Data 8 Foundations of Data Science

Fall 2023

Midterm Exam

INSTRUCTIONS

The exam is worth 90 points. You have 1 hours and 50 minutes to complete the exam.

- The exam is closed book, closed notes, closed computer/calculator, except the provided midterm reference sheet.
- Mark your answers on the exam itself in the spaces provided. We will not grade answers written on scratch paper or outside the designated answer spaces.
- If you need to use the restroom, bring your phone, Cal ID, and exam to the front of the room.

For questions with **circular bubbles**, you should fill in exactly *one* choice.

◯ You must choose either this option

◯ Or this one, but not both!

For questions with **square checkboxes**, you may fill in *multiple* choices.

☐ You could select this choice.

☐ You could select this one too!

Important: Please **fill in** circles and squares to indicate answers and cross out or erase mistakes.

Preliminaries

You can complete these questions before the exam starts.

(a) What is your full name?

(b) What is your student ID number?

(c) Who is sitting to your left? (Write *no one* if no one is next to you.)

(d) Who is sitting to your right? (Write *no one* if no one is next to you.)

1. (23.0 points) True or False

(a) (1.0 pt) All of the work on this exam is your own.

◯ True

◯ False

(b) (2.0 pt) The chance of two events A and B both happening can sometimes be greater than the chance of either A or B happening.

◯ True

◯ False

(c) (2.0 pt) The Law of Averages implies that with high probability, the empirical distribution of a large random sample will resemble the distribution of the population from which the sample was drawn.

◯ True

◯ False

(d) (2.0 pt) The total variation distance can be applied to categorical distributions in which there are only 2 possible categories (e.g. purple or white).

◯ True

◯ False

(e) (2.0 pt) The reason we shuffle labels in an A/B test is to ensure that our subjects are randomly assigned to treatment and control.

◯ True

◯ False

(f) (2.0 pt) Suppose a hypothesis test is proposed and we already know that the null hypothesis is true. If 500 researchers each independently collects a sample of the same size to carry out an experiment and they all use 1% as their p-value cutoff, we should expect around 5 of them to reject the null.

◯ True

◯ False

(g) (2.0 pt) The height of each bar in a histogram always represents the proportion of data within the corresponding bin.

◯ True

◯ False

(h) (2.0 pt) If you are a subject in an experiment, knowing whether you are in the treatment or control group can be considered a confounding variable.

◯ True

◯ False

(i) (2.0 pt) When shuffling labels for a permutation test, sampling must be done without replacement.

◯ True

◯ False

(j) (2.0 pt) The median of a set of 8 integers will always be an integer.

◯ True

◯ False

(k) (2.0 pt) According to the Welcome Survey, the least common way that Data 8 students sleep is on their stomach.

◯ True

◯ False

(l) (2.0 pt) A function in Python must always have a return statement.

◯ True

◯ False

2. (22.0 points) Friday the 13th

Thomas and his friends Andrew, Bianca, and Sheema are getting ready for the Data Scholars Halloween costume party, but they just aren't sure what to wear!

To make things easier, they decide to fill a box with various costume accessories and take turns drawing items at random without replacement from the box.

Assume the box contains the following accessories:

- 3 Oski Masks
- 2 Witch Hats
- 2 Superhero Capes
- 1 Broomstick

(a) (9.0 points)

Sheema, Thomas, Andrew, and Bianca are calculating the probabilities of drawing some combinations of accessories.

i. (2.0 pt) Suppose that Andrew is first to draw, and draws 3 items.

What is the probability that he draws all 3 Oski Masks?

You can write your answer as a mathematical expression without simplifying it.

ii. (2.0 pt) Suppose that Sheema is first to draw, and draws 1 item.

What is the probability she does NOT get a Witch Hat or a Superhero Cape?

You can write your answer as a mathematical expression without simplifying it.

iii. (2.0 pt) Suppose that Thomas is first to draw, and draws 2 items.

What is the probability that his two items are an Oski Mask and a Witch Hat?

You can write your answer as a mathematical expression without simplifying it.

iv. (3.0 pt) Suppose that Bianca is first to draw. She draws 2 items and gets a Witch Hat and a Broomstick.

