Quiz 5 - DSC 10, Winter 2025

Full Name:

PID:

Quiz Time: ◯ 3PM ◯ 3:30PM ◯ 4PM ◯ 4:30PM

Instructions:

- This quiz consists of 3 questions. You have a total of 20 minutes to complete it.
- Please write **clearly** in the provided answer boxes; we will not grade work that appears elsewhere. Completely fill in bubbles and square boxes; if we cannot tell which option(s) you selected, you may lose points.

  ◯ A bubble means that you should only **select one choice**.

  ☐ A square box means you should **select all that apply**.
- If your answer is a string, make sure to put it in quotes. If your answer is a float, make sure to include a decimal point.
- You may use one handwritten sheet of notes. No calculators and no computers.

By signing below, you are agreeing that you will behave honestly and fairly during and after this quiz.

Signature:

Version A

Please do not open your quiz until instructed to do so.

Question 1

UCSD is testing whether a new version of the TritonLink website will improve student engagement. Each student who visits the Tritonlink website receives either the old version or the new version at random, and the number of minutes spent on the site is recorded.

The results of the experiment are stored in the DataFrame `tritonlink`, whose first few rows are shown at right.

student_id	version	minutes
A17924754	old	12
A21827364	new	15
A31765432	new	14
A01900287	old	6
A91856473	new	7

Let's do a permutation test with the following hypotheses:

Null Hypothesis: The amount of time students spend on the TritonLink website is **the same for both versions** of the website.

Alternative Hypothesis: Students spend **more time on the new version** of the site.

a) For our test statistic, we'll use the mean time spent on the new version of the site minus the mean time spent on the old version of the site. The function below should compute one simulated difference of this statistic under the null hypothesis.

```
def one_stat():
    shuffled = np.random.permutation(__(a)__)
    shuffled_df = tritonlink.assign(shuffled = shuffled)
    grouped = (shuffled_df.get(["shuffled", "minutes"])
                          .groupby(__(b)__).mean().get("minutes"))
    return __(c)__
```

(i) Which of the following statements about blank `(a)` is correct?

◯ We can shuffle either `"minutes"` or `"version"` because it doesn't matter which column we shuffle in a permutation test.

◯ We must shuffle `"minutes"` because only numerical data can be shuffled in a permutation test.

◯ We must shuffle `"version"` because only group labels can be shuffled in a permutation test.

◯ We must shuffle `"minutes"` because of the structure of the provided code.

◯ We must shuffle `"version"` because of the structure of the provided code.

(ii) Fill in blanks `(b)` and `(c)` below.

`(b)`:

`(c)`:

b) Suppose we complete our permutation test and calculate a p-value. Which of the following statements are true? Select all that apply.

- ☐ If the p-value is 0.18, this means that if the alternative hypothesis is true, the probability of seeing a difference as or more extreme as the observed difference is about 18%.
- ☐ If the p-value is 0.03, this means that if the null hypothesis is true, the probability of seeing a difference as or more extreme as the observed difference is about 3%.
- ☐ If the p-value is 0, this means that we implemented our permutation test incorrectly.
- ☐ If the p-value is 0.5, this means that exactly half of the simulated test statistics were less extreme than our observed difference.
- ☐ None of the above.

Question 2

For the DataFrame `df` shown in full to the right, both the `"x"` and `"y"` columns have a mean of 0 and a standard deviation of 1. Calculate the correlation between `"x"` and `"y"`.

	x	y
0	-2	-1
1	-1	-1
2	0	-1
3	0	-1
4	0	-1
5	0	1
6	0	1
7	0	1
8	1	1
9	2	1

Question 3

UCSD uses a system called EASy for students to submit requests for enrollment if they need an exception to the enrollment requirements. Naturally, bigger classes generally receive more EASy requests.

Suppose the average number of seats per class (x) is 75, with a standard deviation of 25, and the average number of EASy requests per class (y) is 10, with a standard deviation of 4.

a) For this subpart only, suppose the correlation between x and y is 0.5. Use the regression line to predict the number of EASy requests for a class with 100 seats. Give your answer as an **integer**.

b) A Japanese class has 150 seats and receives 16 EASy requests. Say we use linear regression to predict the number of EASy requests from the number of seats, and find that the **squared** residual for this Japanese class is 4. What is the **smallest** possible correlation between x and y in this case? Give your answer as a **simplified fraction**.

Reminders:

- **Write your PID** on the front page and on the top right corner of page 3.
- If your answer is a string, make sure to put it in **quotes**.
- If your answer is a float, make sure to include a **decimal point**.
- Stay in your seat until the quiz is over. No leaving early.