DSC 10 Practice Final Exam From Spring 2021

This was the final exam for DSC 10 in Spring 2021. This was a 3 hour exam. Students were instructed to do certain problems in their own Jupyter notebook.

This practice exam is different in format to the final you'll take, but it might still be good practice. We recommend primarily focusing on the practice final from Winter 2021 available on Gradescope because it's more similar in style and format to the exam you'll take.

For the real exam, you can reference anything besides other people, so we recommend taking this exam in the same way you'll take the real exam, without collaborating or communicating with anyone else. We also recommend having the DSC 10 Reference Sheet open while you take the exam. You can find that reference sheet on the course website, in the Resources section.

Question 1

A toy company sells stuffed animals in the form of giraffes, gorillas, and ponies. The store would like to sell the toys in equal numbers. In the U.S. the company sold 1,000 toys in total with the proportions 11% - giraffes, 37% - gorillas, and 52% - ponies. Meanwhile, the California proportions are 15% - giraffes, 40% - gorillas, and 45% - ponies.

Question:

The store would like to know if the California proportions are significantly different from uniform. You've been told to use the maximum proportion as your sample statistic. What values of the sample statistic would you consider as evidence in favor of rejecting the null hypothesis?

- ○ Lower than ⅓
- ● Higher than ⅓
- ○ Higher or lower than ⅓
- ○ It's impossible to tell

Question 2

A toy company sells stuffed animals in the form of giraffes, gorillas, and ponies. The store would like to sell the toys in equal numbers. In the U.S. the company sold 1,000 toys in total with the proportions 11% - giraffes, 37% - gorillas, and 52% - ponies. Meanwhile, the California proportions are 15% - giraffes, 40% - gorillas, and 45% - ponies.

Question:

The store would like to know if the proportion of gorillas is significantly different from the proportion of ponies. What test statistic could you use to test this hypothesis?

- ○ The TVD
- ○ The difference between the proportion of gorillas sold and the proportion of ponies sold.
- ○ The ratio of the proportion of gorillas sold and the proportion of ponies sold.
- ● All of these

Question 3

A toy company sells stuffed animals in the form of giraffes, gorillas, and ponies. The store would like to sell the toys in equal numbers. In the U.S. the company sold 1,000 toys in total with the proportions 11% - giraffes, 37% - gorillas, and 52% - ponies. Meanwhile, the California proportions are 15% - giraffes, 40% - gorillas, and 45% - ponies.

Question:

The store would like to know if the California proportions are significantly different from the national proportions. What is the model for the null hypothesis?

- (●) 11% - giraffes, 37% - gorillas, and 52% - ponies
- ( ) 1/3 - giraffes, 1/3 - gorillas, and 1/3 - ponies
- ( ) 15% - giraffes, 40% - gorillas, and 45% - ponies
- ( ) None of these

Question 4

A toy company sells stuffed animals in the form of giraffes, gorillas, and ponies. The store would like to sell the toys in equal numbers. In the U.S. the company sold 1,000 toys in total with the proportions 11% - giraffes, 37% - gorillas, and 52% - ponies. Meanwhile, the California proportions are 15% - giraffes, 40% - gorillas, and 45% - ponies.

Question:

What test statistic could you use to test the null hypothesis that the California stores sold the toys in the same proportions as the U.S. as a whole?

- (●) The total variation distance
- ( ) The average difference between the proportions.
- ( ) The sum of the proportions
- ( ) None of these

Question 5

A toy company sells stuffed animals in the form of giraffes, gorillas, and ponies. The store would like to sell the toys in equal numbers. In the U.S. the company sold 1,000 toys in total with the proportions 11% - giraffes, 37% - gorillas, and 52% - ponies. Meanwhile, the California proportions are 15% - giraffes, 40% - gorillas, and 45% - ponies.

Question:

In your own notebook, using the np.random.multinomial() function, determine if there is sufficient evidence to reject the null hypothesis that the national proportions are uniform. Choose the best answer.