If Thomas draws next, what is the chance he does NOT get an Oski Mask or a Witch Hat?

You can write your answer as a mathematical expression without simplifying it.

(b) (13.0 points)

Suppose that the group decides to have the entire Data 8 course staff take turns drawing from the box *with replacement*.

Andrew is curious how often the broomstick will show up, so he creates the following partially completed code to simulate the number of broomsticks drawn in `n` draws:

```
def simulate_broomsticks(n):

    prob_broomstick = _________
                        (a)

    prob_distribution = make_array(prob_broomstick, _________)
                                                      (b)

    simulated_proportions = _________
                              (c)

    num_broomsticks = _________ * simulated_proportions._________
                        (d)                                 (e)

    return num_broomsticks
```

i. (1.0 pt) Fill in blank (a).

ii. (2.0 pt) Fill in blank (b).

iii. (3.0 pt) Fill in blank (c).

iv. (2.0 pt) Fill in blank (d).

v. (2.0 pt) Fill in blank (e).

vi. (3.0 pt) Suppose that Andrew runs the following code:

```
smalls = make_array()
bigs = make_array()

for i in np.arange(10000):
    smalls = np.append(smalls, simulate_broomsticks(100) / 100)
    bigs = np.append(bigs, simulate_broomsticks(500) / 500)

Table().with_column('Small', smalls).hist('Small')
Table().with_column('Big', bigs).hist('Big')
```

Which of the following will he observe?

Select all that apply.

☐ The `'Big'` distribution will be wider than the `'Small'` distribution.

☐ The `'Big'` distribution and `'Small'` distribution will both be centered around the same value.

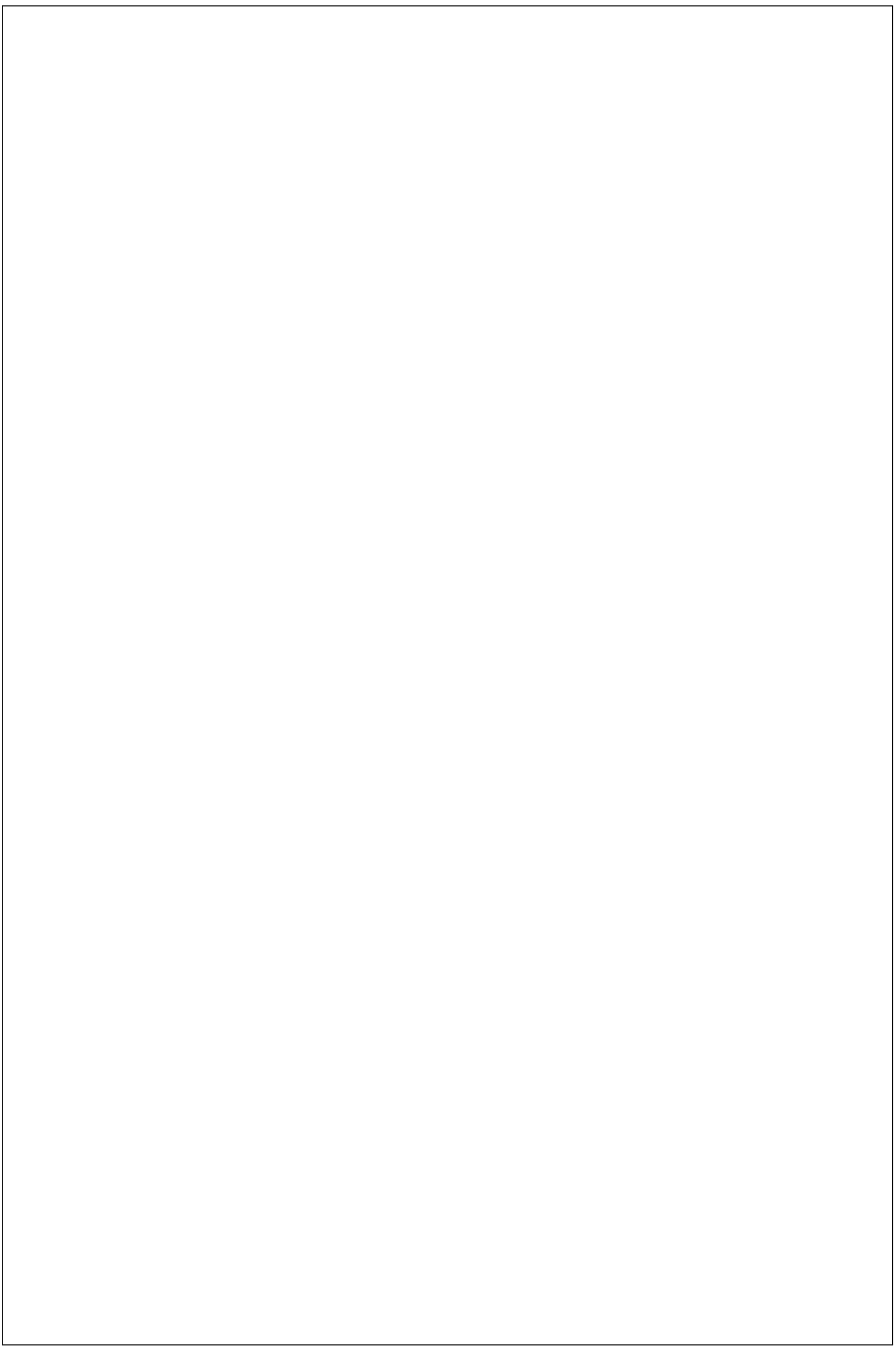
☐ The `'Big'` distribution will be narrower than the `'Small'` distribution.

