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?

The empirical p-value from a permutation test tells us the probability of getting a test statistics as extreme as our observed one if the null hypothesis were true. In this case, it means how likely we are to observe the difference in cholesterol levels between patients with and without heart disease just by random chance rather than a real effect.

Since the p-value is 0.15, which is higher than the threshold (0.01), we fail to reject the null hypothesis. This means the data does not provide strong enough evidence to conclude that cholesterol levels are significantly different between two groups.

0.0.2 Question 1e

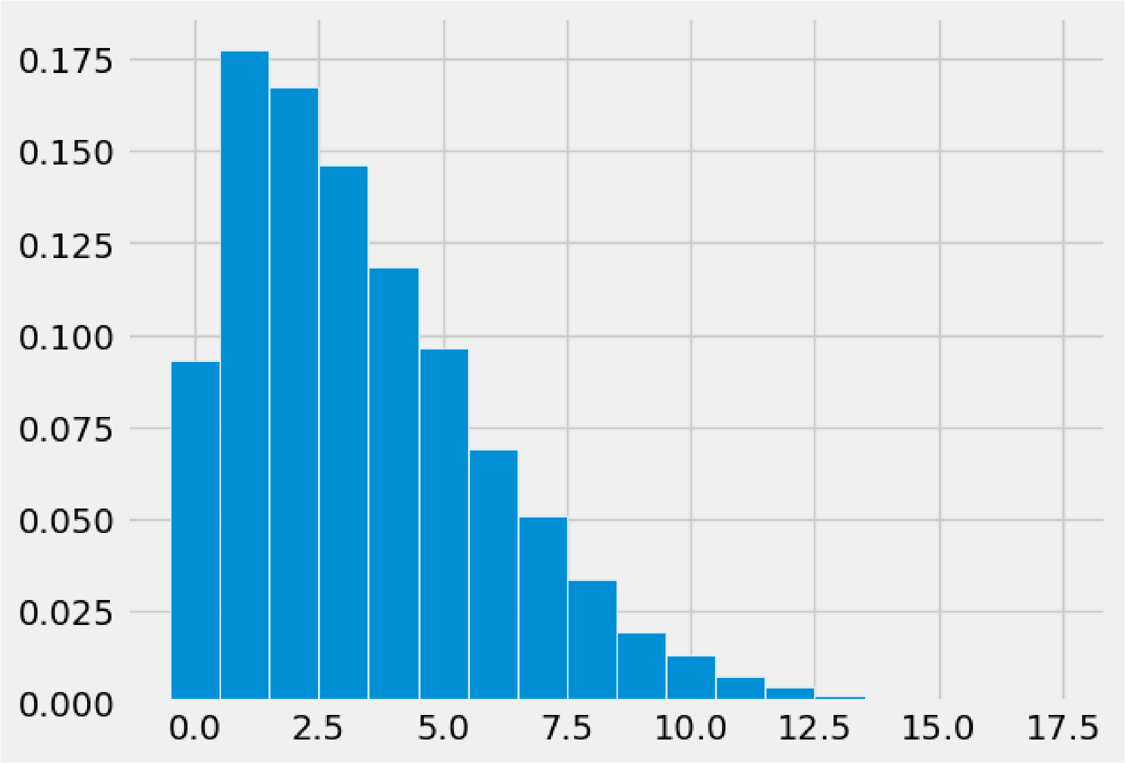
The array `differences` is an empirical distribution of the test statistic simulated under the null hypothesis. This is a prediction about the test statistic, based on the null hypothesis.

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 [26]: def plot_distribution(arr):
             unit_bins = np.arange(round(min(arr)) - 0.5, round(max(arr)) + 1.5, 1)
             plt.hist(arr, bins=unit_bins, density=True, edgecolor="white")
             plt.show()
```

```
In [27]: plot_distribution(differences)
```



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, because p-value is below 0.01