

## CS2065 Spring 2022

### Discussion: Hypothesis Testing (Lab 06)

When we observe something different from what we expect in real life (i.e. four 3's in six rolls of a fair die), a natural question to ask is "Was this observed difference from what we expect due to random chance? Or was it due to something other than random chance?"

*Hypothesis testing* allows us to answer that question in a scientific and consistent manner, using the power of computation and statistics to conduct simulations and draw conclusions from our data.

**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?
2. What is an alternative model for Francie's coin? You don't necessarily have to be able to simulate under this model.
3. 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.*
4. 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 _____
```

5. 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`.

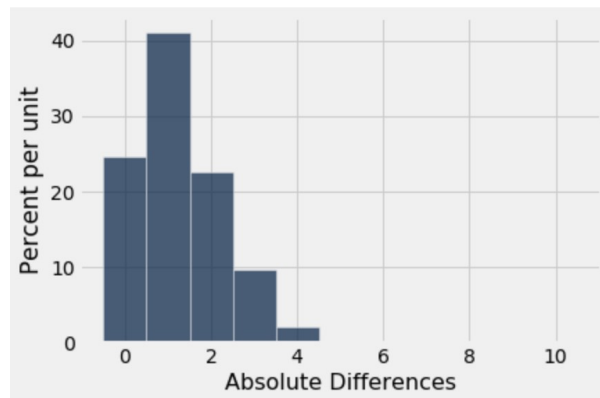
```

trials = _____
abs_differences = _____

for _____:
    abs_diff_one_trial = _____
    abs_differences = _____

```

6. 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?

**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 PL-6. The instructor 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	5
12-1	6
1-2	4

Being a cunning data science student, you would like to test the instructor's claim. Before you design your test, consider: are office hour times numerical data or categorical data?

- a. What is the instructor's hypothesis?
- b. What is the student's hypothesis?
- c. Which hypothesis (instructor or student) can you simulate under?
- d. What is a good statistic to use? *Hint: What is a good statistic for measuring the distance between two categorical distributions?*