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?

Sydnie can simulate under the null hypothesis model, which assumes that the coin is fair and that differences from the theoretical value are due to random chance. From there, she can determine the total variation distance and consider whether the coin is fair based on her observation compared to her experimental simulated testing, which can be conventionally concluded (though subjectively it can be anything) using distribution, p-values, and statistical significance.

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 for Sydnie's coin is the alternative hypothesis model. It assumes that something makes the data disagree with the null hypothesis (the idea that differences in the data may only be due to chance) other than mere chance. In other words, it assumes that the null hypothesis model is wrong and bad--that the coin is unfair.

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.

A good statistic that one could compute from the outcome of her flips is the total variation distance of outcomes between her flips and the theoretical flip outcome of a fair coin (in terms of number of heads). The TVD of a theoretical outcome and hers is |9-5| = 4. From here, a simulated trial can determine the p-value that a fair coin has a TVD of 4 or more.

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(__10, probabilities__________________________)
    num_heads = ____proportions.item(0)*10________________________________
    return _num_heads___________________________________________
```

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

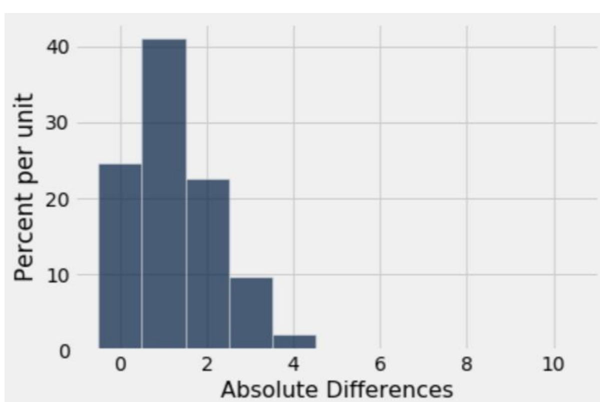
```
def flip_coin_10_times():
    results = np.random.choice(make_array("heads", "tails"), 10)
    return np.count_nonzero(results == "heads")
```

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 = ____flip_coin_10_times()_____________________________
abs_differences = __make_array(abs(trials-5))____________________

for __i in np.arange(9999)_______________:
    abs_diff_one_trial = __abs(flip_coin_10_times()-5)_____________
    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?

The observed statistic is consistent with the model we simulated under in that 9 heads is unlikely (in this case, it did not even happen!). However, we would expect the TVD's mean to be centered around the 0 mark instead of the 1 mark; therefore, this observed statistic is not super consistent with our theoretical value under which we simulated measured in part c and simulated in part e/f.

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?

categorical

a. What is Meghan's hypothesis?

Meghan's hypothesis is that there will be an even distribution of people across these categories (office hours). It is the null hypothesis.

b. What is the student's hypothesis?

The student's hypothesis is the alternative hypothesis, which is that Meghan's hypothesis/model (the null hypothesis) is wrong. There will not be an even distribution of people across all office hour categories.

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

Meghan's hypothesis is a model that can be simulated under.

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

Total variation distance between null hypothesis distribution and observed distribution is a good statistic to use here because data should be mostly centered on the left, so determining p-values is a one-sided ordeal.