0.0.1 Question 1c

Before we write any code, let's review the idea of hypothesis testing with the permutation test. We first simulate the experiment many times (say, 10,000 times) through random permutation (i.e., without replacement). Assuming that the null hypothesis holds, this process will produce an empirical distribution of a predetermined test statistic. Then, we use this empirical distribution to compute an empirical p-value, which is then compared against a particular cutoff threshold in order to accept or reject our null hypothesis.

In the below cell, answer the following questions: * What does an empirical p-value from a permutation test mean in this particular context of birthweights and maternal smoking habits? * Suppose the resulting empirical p-value $p \leq 0.01$, where 0.01 is our p-value cutoff threshold. Do we accept or reject the null hypothesis? Why?

The empirical p-value of the test is the proportion of simulated differences that were equal to or less than the observed difference i.e. difference between the average birth weights of the two groups (i.e., the babies of non-smokers and the babies of mothers who smoke).

The empirical p-value is less 1% and therefore the result is statistically significant. We reject the null hypothesis. The test supports the hypothesis that the average birth weights of the babies of smokers were lesser on average.

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.

0.0.3 Question 1e

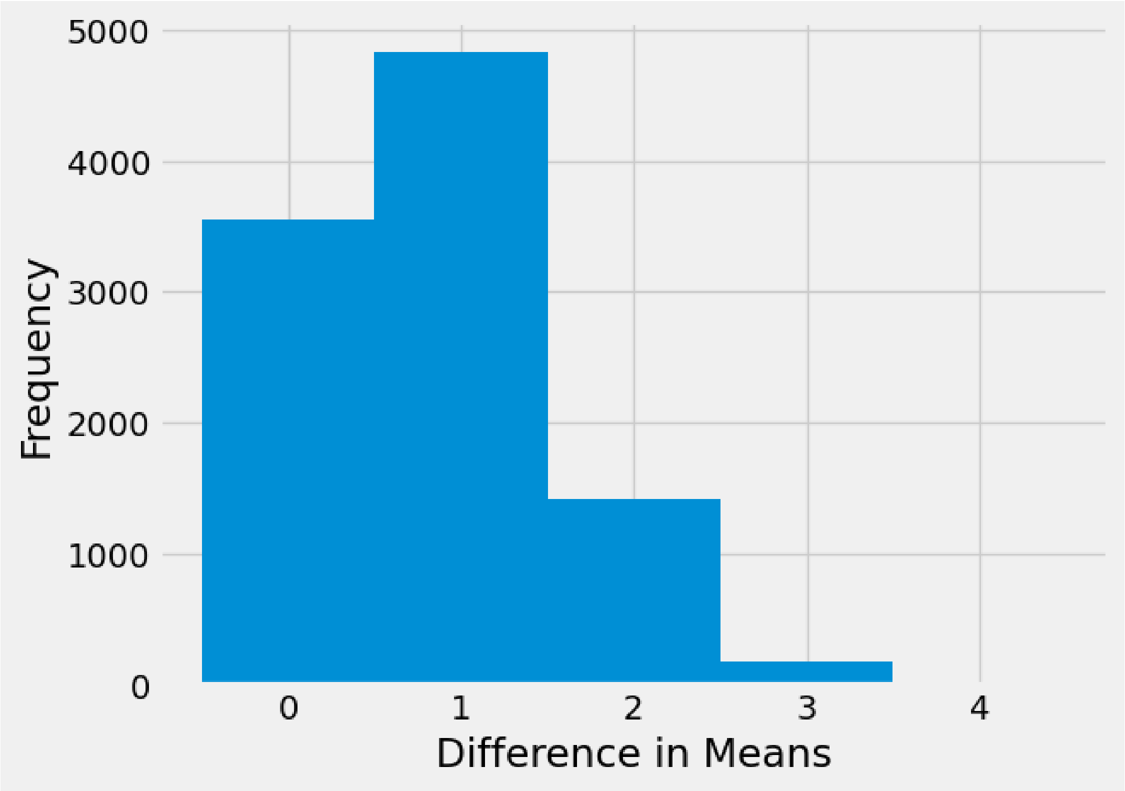
The array differences is an empirical distribution of the test statistic simulated under the null hypothesis. This is a prediction about the test statistic, based on the null hypothesis.

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 [32]: def plot_distribution(differences):
             plt.hist(differences, bins=np.arange(min(differences), max(differences) + 1) - 0.5)
             plt.xlabel('Difference in Means')
             plt.ylabel('Frequency')
             plt.show()

         # Plot the empirical distribution
         plot_distribution(differences)
```


5000
4000
3000
2000
1000
0
Frequency
0
1
2
3
4
Difference in Means

0.0.4 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.