Lab 06: Assessing Models

Data 8 Discussion Worksheet

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 unexpected behavior due to random chance, or something else?"

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

1. Flipping Fun: Sydnie is flipping a coin. She thinks it is unfair, but is not sure. She flips it 10 times, and gets heads 9 times. She wants to determine whether the coin was actually unfair, or whether the coin was fair and her result of 9 heads in 10 flips was due to random chance.

a. What is a possible model that she can simulate under?

b. What is an alternative model for Sydnie's coin? You don't necessarily have to be able to simulate under this model.

c. What is a good statistic that you could compute from the outcome of her flips? Calculate that statistic for your observed data.
Hint: If the coin was unfair, it could be biased towards heads or biased towards tails.

d. Complete the function `flip_coin_10_times`, which takes no arguments and returns the absolute difference between the observed 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 ________________________________________________________
```

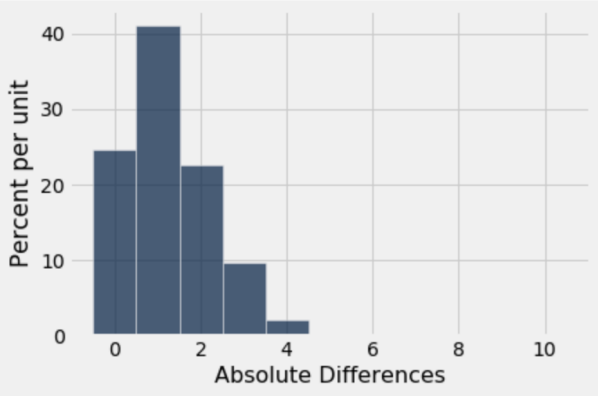
e. Rewrite `flip_coin_10_times` and use np.random.choice instead of sample_proportions this time. You are allowed to create new variables.

f. 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 = _______________________________________________
```

g. Suppose we performed the simulation and plotted a histogram of `abs_differences`. The histogram is shown below.



Is our observed statistic from part c consistent with the model we simulated under?

2. Data 8 Office Hours: As a student curious about office hours waiting times, you scout out the number of people in office hours (OH) from 11-12, 12-1, and 1-2 in SOCS 531. Meghan 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	50
12-1	60
1-2	40

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

a. What is Meghan's hypothesis?

b. What is the student's hypothesis?

c. Which hypothesis (Meghan 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?