**ASSESSING MODELS**

*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.*

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.

* 1. What is a possible model that she can simulate under?
  2. What is an alternative model for Sydnie’s coin? You don’t necessarily have to be able to simulate under this model.
  3. 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.
  4. 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 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* 1. 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 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

A graph with a bar graph

Description automatically generated

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

**Other problems we’ve encountered:**

* **The jury selection**
  + There was a discrepancy between the proportion of black jurors in the panel and the proportions of jurors in the population.
  + The goal: compare the distribution in the population of eligible jurors with the proportion of black panelists in the random sample.
  + The statistic: the number of black panelists in the sample
  + The procedure: create a distribution under the model of random selection and determine how likely it is that the discrepancy is due to chance.
  + The red dot is the observed statistic: the percentage of black jurors selected.

**A graph of a number of objects

Description automatically generated with medium confidence**

* + For multiple categories:
    - The statistic: the difference between two distributions
    - The red dot is the TVD between the distribution of ethnicities in the eligible juror population and the distribution in the panels:

**A graph of a number of blue bars

Description automatically generated with medium confidence**

* **Vaccination across the nation**
  + The goal: to find out each doctor’s position was more correct:
    - Is assigning vaccines 1 to people who show up MWF and vaccine 2 to people who show up TH the same as random selection with chance 3/5 (60%), 2/5 (40%)?
    - The statistic: the distance between the percent of people V1 and 60

**A graph with a red dot

Description automatically generated**

* **Contribution of multiple factors to overall happiness**
  + Multiple categories – calculate TVD

A graph of a number of tvds

Description automatically generated

**A/B TESTING**

*One special kind of hypothesis test we do in this class is called an A/B test. The steps used*

*to run an A/B test are the same as a general hypothesis test, but A/B tests have a specific*

*null hypothesis (that two samples were drawn from the same distribution). We carry out*

*this test by performing a permutation of our data.*

1. When should you use an A/B test versus another kind of hypothesis test?

Kevin, a museum curator, has recently been given specimens of caddisflies collected

from various parts of Northern California. The scientists who collected the caddisflies

think that caddisflies collected at higher altitudes tend to be bigger. They tell him that

the average length of the 560 caddisflies collected at high elevation is 14mm, while the

average length of the 450 caddisflies collected from a slightly lower elevation is 12mm.

He’s not sure that this difference really matters, and thinks that this could just be the

result of chance in sampling.

1. What’s an appropriate null hypothesis that Kevin can simulate under?
2. How could you test the null hypothesis in the A/B test from above? What

assumption would you make to test the hypothesis, and how would you simulate

under that assumption?

1. What would be a useful test statistic for the A/B test? Remember that the direction

of your test statistic should come from the initial setting.

1. Assume flies refers to the following table:

|  |  |
| --- | --- |
| ***Elevation*** | ***Specimen length*** |
| *High elevation* | *12.3* |
| *Low elevation* | *13.1* |
| *High elevation* | *12.0* |

*...*

*(1007 rows omitted)*

Fill in the blanks in this code to generate one value of the test statistic under the null

hypothesis.

def one\_simulation():

shuffled\_labels = flies.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.column(‘Elevation’)

shuffled\_flies =

flies.drop(‘Elevation’).with\_columns(\_\_\_\_\_\_\_\_\_\_,\_\_\_\_\_\_\_\_\_\_)

grouped = shuffled\_flies.\_\_\_\_\_\_\_\_(\_\_\_\_\_\_\_,\_\_\_\_\_\_\_)

means = grouped.column(‘Specimen length mean’)

statistic = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

return statistic

1. Fill in the code below to simulate 10000 trials of our permutation test.

test\_stats = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

repetitions = \_\_\_\_\_\_\_\_\_\_\_

for i in np.arange(\_\_\_\_\_\_\_\_\_\_\_):

one\_stat = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

test\_stats = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

test\_stats

g. The histogram of test\_stats is plotted below with a vertical red line indicating

the observed value of our test statistic. If the p-value cutoff we use is 5%, what is the

conclusion of our test?

A graph of a test

Description automatically generated

h. Suppose that the null hypothesis is true. If we ran this same hypothesis test 1000

times, each time drawing a new random sample from the population and with a

p-value cutoff of 5%, how many times would we expect to incorrectly reject the null

hypothesis?

i. What effect does decreasing our p-value cutoff have on the number of times we

expect to incorrectly reject the null hypothesis?