☐ The `'Big'` distribution will have roughly the same width as the `'Small'` distribution, but it will look smoother.

☐ None of the above.

3. (0.0 points) Quick Break: Just For Fun

(a) (Optional) Take a break and draw a picture of a *Data 8* themed Halloween costume! Make sure to finish the rest of the exam, though!



4. (31.0 points) Coco's Crosswords

Every day, Coco solves a random New York Times crossword and keeps track of how long it takes to solve each one. She stores this data in a table called `crosswords`, with each row representing 1 puzzle.

A sample of the table is provided below:

Date	Published	Solved	Minutes	Seconds
10-05	Thursday	Friday	12	32
10-06	Friday	Tuesday	28	51
10-08	Sunday	Sunday	5	23
10-09	Monday	Saturday	10	5

The table has the following columns:

- *Date*: (string) the month and day that the crossword was published
- *Published*: (string) the day of the week the crossword was published
- *Solved*: (string) the day of the week the crossword was solved
- *Minutes*: (int) the number of minutes the crossword took to solve, rounded down
- *Seconds*: (int) the number of seconds past *mins* that the crossword took to solve

(a) (3.0 pt) Coco writes a function that takes in a row from the `crosswords` table and converts the *Minutes* and *Seconds* columns into only seconds. For example, the first row shown in the sample above would return `732`.

She writes the following partially completed code:

```
def duration_in_seconds(row):
    return ____________
```

Fill in the blank to complete the function.

Reminder: There are 60 seconds in 1 minute.

(b) (5.0 points)

Coco wants to use the `duration_in_seconds` function to generate the duration for every puzzle in the `crosswords` table and add it to her table as a new column name *Total Duration*.

She writes the following partially completed code:

```
durations = crosswords._________
                        (a)

crosswords = _________
              (b)
```

i. (2.0 pt) Fill in blank (a).

ii. (3.0 pt) Fill in blank (b).

(c) (4.0 pt) After running the code in the previous question, Coco next wants to create a table where each unique publishing day of the week gets its own row, each unique solving day of the week gets its own column, and the values inside the table correspond to the average time in seconds it took to solve the puzzles for each combination of published and solved day of the week.

She writes the following partially completed code:

```
crosswords._________
```

Fill in the blank.

Recall: The `crosswords` table has columns *Date*, *Published*, *Solved*, *Minutes*, *Seconds*, and *Total Duration*.

(d) (2.0 pt) Coco's roommate, Aryna, would like to see how *Total Duration* varies between puzzles published on Saturdays and puzzles published on Mondays.

