Quiz 4 - DSC 10, Fall 2024

Full Name:

PID:

Quiz Time: ◯ 4PM ◯ 4:30PM ◯ 5PM ◯ 5:30PM ◯ 7PM

Instructions:

- This quiz consists of 4 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.
- No aids are allowed (no 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.

The DataFrame `movies` contains information on a variety of movies. We have recorded each movie's `"Name"`, `"Genre"`, and `"Rating"`, according to IMDb. The first three rows are shown here, but `movies` has many more rows than pictured.

	Name	Genre	Rating
0	La La Land	Musical	8.0
1	Barbie	Comedy	6.8
2	Spider-Man	Action	7.4

Question 1

We plan to collect a sample of movies and use this sample to estimate the proportion of all movies with the genre `"Musical"`, a population parameter.

a) If we want to create a 95% confidence interval that is at most 0.08 wide, which of the expressions below represents the smallest sample size we should collect?

◯ $\left(\frac{1}{0.02}\right)^2$ ◯ $\left(\frac{1}{0.04}\right)^2$ ◯ $\left(\frac{1}{0.08}\right)^2$ ◯ $\left(\frac{1}{0.16}\right)^2$

b) Let W represent the maximum width of a 95% confidence interval obtained from a sample that is **twice as big** as the sample size you found in part (a). Which of the following is true?

◯ $0 < W < 0.04$ ◯ $W = 0.04$ ◯ $0.04 < W < 0.08$ ◯ $W \geq 0.08$

Question 2

Consider the following pair of hypotheses:

- **Null:** The proportion of `"Comedy"` movies with an average rating above 9 **equals** the proportion of `"Action"` movies with a rating above 9.
- **Alternative:** The proportion of `"Comedy"` movies with an average rating above 9 is **greater than** the proportion of `"Action"` movies with a rating above 9.

Let C_9 be the proportion of `"Comedy"` movies with a rating above 9 and let A_9 be the proportion of `"Action"` movies with a rating above 9. Which of the following are valid test statistics to test these hypotheses? **Select all that apply.**

☐ C_9 - 0.5 ☐ A_9 - 0.5 ☐ $A_9 \cdot C_9$ ☐ $A_9 + C_9$

☐ $C_9 - A_9$ ☐ $A_9 - C_9$ ☐ $|A_9 - C_9|$ ☐ $|C_9 - A_9|$

Question 3

The table below shows how many movies in each of four genres Jack and Eric have seen. What is the total variation distance (TVD) between Jack and Eric's distribution of movies by genre?

	`"Musical"`	`"Comedy"`	`"Action"`	`"Horror"`	Total
Jack	2	14	2	2	20
Eric	55	15	15	15	100

TVD = ☐

PID: _____

Question 4

Suppose that in `movies`, the average rating of `"Action"` movies is 7.4 and the average rating of `"Horror"` movies is 7.2. Based on this data, we decide to test the following hypotheses:

- **Null**: The ratings of `"Action"` and `"Horror"` movies come from the same distribution.
- **Alternative**: On average, `"Action"` movies have a higher rating `"Horror"` movies.

We'll use as our test statistic the mean rating of `"Action"` movies minus the mean rating of `"Horror"` movies.

a) Fill in the blanks so the code below generates 5000 simulated values of this test statistic and calculates the p-value of our test.

```
def one_stat(df):
    group_means = (df.get(["New", "Rating"])
                   .groupby("New").mean().get("Rating"))
    return group_means.loc[__(a)__] - group_means.loc[__(b)__]

action_horror = movies[(movies.get("Genre") == "Action") |
                       (movies.get("Genre") == "Horror")]
diffs = np.array([])
for i in np.arange(5000):
    new_df = action_horror.assign(New = __(c)__)
    diffs = np.append(diffs, __(d)__)

p_value = np.count_nonzero( __(e)__ ) / 5000
```

a: _____ b: _____

c: _____

d: _____

e: _____

b) Suppose that `p_value` evaluates to 0.14. Using the standard p-value cutoff of 0.05, which of the two hypotheses is better supported by the data?

◯ Null ◯ Alternative

c) What kind of hypothesis test did we perform in this question?

◯ Standard hypothesis test ◯ Permutation test

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.