Week 5 Exercises: Statistics for Data Science

Part 1: 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 as opposed 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 2: 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 z-test to test your hypothesis.
4. What p-value did you get?
5. State your conclusion.

Part 3: Permits Wait Time

The file residential_permits.csv contains data on the wait time for new residential construction permits issued in Nashville.

Let's say that you own Beazer Homes Corporation and suspect that your competitor, Legacy South Builders, LLC is getting their permits issued quicker than you are.

In this exercise, we will determine if you have a legitimate complaint.

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

1. Look at the distribution of wait times for both Beazer Homes Corporation (Contact == 'BEAZER HOMES CORP') and for Legacy South Builders (Contact == 'LEGACY SOUTH BUILDERS LLC'). Do these distributions appear to be approximately normal?
2. Conduct a permutation test, looking at the difference between average wait times for Beazer Homes and Legacy South. What p-value do you obtain?
3. Based on this p-value, what is your conclusion?

Part 4: 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?
8. Find the correlation between systolic blood pressure and body mass index for your sample.
9. Conduct a permutation test of whether there is a positive correlation between body mass index and systolic blood pressure. What p-value do you obtain? State your conclusion.