Data C8
Summer 2025

Foundations of Data Science Midterm

12:00-2:00PM, FRIDAY, JULY 18

Berkeley Honor Code [1 point]

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Initials: _____

Full Name: Solutions

Student ID Number: _____

Exam Location: _____

Name of person to your left: _____

Name of person to your right: _____

GSI/TA's Name: _____

INSTRUCTIONS

- You may only have with you: a pencil(s), an eraser(s), your student ID, a water bottle, and your midterm reference sheet, unless you have received pre-approved accommodations otherwise.
- If you need to use the restroom, bring your phone, exam, reference sheet, and student ID to the front of the room.
- Do not open the exam until you are instructed to do so.
- Write your initials at the top of each page.
- There are **4** questions and **15** pages on this exam, including cover page. **Read the instructions and point values carefully** for each question, part and subpart.
- Multiple choice questions with bubbles ◯ have one correct answer. Multiple choice questions with squares ☐ have one or more correct answers.
- Where relevant, you may assume that all necessary Python modules have been imported. Data C8 is an introductory course. Out of fairness to all students, use of any code which has not been taught in this iteration of the course is prohibited and it will not be graded.
- Where a written (English) answer is expected, you must use complete sentences. Your work will not be graded otherwise.
- **Each coding blank may include multiple arguments/methods/functions.** However, your solution must use every blank available.

1 General [20 points]

Read the instructions for each question carefully and answer accordingly.

a. (5 points) Which of the facets of data science have been the primary focus of the first half of the course? *Select all that apply.*

☐ Prediction ☐ Coding ☐ Conceptual
☑ Inference ☑ Exploration ☐ A facet(s) not listed here.

b. (4 points) What will the following Python expression output?

```
make_array(0,0,0) == False
```

◯ `False` ◯ `array([False,False,False])`
◯ `True`
◉ `array([True,True,True])` ◯ An error will be produced.

c. (3 points) Below is a pivot table made from the `cones` table seen in lecture, using the `"Color"` and `"Flavor"` columns in that table. In each cell of this table lies the number of ice cream cones belonging to a particular color-flavor combination. What would be the dimensions (rows x columns) of a grouped table made from `cones` that displays the same information as this pivot table?

Color	bubblegum	chocolate	strawberry
dark brown	0	2	0
light brown	0	1	0
pink	1	0	2

Color Flavor Count

◉ 4 x 3 ◯ 9 x 3 ◯ 3 x 4 ◯ 2 x 4
◯ 4 x 2 ◯ 9 x 2 ◯ 3 x 9 ◯ 2 x 9

d. (1 point) Prof Jeremy is interested in seeing whether removing multiple choice as a question format causes a difference in student performance on a midterm exam. He randomly assigns half of the 200 students in his class to receive midterm version A, which contains both multiple choice and free-response questions, and assigns the other half to receive version B, which contains only free-response questions. How many potential outcomes does Prof Jeremy end up observing in his experiment?

◯ 400 ◯ 100 ◯ 25
◉ 200 ◯ 50 ◯ None of these.

e. (4 points) Every semester after the midterm, Data C8 staff solicits and records feedback from each student on their experiences with the course using a Google Form. The `feedback` table contains student responses to a select set of questions asked on the form from the Spring 2025 semester.

- **Lab** (string): A student's choice of lab format: in-person (Regular) or virtual (Self-Service).
- **Lecture** (string): Whether a student regularly interacts with the Data C8 lecture, either through live attendance or through the recording (Yes or No).
- **Ed** (integer): How satisfactory a student found interacting with Ed, the class forum, to be, on a scale of 1 (strongly disagree) to 5 (strongly agree).
- **Assignments** (integer): How many assignments that a student had turned in on Gradescope before the midterm exam.

