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?

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.

The heads are like independent and identically distributed draws at random from a distribution in which 50% are *Heads* and 50% are *Tails*.

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

An alternative model that Sydnie 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 model, because the statement "the coin is unfair" is not testable (we can ask questions like "how unfair?" or "biased towards heads or tails?")

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.

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.

We could also use proportions, ie |number of heads/10 - 0.5|

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 ________________________________________________________
```

```
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)
```

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

```
def flip_coin_10_times():
    choices = make_array("Heads", "Tails")
    flips = np.random.choice(choices, 10)
    num_heads = np.count_nonzero(flips == "Heads")
    return abs(num_heads - 5)
```

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`.

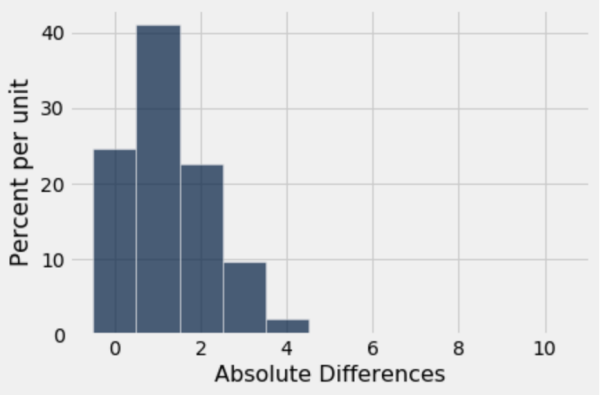
```
trials = ____________________________________________________________
abs_differences = ___________________________________________________

for ___________________________________:
    abs_diff_one_trial = ____________________________________________
    abs_differences = _______________________________________________
```

```
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)
```

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?

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 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?

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

b. 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

c. Which hypothesis (Meghan or student) can you simulate under?
   You could simulate under Meghan'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.

d. 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.