0.1 QUESTION 1 - Scientists vs. P-Values

Read the following article AND watch the following video. Then answer the following questions below.

Step 1). Read the following article from **FiveThirtyEight**: Statisticians Found One Thing They Can Agree On: It's Time to Stop Misusing P-Values

Step 2). Watch this video (11 min): P-Hacking

Based on the article:

Question 1.1. In what ways are scientists misusing p-values? For full credit list **at least 3 ways** mentioned in the article.

Question 1.2. What suggestions are made in the article to use them properly?

Based on the video:

Question 1.3. Suppose the null hypothesis is true. If you're conducting multiple hypothesis tests at the 5% significance level, what's the minimum number of tests you need to do before it's more than 50% likely that at least one of the tests will incorrectly reject the null hypothesis? Show work justifying your answer.

Question 1.4. What is the Bonferroni correction as described in the video? Give an example from the video as to how it could be used.

Answer all 4 parts in the same Markdown cell below:

1.1: One way that scientists are misusing p-values is that they believe p-values prove your hypothesis. This is not true. Instead, p-value is the probability of getting the data you observed based on your hypothesis. Another way that scientists are misuing p-values is that it tells you the probabilty of the results being true. Finally, scientists were misusing p-values to separate false findings from true findings by cherry picking their p-value to fit their data.

1.2: The ASA recommended to rely on other measures like confidence intervals, but the overall goal is to move towards embracing uncertainty and variation.

1.3: Since each test is using the p-value of 5%, there is a 95% chance the data is not significant. Thus, the minimum number of tests you need to do before it's more than 50% likely that at least one of the tests will incorrectly reject the null is 13 tests. This is because the probability of having none of the 14 tests come up significant is $0.95^{14} = .4877$ and $(1 - .4877) * 100 = 51.23$ of the time one or more of the tests will reject the null by chance.

1.4: The Bonferroni correction is instead of using the usual significance values to decide when results are significant or not, divide the significant value by the number of tests and use that as your p-value.

```
In [3]: 1-.95**14
```

```
Out[3]: 0.5123250208844705
```

Question 2.1. Suppose we want to test whether or not each factor contributes the same amount to the overall Happiness Score. Define the null hypothesis, alternative hypothesis, and test statistic in the cell below.

Note: Please format your answer as follows: - Null Hypothesis: ...
- Alternative Hypothesis: ...
- Test Statistic: ...

- Null Hypothesis: Each factor contributes the same amount to the overall Happiness Score
- Alternative Hypothesis: Each factor does not contribute the same amount to the overall Happiness Score
- Test Statistic: Total Variation Distance

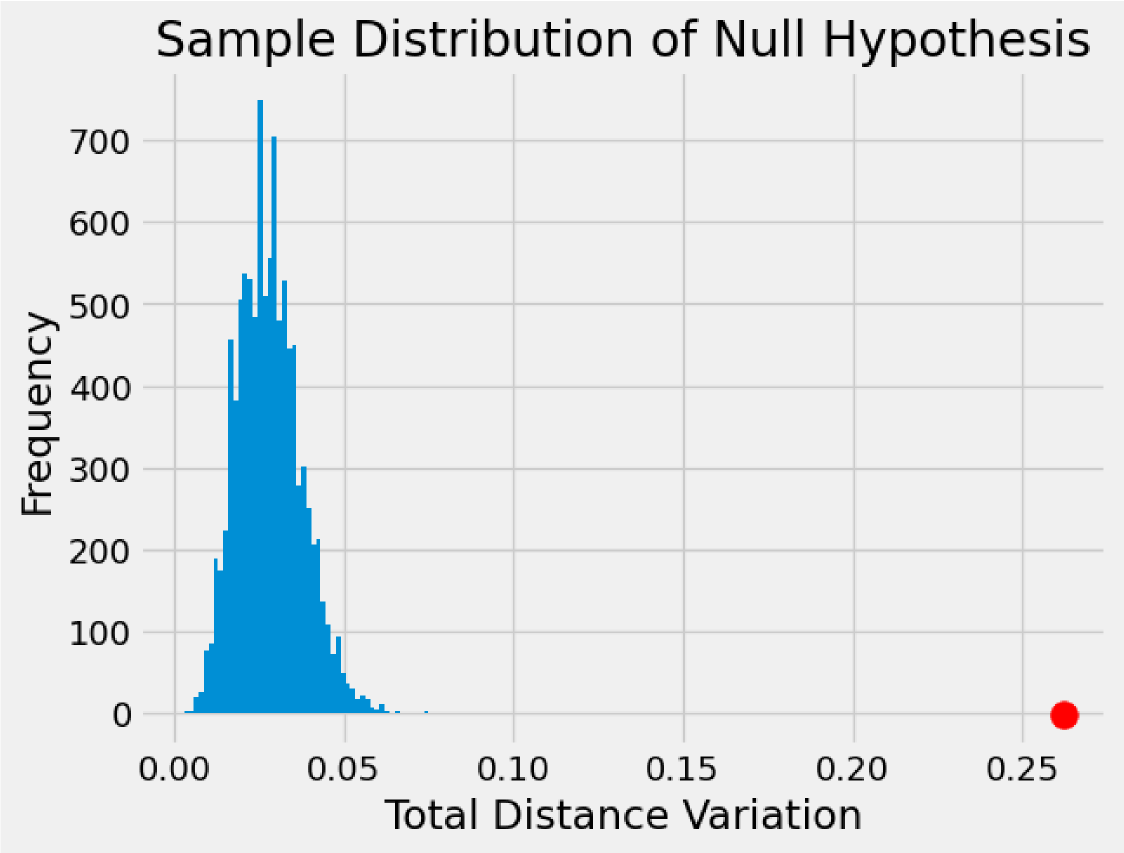
Question 2.3. Create an array called `simulated_tvds` that contains 10,000 simulated values under the null hypothesis. Assume that the original sample consisted of 1,000 individuals.

