Week 4 Exercises: Statistics for Data Science

Part 1: HINTS

The file hints_sample.csv contains a subset of the 2019 Health Information National Trends Survey. The HINTS is a nationally-representative survey which has been administered every few years by the National Cancer Institute (NCI) since 2003. Its target population is civilian, non-institutionalized adults aged 18 or older living in the United States.

In our subset, we will be looking at 3 variables from the survey: AverageTimeSitting: During the past 7 days, how much time did you spend sitting on a typical day at home or at work?

AvgDrinksPerWeek: Average number of drinks per week.

MorningNightPerson: Someone might describe themselves as a morning-person or night-person. Which do you consider yourself to be?

Read in this dataset as a dataframe named *hints*.

1. Plot the distribution of the AverageTimeSitting variable.
2. Find the mean and standard deviation of the AverageTimeSitting variable.
3. Would you say that the AverageTimeSitting variable is approximately normally distributed?
4. Plot the distribution of the AvgDrinksPerWeek variable.
5. Find the mean and standard deviation of the AvgDrinksPerWeek variable.
6. Would you say that the AvgDrinksPerWeek variable is approximately normally distributed?
7. What differences do you notice between these two variables?
8. Find the coefficient of variation between these variables.
9. Find a 95% confidence interval for the mean for both of these variables. What do you notice about the margin of error for these variables?
10. The MorningNightPerson has five possible responses:

    (1): I'm definitely a morning-person

    (2): I'm more of a morning-person than a night-person.

    (3): I'm neither a morning-person nor a night-person.

    (4): I'm more of a night-person than a morning-person.

    (5): I'm definitely a night-person.

Create a bar chart showing the percentage of our sample which falls into each category. What do you notice?

11. Create a 95% confidence interval for the proportion of the population which consider themselves to be a night-person (for purposes of this exercise, we'll consider a response of either 4 or 5 to be a night-person). If I claim that 1/3 of the population are night-people, does the data back up my claim?

Part 2: The Taylor Perkins Problem

In bootcamps at NSS, there are numerous group projects, and groups get rearranged frequently. The most recent part-time data analytics bootcamp had 24 students. Groups were assigned 5 times, with each group having 4 members. Despite trying to make sure that groups stay varied, there was a pair of students who were assigned to the same group 3 times. Let's say that you're interested in how likely it is for there to be a pair of students who are assigned to the same group at least 3 times if assignments are completely random (where the class size/number of assignments/group size is the same as that of the data analytics class). Since there are a ridiculous number of possible group assignments and combinatorics is hard, you decide to instead run a

simulation to estimate this probability (i.e., proportion). This means that you will create a large number of random assignments and check those assignments to see if there are any pairs of students who work together at least 3 times. This process is essentially the same as drawing a random sample from the population of all possible group assignments.

1. Based on a simulation of 100,000 random assignments, you see that in 99.378% of cases, there is at least one such pair of student. Using this information, create a 99% confidence interval for the true proportion of random assignments that result in at least one pair of students working together 3 times.
2. By how much does the margin of error shrink if 500,000 simulations are run instead of 100,000?
3. By how much does the margin of error shrink if 1,000,000 simulations are run instead of 100,000?

Part 3: Restaurant Permits

You are planning to open a new restaurant in Nashville, and are going to have an entirely new building built for this purpose. Before construction can start, you need to have a building permit issued. For planning purposes, you are trying to estimate how long you might have to wait for your permit.

The file restaurant_permits.csv contains information about all permits for new restaurants issued in Nashville for the last three years. The wait_time column contains the number of days between when a permit application was entered and when the permit was issued.

Read in this dataset as a dataframe named *permits*.

Based on this data, construct 95% confidence intervals for the following:

1. Average scenario: the mean wait time
2. Quick scenario: the 25th percentile of wait times
3. Worst-case scenario: the 90th percentile of wait times