Week 5 Exercises: Statistics for Data Science

Part 1: 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. Quick scenario: the 25th percentile of wait times
2. Worst-case scenario: the 90th percentile of wait times
3. Average scenario: the mean wait time. For this case, build a confidence interval using a t-interval (what we did last week) and also using a bootstrap interval. What do you notice about the results of both of these methods?

Part 2: Pulse Rates

The file nhanes_pulse_sample.csv contains a sample of 100 men between the ages of 30 and 40 from the 2015 National Health and Nutrition Examination Survey.

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

You suspect that the average pulse rate for men between the ages of 30 and 40 is greater than 70 and want to use this data to test this hypothesis.

1. What are your null and alternative hypotheses? Is this is one-tailed test or a two-tailed test?

2. Find the mean and standard deviation of the pulse rate in your sample.

3. Conduct a t-test to test your hypothesis.

4. What p-value did you get?

5. State your conclusion.

Part 3: Crashes on Weekend vs Weekday

We are interested in studying the difference between weekends and weekdays in terms of number of reported crashes in Davidson County. We speculate that there are a larger number of reported crashes on average on weekdays compared to weekends.

The file crashes_sample.csv contains a random sample of 65 randomly selected days.

Read in this data as a dataframe named *crashes*.

1. Look at the distribution of Accident_Number for weekends vs non-weekends. Do this distributions appear to be approximately normal?

2. What are your null and alternative hypotheses?

3. What is the observed difference in the average number of crashes on weekends vs. weekdays?

4. Conduct a t-test to test your hypotheses.

5. What p-value did you get?

6. State your conclusion.

7. Repeat this test, but use a permutation test this time. How do your results compare to the t-test?

Part 4: Russia Constitutional Referendum

A 2016 paper published in the Annals of Applied Statistics (https://projecteuclid.org/euclid.aoas/1458909907) suggests that election falsification can be indicated by the presence of higher-than-usual reported number of integer results. The paper suggests that this may be due to the well-known psychological phenomenon of attraction to round numbers.

Recently in Russia, a constitutional referendum passed, which included provisions allowing President Vladimir Putin to run again for two more six-year presidential terms.

The file results_2020_clean.csv contains a cleaned up version of these refenendum vote results by polling location. Note that 11 polling locations were dropped for having zero reported votes. The source for this data is https://github.com/khakhalin/Sketches/tree/master/ru_vote_2020.

The first three columns (region, tik, and uik) together identify the polling location.

The next three columns (yes, no, and total_votes) give the number of yes votes, number of no votes, and total number of votes cast for that location.

The winning_pct column is equal to yes / (yes + no), and the winning_pct_rounded column is the winning percentage rounded to one decimal place.

The final column, which is what we are most interested in, is the decimal value of the rounded winning percentage.

1. What proportion of the time would you expect the decimal part of random number rounded in the manner as above to be 0?
2. Create a bar chart showing the distribution of the occurences of each digit in the decimal column of the dataset. That is, you should have a bar for 0, 1, 2, 3, . . . , 9. What do you notice from your bar plot?
3. Assume, as a null hypothesis that each digit is equally likely to appear as the decimal for any polling location (i.e., the probability of the decimal being 0 is 0.1). Under the assumption of this null hypothesis, what is the probability of seeing as extreme a proportion of 0's in the decimal position as was observed in the referendum?
4. Does the result of this hypothesis test call into question to reported results of this referendum?

Part 5: Late Night Hit and Runs

You speculate that crashes ocurring late at night are more likely to be hit and run crashes. For purposes of this exercise, we have defined "late at night" to mean occurring between midnight and 5:00 AM.

The file hit_and_run_sample.csv contains a random sample of 50 car crashes that took place in Davidson County.

Read in this data as a dataframe named *hit_and_run*.

1. Find the proportion of hit and run crashes for both late at night and not late at night. What is the observed difference between these two?

2. State the null and alternative hypothesis.

3. Conduct a fisher exact test to test your hypothesis.

4. What p-value did you get?

5. State your conclusion.