Then plot a density histogram of your simulated test statistics, as well as a red dot representing the observed value of the test statistic. Include a title and label your x and y axes.

```
In [34]: simulated_tvds = np.array([])
         for i in range (10000):
             sample = np.random.multinomial(1000, null_distribution);
             sample_proportions = sample/1000
             simulated_tvds = np.append(simulated_tvds, calculate_tvd(sample_proportions, \
                                                                        null_distribution))

         plt.hist(simulated_tvds, bins=50)
         # your code for the histogram above here.  The code below plots a red marker at the mean:
         plt.scatter(observed_tvd, -1, marker='.', color='red', s=500)
         plt.title("Sample Distribution of Null Hypothesis")
         plt.ylabel("Frequency")
         plt.xlabel("Total Distance Variation")
         # your code above this line
```

```
Out[34]: Text(0.5, 0, 'Total Distance Variation')
```



```
In [10]: grader.check("q2_3")
```

```
Out[10]: q2_3 results: All test cases passed!
```

Question 2.5. What can you conclude about how each factor contributes to the overall happiness score in the US? Explain your answer using the results of your hypothesis test. Assume a significance level (i.e. p-value cutoff) of 5%.

We can conclude that each factor does not contribute the same amount to the overall Happiness Score because the empirical p-value is 0 which is less than the significance level of 5%.

0.2 QUESTION 3: A/B Tests

Answer all 4 parts to this question in the same Markdown cell below.

Question 3.1. When should you use an A/B test versus another kind of hypothesis test?

Question 3.2. Kevin, a museum curator, has recently been given specimens of caddisflies collected from various parts of Colorado. 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.

- **Question 3.2.a** What's an appropriate null hypothesis that Kevin can simulate under?
- **Question 3.2.b** 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?
- **Question 3.2.c** 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.

3.1: You should use an A/B test when you are comparing 2 samples compared to other hypothesis tests that are looking at 1 sample

3.2a: An appropriate null hypothesis that Kevin can simulate under is that there is no difference in size between caddisflies collected at higher altitudes and caddisflies collected at lower altitudes.

3.2b: To test the null hypothesis in the A/B test, you assume the distribution of both caddisflies at higher elevation and at lower elevation are the same and would use random permutations to simulate under the null hypothesis.

3.2c: A useful test statistic for the A/B test is difference between average length. If the difference is positive then it would mean higher elevation caddisflies are truly bigger.

Question 4.8.

In the first cell below:

- Define a function `simulate_one_statistic` that takes no arguments and returns one simulated value of the test statistic. Refer to the code you have previously written in this problem, as you might be able to re-use some of it.

In the 2nd cell below:

- Complete the code to simulate 10,000 values of the statistic and store it in the array `simulated_statistics_ab`.
- Then draw a density histogram with the empirical distribution of the statistic
- Include a red dot on your histogram at the value of `observed_statistic_ab`.
- Include a title for your histogram and label the x and y-axes.

```
In [35]: def simulate_one_statistic():
             shuffled_labels = football["Team"].sample(frac=1, replace=False).values
             original_and_shuffled["ShuffledLabel"] = shuffled_labels

             patriots_shuffled = original_and_shuffled[original_and_shuffled["ShuffledLabel"] == \
                                                       "Patriots"]
             colts_shuffled = original_and_shuffled[original_and_shuffled["ShuffledLabel"] == "Colts"]

             avg_drop_for_colts = colts_shuffled["PressureDrop"].mean()
             avg_drop_for_patriots = patriots_shuffled["PressureDrop"].mean()

             return avg_drop_for_patriots - avg_drop_for_colts

         # Your code above this line


         simulate_one_statistic()
```

```
Out[35]: 0.060227272727273684
```

