**Data 8 Spring 2020**

**Discussion: Hypothesis Testing (Lab 06)**

**Question 1.** Francie is flipping a coin. She thinks it is fair, but is not sure. She flips it 10 times, and gets heads 9 times.

She wishes to determine whether the coin was actually unfair, or whether the coin was fair and her result of 9 heads in 10 flips was by random chance.

1. What is a possible model that you can simulate under?

A possible model that you could simulate under could be that on each flip, there is a 50% chance of heads and a 50% chance of tails. Any difference is due to chance.

1. What is an alternative model for Francie’s coin? You don’t necessarily have to be able to simulate under this model.

An alternative model that Francie might suggest is that the coin is unfair. The difference in the observed data is due to something other than chance. We wouldn’t be able to simulate under this hypothesis, because the statement “the coin is unfair” is not testable (we can ask questions like “how unfair?” or “biased towards heads or tails?”)

1. What is a good statistic that you could simulate? Calculate that statistic for your observed data.

*Hint: If the coin was unfair, it could be biased towards heads or biased towards tails.*

The **absolute difference between the number of heads we observe and the expected number of heads (5)**. For our data, this is |9-5| = 4. Notice that this statistic is large for both a large number of heads and a small number of heads.

1. Complete the function flip\_coin\_10\_times, which takes no arguments and returns the absolute difference between the number of heads in 10 flips of a fair coin and the expected number of heads in 10 flips of a fair coin.

def flip\_coin\_10\_times():

probabilities = make\_array(0.5, 0.5)

proportions = sample\_proportions(\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)

num\_heads = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

return \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

def flip\_coin\_10\_times():

probabilities = make\_array(0.5, 0.5)

proportions = sample\_proportions(10, probabilities)

num\_heads = proportions.item(0)\*10

return abs(num\_heads - 5)

1. Complete the code below to simulate the experiment 10000 times and record the statistic in each of those trials in an array called abs\_differences.

trials = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

abs\_differences = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

for \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_:

abs\_diff\_one\_trial = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

abs\_differences = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

trials = 10000

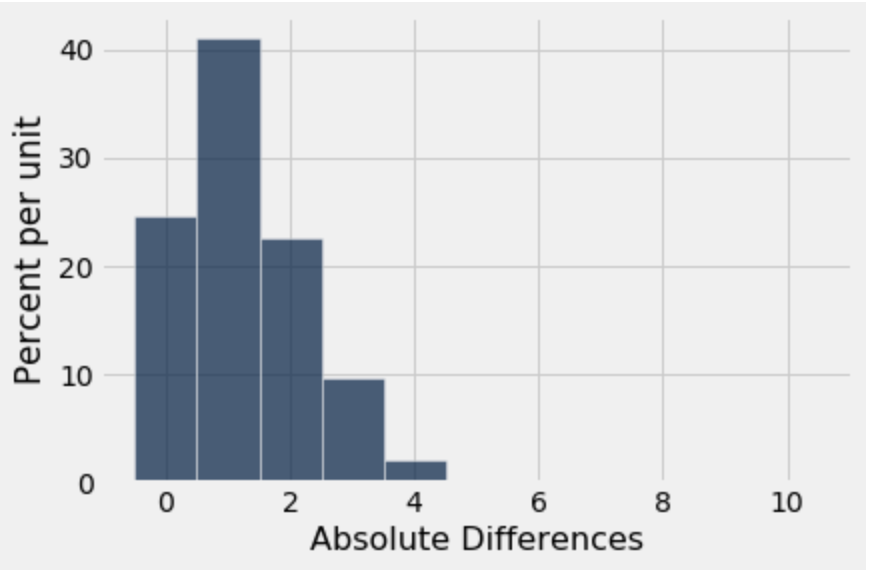
abs\_differences = make\_array()

for i in np.arange(trials):

abs\_diff\_one\_trial = flip\_coin\_10\_times()

abs\_differences = np.append(abs\_differences, abs\_diff\_one\_trial)

1. Suppose we performed the simulation and plotted a histogram of abs\_differences. The histogram is shown below.

****

Is the observed statistic described in the question consistent with the model we simulated under?

No, the observed statistic is not consistent with the model we simulated under. If we look for the observed statistic (4), we’ll see that it rarely ever happened in our simulation. Therefore, we would say that it is inconsistent with the hypothesis we simulated under.

**Question 2.** As a student fed up with waiting times at office hours, you scout out the number of people in office hours (OH) from 11-12, 12-1, and 1-2 in B6 Evans. The Head GSI claims that the distribution of students is even across the three times, but you do not believe so. You observe the following data:

|  |  |
| --- | --- |
| OH Time | Number of Students |
| 11-12 | 250 |
| 12-1 | 300 |
| 1-2 | 200 |

Being a cunning Data 8 student, you would like to test the Head GSI’s claim. Before you design your test, consider: are office hour times numerical data or categorical data?

1. What is the Head GSI’s hypothesis?

The distribution of students in office hours is equal, with ⅓ probabilities per time. Any difference is due to chance.

1. What is the student’s hypothesis?

The difference is not due to chance - the number of students are not evenly distributed among times, with some times having more students

1. Which hypothesis (Head GSI or student) can you simulate under?

You could simulate under the Head GSI’s hypothesis. This is because it is a fully defined model, meaning we are able to describe the parameters of an experiment surrounding it. The student hypothesis is simply that the distribution is not even among office hour times, but doesn’t give us any details that mean we can test it.

1. What is a good statistic to use? *Hint: What is a good statistic for measuring the distance between two categorical distributions?*

**TVD from expected distribution [⅓, ⅓, ⅓].** When we are observing **categorical distributions** of data and want to compare them, we should use TVD. Note, this is a good example because we have three different components in the distribution that we would like to test.