Lab 07: A/B Testing

Data 8 Discussion Worksheet

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. Mid-semester Check In

a. What has been your favorite topic/assignment/lecture/anything so far with the first half of the class done?

   My favorite assignment/topic/lecture was [....]
   My favorite assignment so far was [...]

   If you have any concerns about your performance in the class so far, feel free to bring it up to your lab TA.

2. A/B Testing and Error Probabilities

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

   You should use an AB test when determining whether two samples, also known as an A group and a B group, were sampled from the same underlying distribution/population.

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.

b. What's an appropriate null hypothesis that Kevin can simulate under?

The distribution of specimen lengths is the same for caddisflies sampled from high elevation as those sampled from low elevation.

c. 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?

If the null hypothesis is true - the caddisflies *did not* come from different distributions - then it shouldn’t matter how the samples were labeled (high elevation or low elevation). Under this assumption, you could shuffle the labels of the caddisflies and calculate your test statistic from this “relabeled” data.

d. 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

Difference in mean lengths between the two groups. Note that this is not an absolute difference-- we could choose either order for subtraction, but that would affect the direction of our alternative hypothesis so we need to be careful!

e. 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.

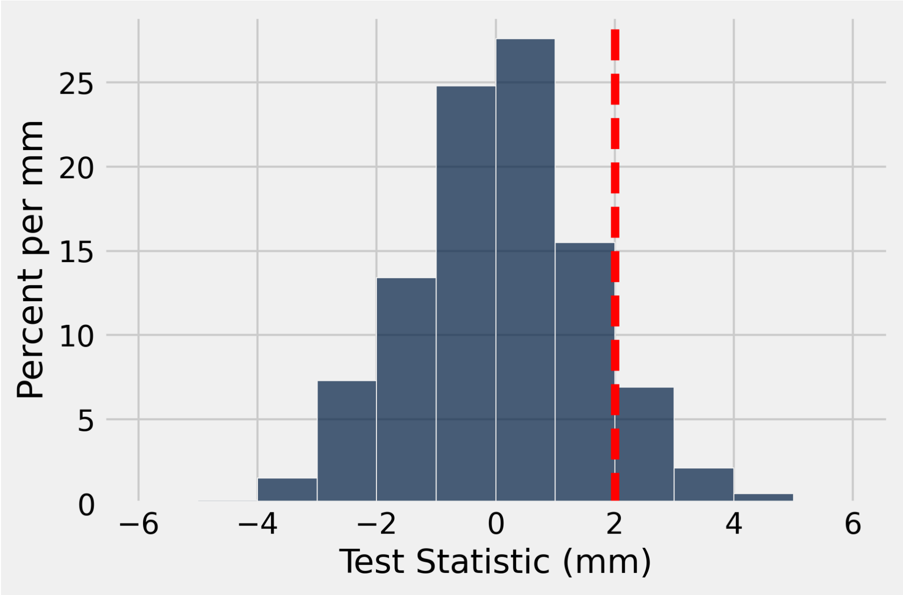
```
def one_simulation():
    shuffled_labels = flies.________________.column(‘Elevation’)

    shuffled_flies =
    flies.drop(‘Elevation’).with_columns(___________,
    ___________)
```

```
        grouped = shuffled_flies.________(_______,_______)

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

def one_simulation():
     shuffled_labels = flies.sample(with_replacement =
     False).column(“Elevation”)
     shuffled_flies =
     flies.drop(‘Elevation’).with_columns(‘Elevation’,
     shuffled_labels)
     grouped = shuffled_flies.group(‘Elevation’, np.mean)
     means = grouped.column(‘Specimen length (mm) mean’)
     statistic =  means.item(0) - means.item(1)
     return statistic
```

f. 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

   test_stats = make_array()
   repetitions = 10000

   for i in np.arange(repetitions):

        one_stat = one_simulation()

        test_stats = np.append(test_stats, one_stat)

   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?



We can inspect the histogram above to see that the area to the **right** of the observed value (which is our **p-value**) is greater than 5%. Since our p-value is greater than our p-value cutoff, we fail to reject the null hypothesis and conclude that the data tend to favor the null hypothesis.

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?

Recall: If you use a p-% cutoff for the p-value, and the null hypothesis happens to be true, then there is about a p-% chance that your test will conclude that the alternative is true.

Using this logic and a 5% cutoff, there's a 5% chance that a given test will incorrectly reject the null. If we run 1000 tests, then we'll expect to incorrectly reject on 1000 * 0.05 = 50 tests.

i. What effect does *decreasing* our p-value cutoff have on the number of times we expect to *incorrectly reject* the null hypothesis?

If we *decrease* our p-value cutoff, we are reducing the expected number of times we'll incorrectly reject the null.