Week 5 Exercises: Statistics for Data Science

Part 1: 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 2: 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.

Part 3: 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 4: 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.

Part 5: NHANES Blood Pressure

The file nhanes_blood_pressure.csv contains data coming from the 2013 National Health and Nutrition Examination Survey. Specifically, it contains three variables:

- SEQN: an identifier number per participant
- add_salt_rarely: Whether that participant indicated that they rarely added salt to their food.
- systolic_blood_pressure: Systolic blood pressure measurement.
- body_mass_index: Body mass index

You suspect that people who rarely add salt to their food will have lower blood pressures on average than those who add it more than rarely. Let's test this claim.

Read in the data as a dataframe named *nhanes*.

1. Create a boxplot showing systolic blood pressure vs whether a person rarely adds salt to their food. What do you notice from the boxplot.
2. What are the null and alternative hypothesis?
3. What is the observed difference in the average systolic blood pressure between groups?
4. Conduct a t-test to test your hypotheses.
5. What p-value did you get?
6. State your conclusion.

7. Find the effect size for the difference. What does this say about the difference you found between the two groups? blood pressure. What p-value do you obtain? State your conclusion.