adjustments to account for the imperfection and randomness in the sampling process.

   In this question, we're going to use survey data from a July 2019 survey about American politics ("july-2019-sm-poll.sav). To keep things a bit simpler, the data we've provided for this problem set doesn't include every question on the survey.

   a. We'll begin by looking at the `weight` variable in the dataset. What is the average weight given to people in the dataset? Why is it generally a good idea for survey weights to have this average?

   b. Sometimes we trim survey weights so that no single respondent has a huge amount of influence on the conclusions we draw from the data. Does it appear that the survey weights in the data have been trimmed? If so, what value were they trimmed to?

   c. Identify the person in the data that has the highest survey weight. What is the race and gender of this person? Why might this person have such a high survey weight?

   d. What is the unweighted average age of people in the dataset? Using the survey weights, what is the weighted average age of people in the data? The `weighted.mean()` function might be helpful here. You can ignore the people who did not provide their age in the survey, and leave them out of the denominators. What does the difference between these two numbers indicate to you?

   e. Compare the unweighted versus weighted percentages of people in the survey who are white, black, Hispanic, and Asian. When you apply the weights, which groups increase in size and which decrease? By how much do each of the group sizes change (in terms of percentage points)? You might find the `wpct()` function in the 'weights' package to be useful.

   f. What percentage of people said that they would be somewhat or very willing to pay higher taxes to pay for infrastructure improvements? What is this percentage when you only look at Republicans? What about when you only look at Democrats? You should use the survey weights and omit people who did not answer the question from your calculations.

   g. Analyze the questions about trust in the state and federal governments. Use those variables, and others in the dataset, to find some interesting pattern or result in the data. What did you find?