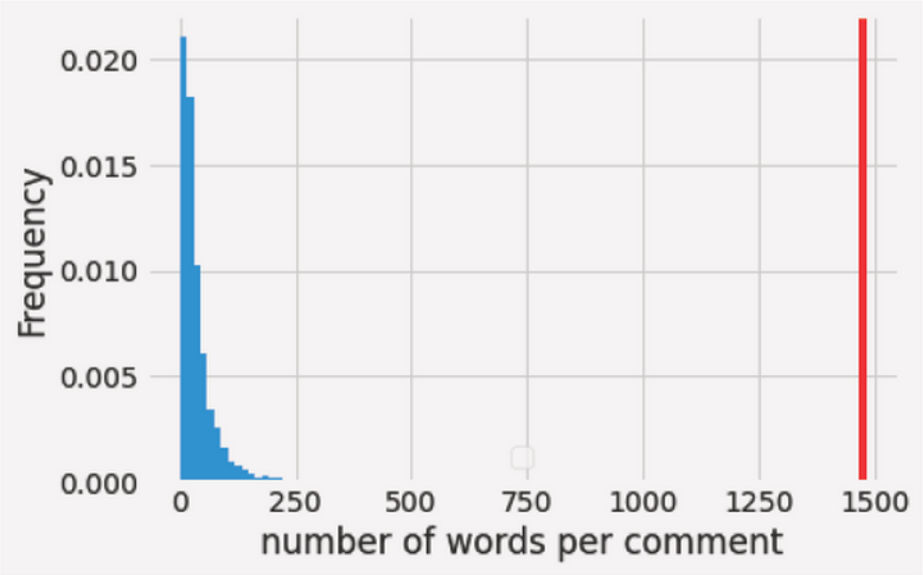
- ○ No, the P-value is <95%
- ○ No, the P-value is >95%
- ● Yes, at the 99% level of significance
- ○ Yes, at the 95% level of significance

Question 6

You want to construct a confidence interval for the sample median of a random variable that has values anywhere between 0 and 1. Which of the following are valid for doing so? Choose all that apply

- [x] Draw bootstrap samples and use np.percentile()
- [ ] Invoke the central limit theorem and use p*(1-p) as the population standard deviation)
- [ ] Draw bootstrap samples and use the sample mean of the medians plus and minus 2 standard deviations
- [ ] Draw samples using np.random.multinomial() and convert to proportions

Question 7



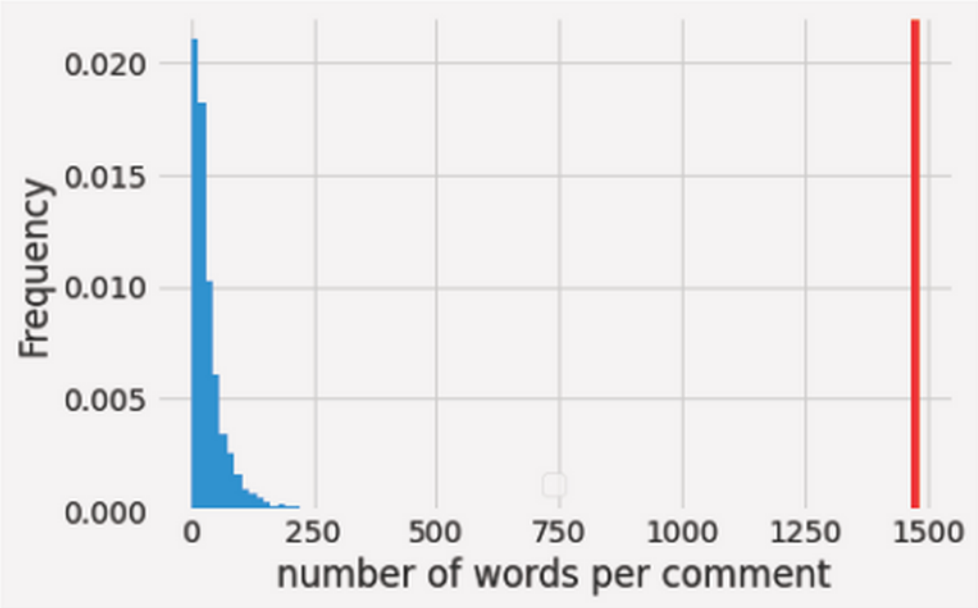
This is a histogram of comment lengths (in words) from a sample of 5000 comments randomly sampled from YouTube comments. The population mean of this distribution is 39.065. The sample mean is 39.45. The population standard deviation is 49.45. The red line indicates the maximum of this sample.

