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:

Q1.1 Scientists may misuse p values through p-hacking, misinterpreation of pvalues, and overreliance on pvalues to see the size of an effect. Through p-hacking, researchers may perform multiple tests and selectively report only that produces statsistacally significant results, while not reporting the others that didn't show significance. In misinterpreation of pvalues, they may use pvalues as a direct measure of importance or proabbiltiy of a hypothesis being true. Finally, some researchers may rely soley on p-values to make decisions without considering other statisical information.

Q1.2 To use p-values properly, the article suggests that researchers must report effect size effect with p-values. This provides information about the significance of the results, and helps determine if a result is statistically significant. Another way way is to not use p-values to determine if the hypothesis is correct or incorrect. They shouldn't be used as a base to determine conclusions. Finally, the misinterpretation of

p-values and misuse of p-values needs to be fixed, and needs to be understood correctly.

Q1.3 In null hypothesis for a single test, the probability of not incorrectly rejcting the null hypothesis is 1 - 0.05 = 0.95. There are n number of tests. And we want to see the minimum number of tests occured that it is 50% likely that at least one of the tests will incorrectly reject the null hypothesis. Thus we can use the equation:

$(0.95)^n > 0.5$

$n \cdot ln(0.95) > ln(0.5)$

$n > (\frac{ln(0.5)}{ln(0.95)})$

$n > 13.5134$

Thus it will take 14 number of tests.

Q1.4 The Bonferroni correction is a method to address the problem of multiple comparisons by adjusting the significance level for individual tests to maintain a desired overall significance level.

For example, if we are conducting 10 tests and you want to maintain an overall significance level of 5%, you would adjust the significance level for each individual test to 0.5% (0.05 / 10). If any individual test yields a p-value less than 0.5%, it would be considered statistically significant. This correction helps control the family-wise error rate, ensuring that the overall probability of making at least one Type I error remains low.

In [ ]:

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: The contribution of each factor to the overall Happiness Score is the equal.
- Alternative Hypothesis: The contribution of at least one factor the overall Happiness score is different.

- Test Statistic: TVD measures the difference between the expected distrubution and observed distribution of proportions to each facotr in the Happiness score.

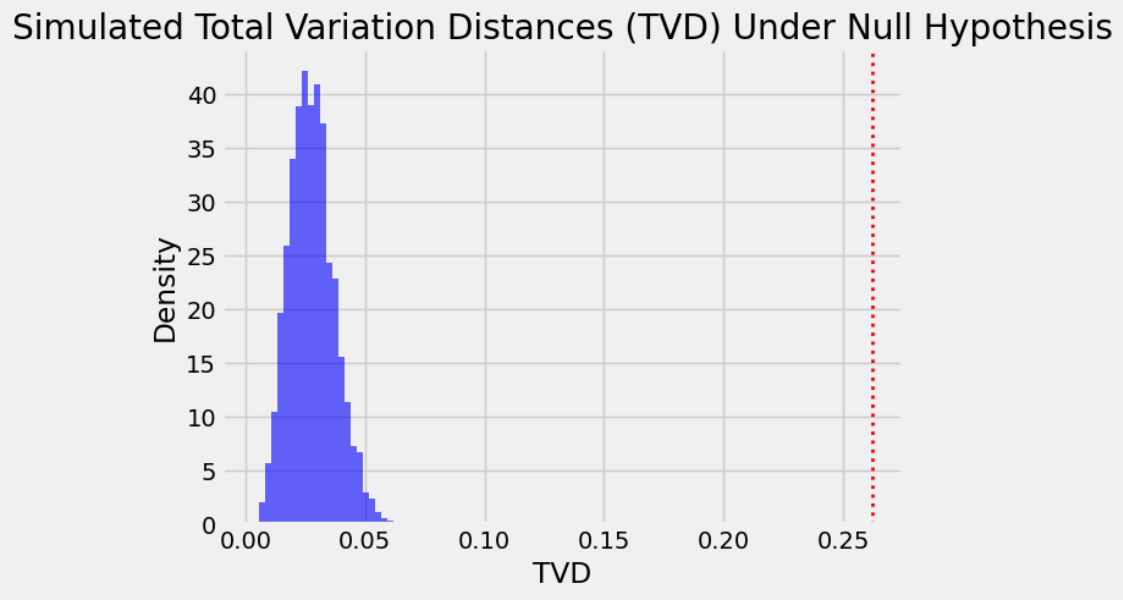
Question 2.3. Create an array called `simulated_tvds` that contains 10,000 simulated values under the null hypothesis. Assume that the original sample consisted of 1,000 individuals.

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 [8]: simulated_tvds = []
        sample = 1000
        for i in range(10000):
            s = np.random.multinomial(sample, null_distribution) / 1000
            simulated_tvds.append(calculate_tvd(s, null_distribution))

        simulated_tvds = np.array(simulated_tvds)

        plt.hist(simulated_tvds, bins=30, density=True, color='blue', alpha=0.6)
        plt.axvline(observed_tvd, color='red', linestyle='dotted', linewidth=2)
        plt.title('Simulated Total Variation Distances (TVD) Under Null Hypothesis')
        plt.xlabel('TVD')
        plt.ylabel('Density')
        plt.show()
```




```
In [9]: grader.check("q2_3")
```

```
Out[9]: 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%.

Based on the significance level of 5%, we can conclude that not all factors contribute equally to the overall happiness score in the United States.

Since the pvalue is less than 5%, we can reject the null hypothesis, that each factor contributes the same amount to the overall happiness score, as we can conclude that at least 1 factor contributes to the overall happiness score differently than others.

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

We use A/B testing when we want to compare two different versions of something to determine which is better. They are used when you have a change to the current original version, and determine whether or not it has a significant impact on a specific outcome.