Which of the following lines of code could help with this?

Select all that apply.

☐ `crosswords.scatter('Published', 'Total Duration')`

☐ `crosswords.hist('Total Duration')`

☐ `crosswords.pivot('Total Duration', 'Published')`

☐ `crosswords.hist('Total Duration', group='Published')`

☐ `crosswords.scatter('Published', 'Total Duration', group='Published')`

(e) (6.0 points)

Aryna has been secretly reviewing Coco's puzzles and scoring them for correctness.

She's created a separate table called `puzzles` that contains every puzzle that Coco has solved. It has the following columns:

- *Edition*: (string) the month and day that the crossword was published (for example, `'10-05'` for October 5)
- *Score*: (float) the percent of letters that Coco answered correctly

Aryna suspects that Coco doesn't do that well on her crosswords when solving them on Saturdays because she's busy playing sports all day.

She writes the following partially completed code, which assigns `min_score` to the minimum score that Coco ever received when solving a puzzle on a Saturday.

```
min_score = np.min(crosswords._____(a)_____.where(_____(b)_____)._____(c)_____)
```

Recall: The `crosswords` table has columns *Date*, *Published*, *Solved*, *Minutes*, *Seconds*, and *Total Duration*.

i. (2.0 pt) Fill in blank (a).

ii. **(2.0 pt)** Fill in blank (b).

iii. **(2.0 pt)** Fill in blank (c).

(f) (5.0 points)

Suppose Aryna wants to test the hypothesis that Coco gets lower scores when solving puzzles on Saturdays as opposed to solving puzzles on other days of the week.

She proceeds to take a random sample from the `crosswords` table.

i. (2.0 pt) Which of the following test statistics could she use? Skip Question

Select all that apply.

- ☐ Mean *Score* for Saturdays
- ☐ Mean *Score* for non-Saturday days minus mean *Score* for Saturdays
- ☐ Absolute difference between mean *Score* for Saturdays and mean *Score* for non-Saturday days
- ☐ Mean *Score* for Saturdays minus mean *Score* for non-Saturday days
- ☐ None of the above because the sample size of Saturday scores is different from the sample size of non-Saturday scores

ii. (3.0 pt) Suppose Aryna carries out her hypothesis test and calculates a p-value of 2%.

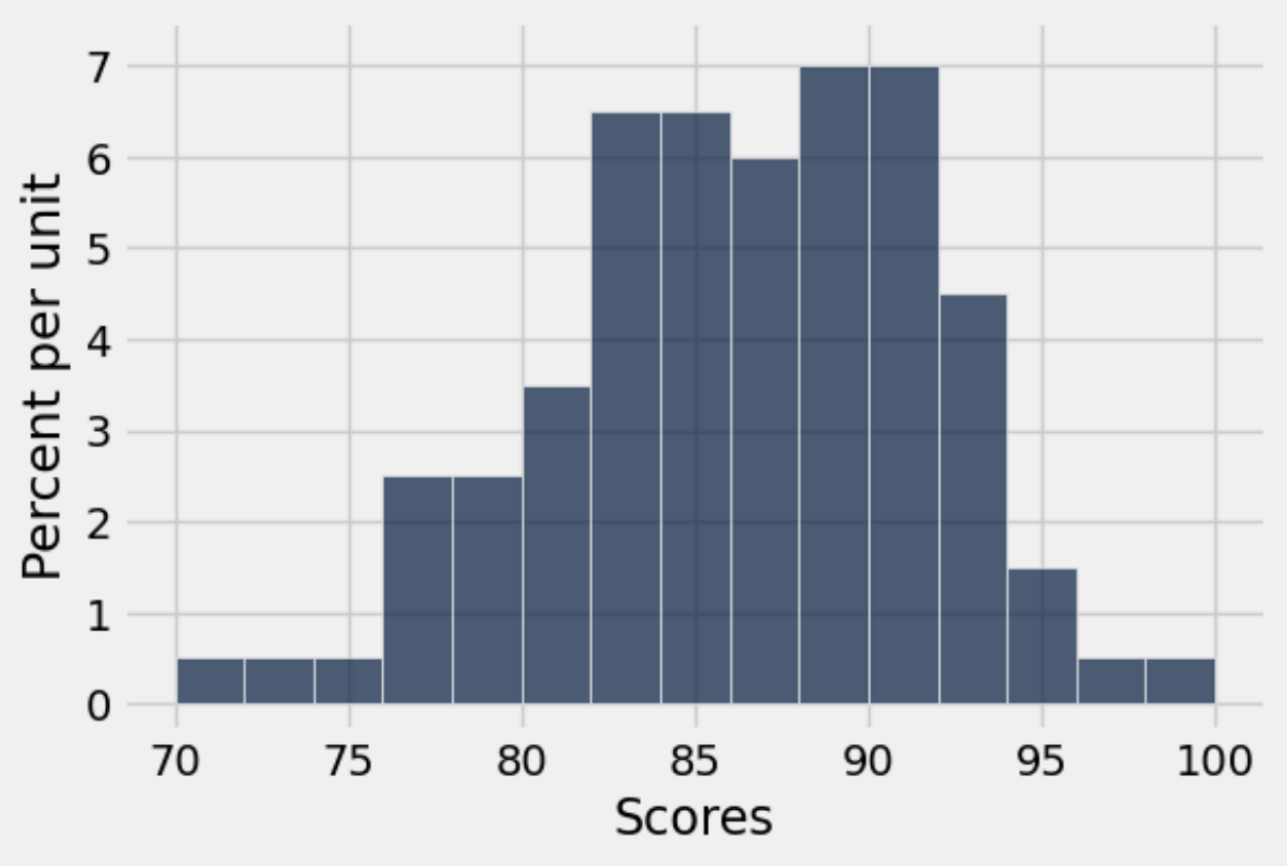
Using a **5%** cutoff, which of the following conclusions can she make?