Question:

Which of the following is true of the sample mean of this distribution

- ○ It is a parameter of the distribution
- ● It is a statistic.
- ○ It is less than the sample median.
- ○ All of these
- ○ None of these

Question 8

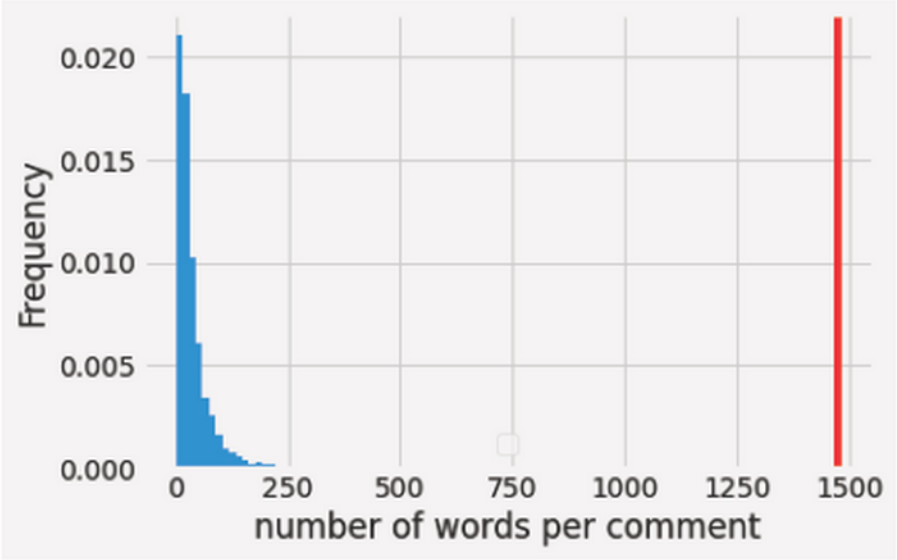


This is a histogram of comment lengths (in words) from a sample of 5000 comments randomly sampled from YouTube comments. The population mean of this distribution is 39.065. The sample mean is 39.45. The population standard deviation is 49.45. The red line indicates the maximum of this sample.

Question:
Which of the following would you expect to be true of the population distribution?

- ○ The population mean is less than the population median
- ○ The sample mean is less than the sample median
- ○ The sample mean is less than the population median
- ● None of these

Question 9



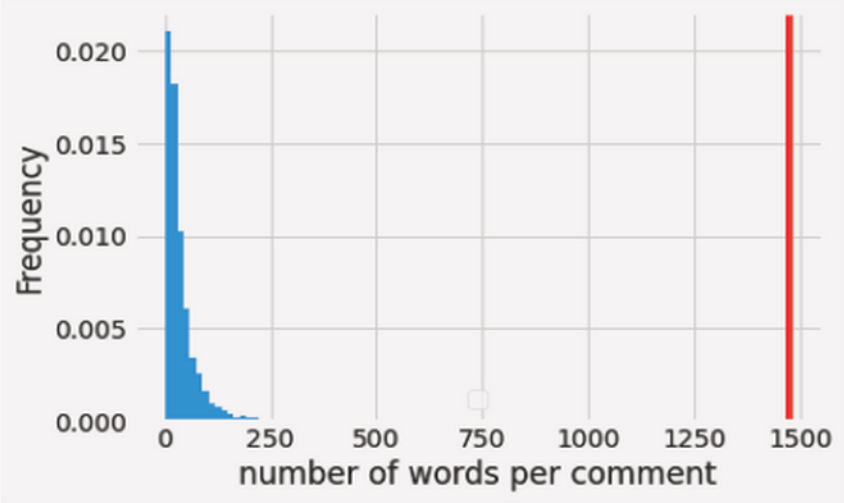
This is a histogram of comment lengths (in words) from a sample of 5000 comments randomly sampled from YouTube comments. The population mean of this distribution is 39.065. The sample mean is 39.45. The population standard deviation is 49.45. The red line indicates the maximum of this sample.

Question:

We will find 96% of the data no more than z standard deviations from the mean. What does z equal?

- ○ 2
- ● 5
- ○ We can't know from this information
- ○ None of these

Question 10



This is a histogram of comment lengths (in words) from a sample of 5000 comments randomly sampled from YouTube comments. The population mean of this distribution is 39.065. The sample mean is 39.45. The population standard deviation is 49.45. The red line indicates the maximum of this sample.

Question:

Suppose you are given these data in a dataframe called `words` with a single column named `NumWords`. You then run the following code:

```
n_resamples = 5000

boot_medians = np.array([])
for i in range(n_resamples):

    resample = words.sample(5000, replace=True)
    boot_median = resample.get('NumWords').median()
    boot_medians = np.append(boot_medians, boot_median)
```

Which of the following are true about the result?