3.2

a) The population mean length of caddisfilies collected at high and low elevations should be equal.

b) We could test the null hypothesis from above by collecting and recording data of the lengths of caddisflies of high and low elevations. We already defined the null hypothesis. But the alternative hypothesis is: The population mean length of caddisflies collected at high elevations are greater length than the ones in lower elevations. We then use the permutation test to test the null when we shuffle. We first calculate the mean length of caddisflies at higher altitudes, then randomly shuffle out data and select without replacement. Then we do the same for the lower altitude. We see that the higher altitude is our observed statistic. We then take abs(shuffed_high_alt - shuffeled_low_alt). We choose the appropriate amount of significance level, and determine the threshold for statsical significance, like (0.05). Then we calculate the means and standard deviations of both group's samples. by sum(data > observed data)/num simulations. From this data we determine whether or not to reject the null.

c) A good test statistic would be calculating the absolute value of difference in means between the high altitude flies and low altitude flies.

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 [27]: def simulate_one_statistic():
             shuffled_pressure_drops = np.random.permutation(football['PressureDrop'])
             colts_mean = np.mean(shuffled_pressure_drops[:4])  # First 4 values are Colts
             patriots_mean = np.mean(shuffled_pressure_drops[4:])  # Last 11 values are Patriots
             return patriots_mean - colts_mean

         # Your code above this line


         simulate_one_statistic()
```

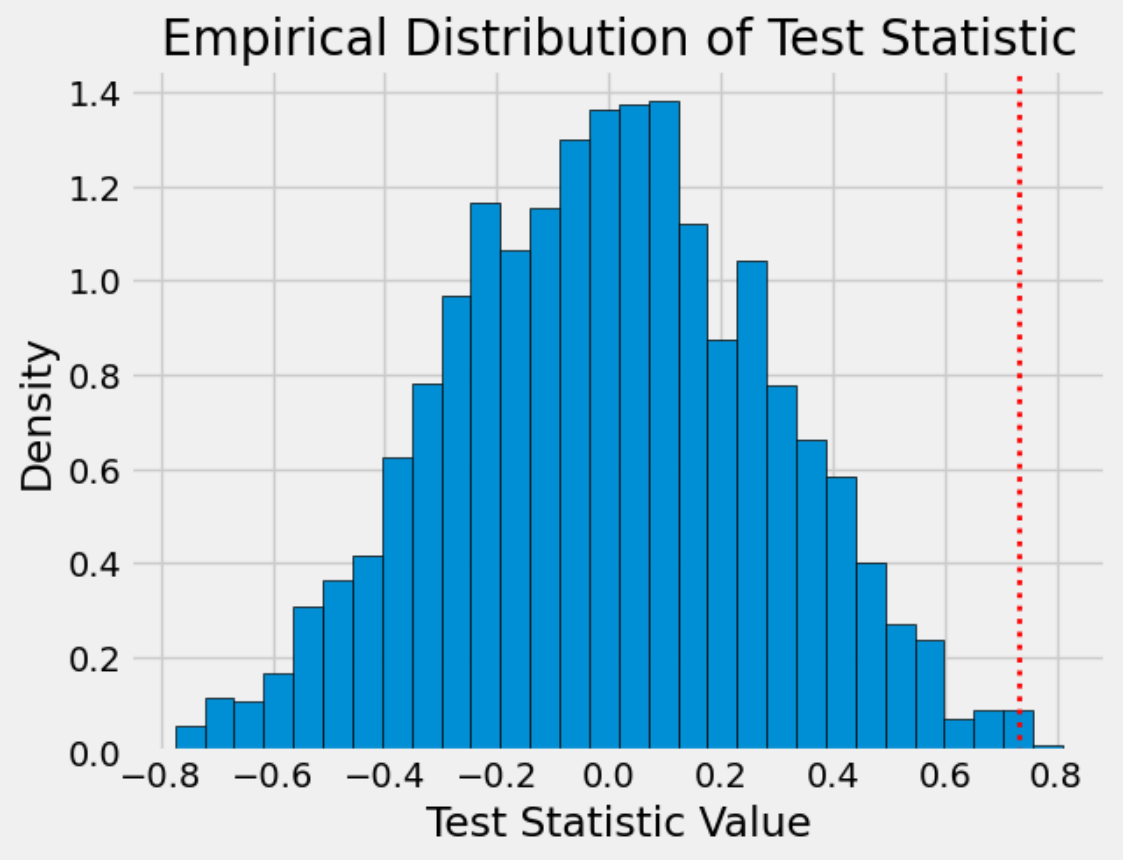
```
Out[27]: 0.043181818181817544
```

```
In [28]: repetitions = 10000
         simulated_statistics_ab = []

         for _ in range(repetitions):
             statistic = simulate_one_statistic()
             simulated_statistics_ab.append(statistic)

         plt.hist(simulated_statistics_ab, density=True, bins=30, edgecolor = 'k')
         plt.axvline(observed_statistic_ab, color='red', linestyle='dotted', linewidth=2)
         plt.title('Empirical Distribution of Test Statistic')
         plt.xlabel('Test Statistic Value')
         plt.ylabel('Density')
         plt.show()

         # your code for histogram and observed statistic above this line
```


Empirical Distribution of Test Statistic
Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Test Statistic Value

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 see that the observed difference in pressure drop between the Patriots's and Colt's footballs is statistically significant at a 5% significance level, thus we can reject the null hypothesis. We can say that on average, the Patriots have a larger drop in pressure compared to Colt's football, and the difference is unlikely due to chance. However, the observed difference could be due to various factors, and the test only tells us that there is a statistical difference but doesn't show the causes of the difference. We will need to know more in order to identify the reaons for the observed difference in pressure drop between the two team.