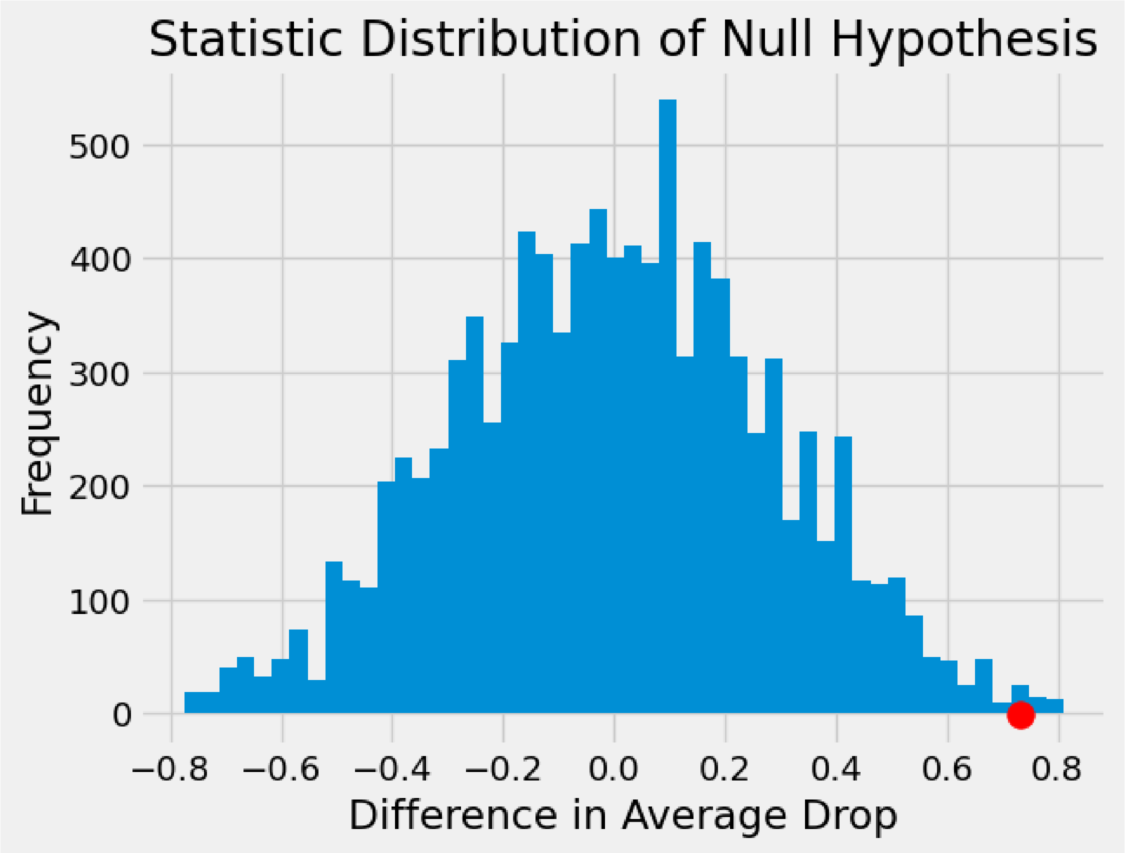
```
In [29]: repetitions = 10000

         simulated_statistics_ab = np.array([])

         for i in range(repetitions):
             stat = simulate_one_statistic()
             simulated_statistics_ab = np.append(simulated_statistics_ab, stat)
```

```
plt.hist(simulated_statistics_ab, bins=50)
plt.scatter(observed_statistic_ab, -1, marker='.', color='red', s=500)
plt.title("Statistic Distribution of Null Hypothesis")
plt.ylabel("Frequency")
plt.xlabel("Difference in Average Drop")
# your code for histogram and observed statistic above this line
```

```
Out[29]: Text(0.5, 0, 'Difference in Average Drop')
```



Question 4.10. What is the conclusion of your test? Explain what this means in the context of this particular problem. Can we make any casual conclusions from this test? Why or why not?

We can conclude that we reject the null hypothesis and accept the alternative hypothesis. This means that the Patriots' pressure drops are too large, on average, to resemble a random sample drawn from all the drops because the empirical p-value of .3% is less than the 5% significance cuttoff for the p-value. However, we cannot make any causal conclusions from this test because teams were not randomly assigned balls.