Select all that apply.

- ☐ The alternative hypothesis is true.
- ☐ Coco gets lower scores on average when solving puzzles on Saturday compared to solving on other days of the week.
- ☐ Coco gets lower scores on average when solving puzzles on the weekend compared to solving on weekdays.
- ☐ Solving puzzles on Saturdays causes Coco to get lower scores on average than when solving on other days of the week.
- ☐ The distribution of scores is the same between puzzles that Coco solves on Saturdays and puzzles that she solves on other days of the week.

(g) (6.0 points)

Aryna generates the following histogram of Coco's scores across all puzzles she solved on Saturdays.



Note: The histogram was generated using equally spaced bins.

i. (2.0 pt) Based on the histogram, which of the following are possible values of `min_score` from the previous question?

Select all that apply.

☐ 69.8

☐ 70

☐ 71.9

☐ 72

☐ 75

☐ 100

☐ Not enough information to tell

ii. **(2.0 pt)** Based on the histogram, what percentage of Coco's scores across all puzzles solved on Saturdays are below or equal to 78?

- ◯ 8%
- ◯ 4%
- ◯ 2.5%
- ◯ 5%
- ◯ Not enough information to tell

iii. **(2.0 pt)** Aryna concludes that Coco received a score of 87 or higher in more than 50% of the crosswords she solved on Saturdays.

Based on the histogram above, is this statement correct?

- ◯ Yes
- ◯ No
- ◯ Not enough information to tell

5. (24.0 points) Waystar

Waystar is an organization that owns businesses and trades stocks in a variety of sectors including media, entertainment, tech, etc. Its executive team wants to understand the performance of its businesses.

Waystar executives Tom and Greg put together a table called `performance`, which contains randomly sampled public information about the various businesses' performance over the last 40 years. Here are the first few rows:

Name	Year	Revenue	Sector	Advice
NY Globe	2019	18.8	Media	Hold
Brightstar	2018	92.8	Leisure	Sell
Vaulter	2021	150.9	Tech	Buy
ATN	2020	61.7	Media	Hold
Brightstar	2019	89.2	Leisure	Sell

... (155 rows omitted)

The table contains the following columns:

- *Name*: (string) the name of the business
- *Year*: (int) the year of the financial performance
- *Revenue*: (float) the business's revenue (in millions of USD)
- *Sector*: (string) the business's sector in the industry
- *Advice*: (string) Wall Street's stock purchase advice (`'Buy'`, `'Hold'` or `'Sell'`)

Greg notices that businesses in the `'Tech'` sector seem to have higher revenue than those in `'Media'`. Tom thinks the differences observed in the sample are only due to chance.