- (●) You can use array boot_medians to construct a confidence interval over the median.
- ( ) It is valid to invoked the central limit theorem instead of sampling for this statistic.
- ( ) You would have gotten an error.
- ( ) None of these

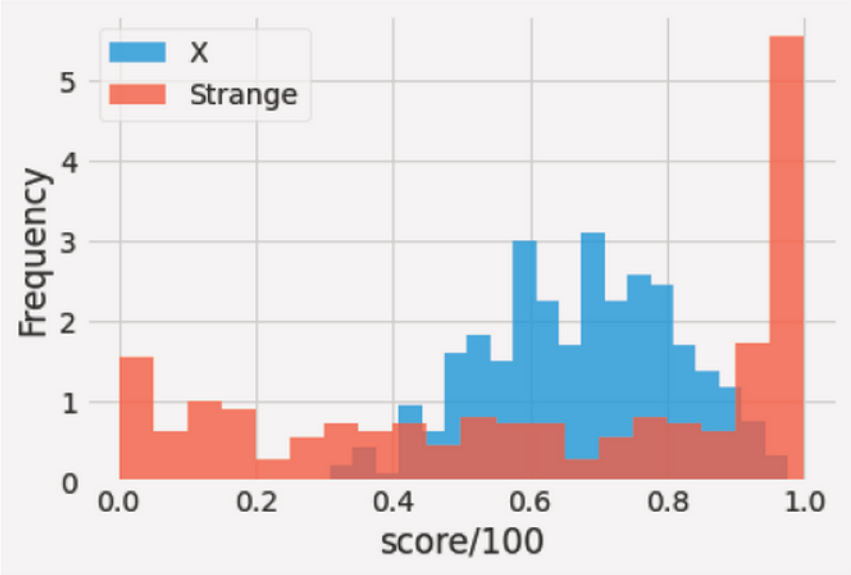
Question 11

You toss a coin 500 times with the hope of determining if the coin is fair. You want to make sure that you can detect if the coin deviates as little as 1% from fair. Have you tossed this coin enough times? *(hint: You will need to use the central limit theorem)*

- ○ No, you will need to toss at least 2000 times
- ● No, you will need to toss at least 10,000 times
- ○ Yes
- ○ None of these

Question 12

The 2 worst teachers for DSC10 in previous quarters were Professor X and Dr. Strange. The two instructors took very different strategies to teaching the class. Professor X went a little too slow for the most advanced students but made sure everyone was learning. Dr. Strange only paid attention to his brightest students. In one quarter, there were 500 students that took the course. 280 of the students were enrolled in Professor X's sections. The rest were enrolled in Dr. Strange's section. The final exam scores for both professors were arranged in a dataframe called scores, which has 2 columns (section, and score). You have been asked to determine whether the scores were statistically different for the two teachers. The average test score for Professor X was 66.9% and the average score for Dr. Strange was 61.4%. A histogram of the exam scores is below.



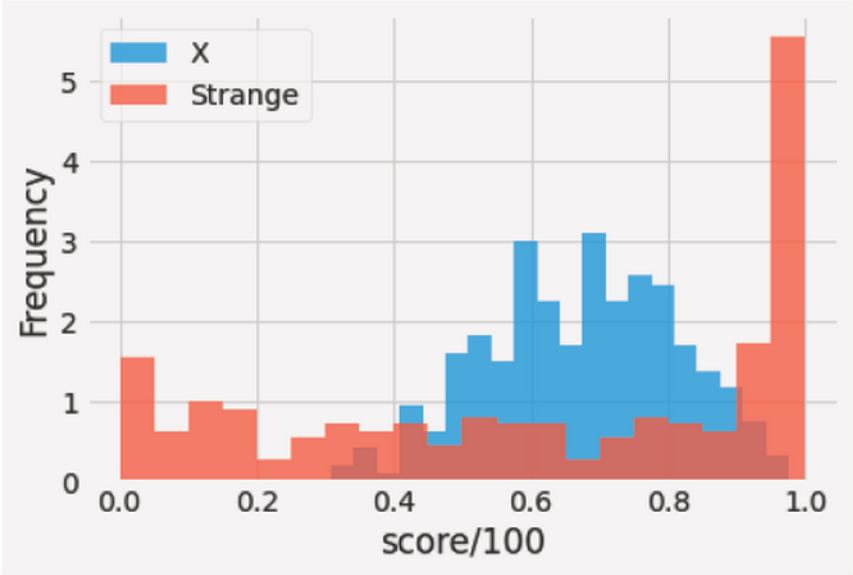
Question:

Which of the following would be true of a permutation test for these data?

