
0.0.1 Question 1c

Before we write any code, let's review the idea of hypothesis testing with the permutation test. It follows the procedure below: 1. We first simulate the experiment many times (say, 10,000 times) using random permutation (i.e., without replacement) (i.e., under the assumption that the null hypothesis is true). This simulated sampling process produces an empirical distribution of many values of a predetermined test statistic (say, 10,000 values). 2. Then, we compare our one true observed test statistic to this empirical distribution of simulated test statistics to compute an empirical p-value. 3. Finally, we compare this p-value to a particular cutoff threshold (often, 0.05) to decide whether we fail to reject the null hypothesis.

In the cell below, answer the following questions: * What does an empirical p-value from a permutation test mean in this particular context of serum cholesterol and having heart disease? * Suppose the empirical p-value is $p = 0.15$, and our p-value cutoff threshold is 0.01. Do we reject or fail to reject the null hypothesis? Why?

An empirical p-value from a permutation test is the proportion of simulated experiments that create a test statistic as extreme or more extreme than our observed statistic when assuming the null hypothesis is true. If the empirical p-vale is 0.15, and our p-value cutoff is 0.01, we fail to reject the null hypothesis. 15% of the simulated experiments showed differences as extreme as our observed statistic, which is higher than the expected 1% cutoff, there is not enough evidence that the average serum cholesterol of patients with heart disease is significantly higher or lower than that of patients without heart disease.

0.0.2 Question 1e

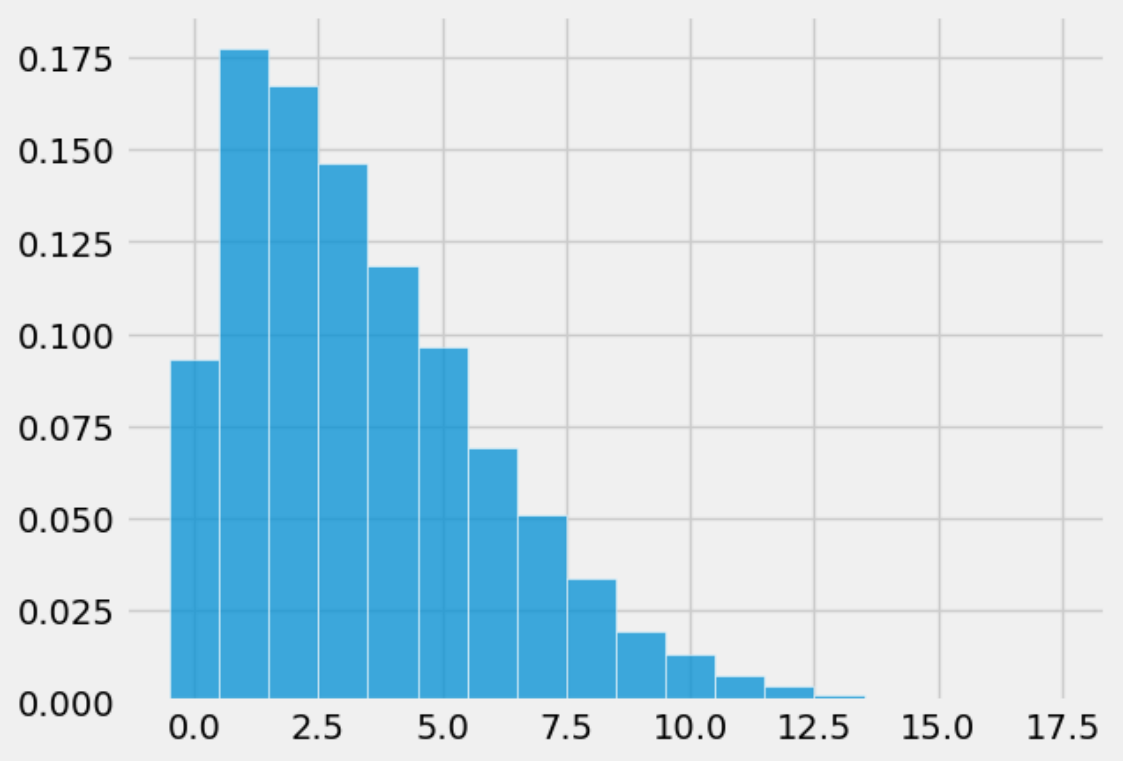
The array `differences` is an empirical distribution of the test statistic simulated under the null hypothesis. This is a prediction about the test statistic, based on the null hypothesis.

Use the `plot_distribution` function you defined in an earlier part to plot a histogram of this empirical distribution. Because you are using this function, your histogram should have unit bins, with bars centered at integers. No title or labels are required for this question.

Hint: This part should be very straightforward.

```
In [42]: plot_distribution(differences)
```

```
Out[42]: (array([9.330e-02, 1.775e-01, 1.672e-01, 1.462e-01, 1.185e-01, 9.650e-02,
                 6.910e-02, 5.110e-02, 3.340e-02, 1.930e-02, 1.320e-02, 7.300e-03,
                 4.600e-03, 1.800e-03, 8.000e-04, 0.000e+00, 1.000e-04, 1.000e-04]),
          array([-0.5,  0.5,  1.5,  2.5,  3.5,  4.5,  5.5,  6.5,  7.5,  8.5,  9.5,
                 10.5, 11.5, 12.5, 13.5, 14.5, 15.5, 16.5, 17.5]),
          <BarContainer object of 18 artists>)
```



0.0.3 Question 1g

Based on your computed empirical p-value, do we reject or fail to reject the null hypothesis? Use the p-value cutoff proposed in Question 1c of 0.01, or 1%.

We reject the null hypothesis that serum cholesterol of non-patients is the same for heart disease patients.