(a) (3.0 pt) Which of these are **null** hypotheses that Greg could use to assess his claims?

Select all that apply.

☐ Businesses with high revenue have the same *Sector* distribution as businesses with low revenue.

☐ `'Tech'` businesses have the same revenue distribution as `'Media'` businesses.

☐ `'Tech'` businesses have the same revenue distribution as businesses that are not `'Tech'`.

☐ `'Media'` businesses have the same revenue distribution as `'Tech'` businesses.

☐ None of the above.

(b) (3.0 pt) Which of the following function calls would be helpful when simulating data from the null hypothesis?

Assume `props` is an array containing the categorical distribution of *Sector* in `performance`.

Select all that apply.

☐ `performance.sample()`

☐ `performance.sample(with_replacement=False)`

☐ `performance.sample(200, with_replacement=False)`

☐ `sample_proportions(200, props)`

☐ None of the above.

(c) (3.0 pt) Suppose Greg decides to use a test statistic such that higher values are in favor of the alternative hypothesis.

He simulates 1,000 values of the test statistic and finds that 40 of them are greater than the observed test statistics.

Assuming he uses a **5%** cutoff, which of the following can he conclude?

Select all that apply.

☐ In the population, `'Tech'` businesses and `'Media'` businesses have the same *Revenue* distribution.

☐ In the population, businesses with high revenue have a different *Sector* distribution than businesses with low revenue.

☐ In the population, `'Tech'` businesses have higher *Revenue* on average compared to `'Media'` businesses.

☐ In the population, being in `'Tech'` causes businesses to have higher *Revenue* on average compared to `'Media'` businesses.

☐ In the population, `'Tech'` businesses have higher *Revenue* on average compared to businesses that are not `'Tech'`.

(d) (9.0 points)

Greg notices that businesses in the `'Tech'` sector have a different *Advice* distribution than those in `'Leisure'`.

Tom thinks the differences observed in the sample are only due to chance.

Greg wants to make a function to calculate the total variation distance of the *Advice* distributions between `'Tech'` and `'Leisure'` businesses.

He writes the following partially completed code:

```
def test_stat(data, category_a, category_b):

    dist_a = data.where(_________, category_a)._________
                           (a)                    (b)

    counts_a = dist_a.sort(0)._________
                                 (c)

    dist_b = data.where(_________, category_b)._________
                           (a)                    (b)

    counts_b = dist_b.sort(0)._________
                                 (c)

    props_a = counts_a / np.sum(counts_a)

    props_b = counts_b / np.sum(counts_b)

    return _________
              (d)
```

Note: The function's arguments are `data` *(Table),* `category_a` *(string) and* `category_b` *(string).*

i. **(1.0 pt)** Fill in blank (a).

ii. **(2.0 pt)** Fill in blank (b).

iii. **(2.0 pt)** Fill in blank (c).

iv. **(4.0 pt)** Fill in blank (d).

(e) (3.0 pt) Greg simulates 1,000 values of the test statistic under the null and stores these in an array called `test_stats`. Suppose the observed value of the test statistic is `0.52`.

Write a Python expressions that returns the p-value for this hypothesis test.

(f) (3.0 pt) Greg creates a histogram of `test_stats` and uses the area principle to calculate that at least 6% of the values are greater than `0.55`.

If his p-value cutoff is **5%** and the observed test statistic is `0.52`, which of the following can he conclude?

Select all that apply.

☐ The data are consistent with the null hypothesis.

☐ The data are consistent with the alternative hypothesis.

☐ The null hypothesis is true.

☐ The null hypothesis is false.

☐ There is not enough information to make any of these conclusions.

6. (0.0 points) Optional

(a) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question (including the title of the section, e.g., Experiments) and state your assumptions. Be warned: We only plan to consider this information if we agree that the question was erroneous or ambiguous and we consider your assumption reasonable.