- ○ It would average out all of the confounding variables.
- ○ It would prove whether Dr. Strange's teaching strategy was as effective as Professor X's.
- ○ It would prove whether Dr. Strange is a worse teacher than Professor X.
- ● None of these

Question 13

The 2 worst teachers for DSC10 in previous quarters were Professor X and Dr. Strange. The two instructors took very different strategies to teaching the class. Professor X went a little too slow for the most advanced students but made sure everyone was learning. Dr. Strange only paid attention to his brightest students. In one quarter, there were 500 students that took the course. 280 of the students were enrolled in Professor X's sections. The rest were enrolled in Dr. Strange's section. The final exam scores for both professors were arranged in a dataframe called scores, which has 2 columns (section, and score). You have been asked to determine whether the scores were statistically different for the two teachers. The average test score for Professor X was 66.9% and the average score for Dr. Strange was 61.4%. A histogram of the exam scores is below.



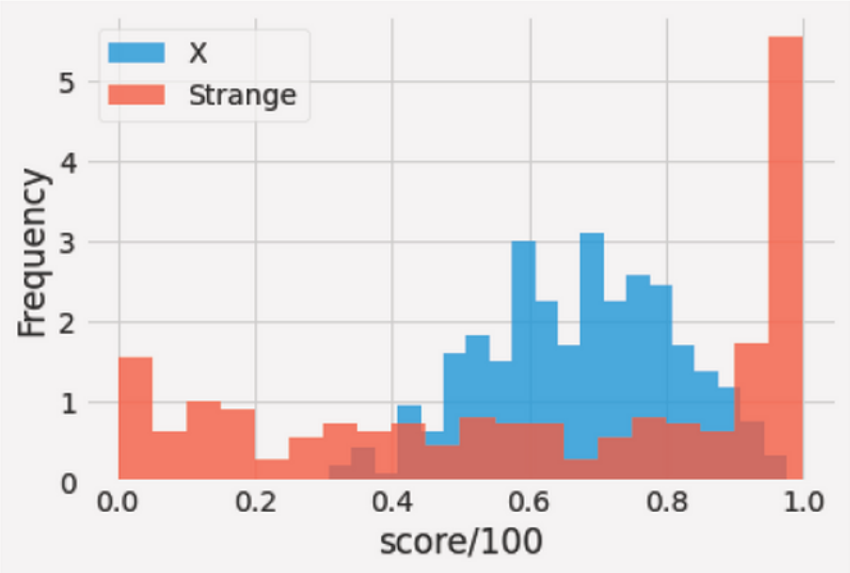
Question:

You decide to test the null hypothesis by sampling. Which of the following are an appropriate way to sample from the distribution corresponding to the null hypothesis?

- (●) Permutation testing (shuffling either the scores or the labels)
- ( ) Permutation testing (shuffling the student's scores)
- ( ) Permutation testing (Shuffling the class label)
- ( ) Using np.random.multinomial(sample_size, pop_distribution)

Question 14

The 2 worst teachers for DSC10 in previous quarters were Professor X and Dr. Strange. The two instructors took very different strategies to teaching the class. Professor X went a little too slow for the most advanced students but made sure everyone was learning. Dr. Strange only paid attention to his brightest students. In one quarter, there were 500 students that took the course. 280 of the students were enrolled in Professor X's sections. The rest were enrolled in Dr. Strange's section. The final exam scores for both professors were arranged in a dataframe called scores, which has 2 columns (section, and score). You have been asked to determine whether the scores were statistically different for the two teachers. The average test score for Professor X was 66.9% and the average score for Dr. Strange was 61.4%. A histogram of the exam scores is below.



Question:

Suppose you randomly selected the section that each student was enrolled in before the start of the quarter. What effect would this have on the results of the study?

- ○ You would cancel out all effects except for those factors related to the teachers.
- ○ You would cancel-out any factors that could lead better or worse students to choose Dr. Strange or Professor X.
- ○ You would cancel-out (on average) factors that could bias the results
- ● All of these

Question 15

You are handed data with pairs of data points (x and y) and you calculate the least-squares regression line. You find that the slope is close to zero. Which of the following CANNOT be a culprit for why this is the case.

- [x] We need to convert to standard units to get the correct correlation.
- [ ] The correlation is zero
- [ ] The relationship between x and y is nonlinear with no trend.
- [ ] There could be an outlier in your data that pulls the slope toward zero.

Question 16

You been given the equation of a regression line to predict *x* from *y* in standardized units. Which of the following are true?

- [ ] The slope is 1
- [ ] The error in predicting of y from x is equal to the regression coefficient
- [x] The y-intercept is zero
- [ ] All of these

Question 17

You've been given the equation of a least-squares regression line. What do the residuals represent?

- (●) The errors in predicting y from x.
- ( ) The amount you need to adjust the line to better fit the data.
- ( ) The distance along the x-axis from the data to the line.
- ( ) All of these