Lab	Lecture	Ed	Assignments
Regular	Yes	5	18
Regular	No	3	20
Self-Service	No	4	17

...(1397 rows omitted)

(i) (2 points) Cyrus is interested in seeing whether there is a difference in Spring 2025 lecture attendance between the two lab format groups. Which of the following procedures/lines of code is appropriate to complete this task? *Select all that apply.*

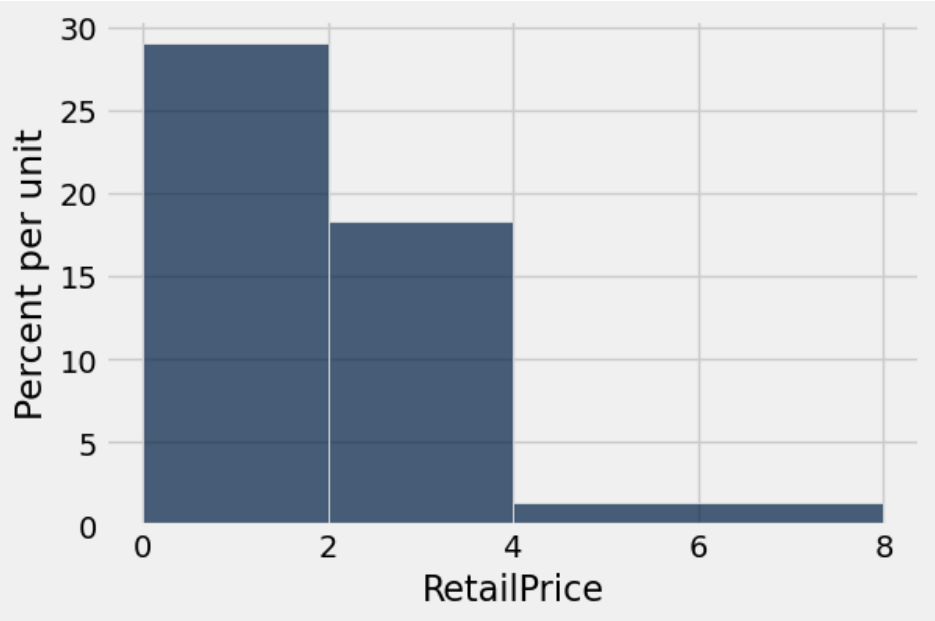
- ☐ Perform a hypothesis test where the null hypothesis is that the Spring 2025 students in Regular and Self-Service labs belong to the same, underlying lecture attendance distribution.
- ☑ `feedback.group(['Lab','Lecture'])`
- ☐ `feedback.pivot('Lab','Lecture', 'Ed', np.median)`
- ☑ Visualize the **Lab** and **Lecture** attributes using a bar chart.
- ☐ Visualize the **Lab** and **Lecture** attributes using a histogram.

(ii) (2 points) Which of the following statements is true regarding the **Ed** and **Assignments** attributes?

- ◯ Both attributes are equally appropriate for visualization with a histogram.
- ◯ Neither attribute is appropriate for visualization with a histogram.
- ◯ **Ed** is more appropriate than **Assignments** for visualization with a histogram.
- ● **Assignments** is more appropriate than **Ed** for visualization with a histogram.

f. (3 points) The following visualization displays the prices (per gram) of 93 vegetable products as estimated by the U.S. Department of Agriculture. *True or False:* Roughly 75 percent of these vegetable products are priced between 0 and 3 dollars per gram.

◯ True ◯ False ● We cannot determine an answer.



2 Data Tea-8 [18 points]

Richard is studying well into the night inside Evans Hall, and can't stop thinking about the iced black tea he is going to get afterward. Richard's plan is to place an online order so that he can walk to a tea shop and have the tea ready for pickup. He has access to two tables: `drinks` and `businesses`, that might help him select an iced black tea to order.

`drinks` table:

- **Shop** (string): Name of the tea shop.
- **Price** (float): The base price of an iced black tea, as obtained from the online ordering site for the tea shop.
- **Rating** (float): The aggregated Yelp rating for the tea shop, out of 5.0 stars.

Shop	Price	Rating
Asha Tea House	5.75	4.0
Ti-Bear	6.75	4.5
Happy Lemon	7.75	4.0

...(some rows omitted)

`businesses` table:

- **Business** (string): Name of the business.
- **Neighborhood** (string): The neighborhood within Berkeley that the business is located in.
- **Minutes** (integer): The estimated number of minutes that it takes to walk from Evans Hall to the business, according to Apple Maps.
- **Late Night** (boolean): Indicates whether or not the business is open past 9pm.

Business	Neighborhood	Minutes	Late Night
Asha Tea House	Downtown	16	`False`
TP Tea	Southside	11	`False`
Plentea	Southside	11	`True`

...(some rows omitted)

Fill in the blanks in the Python expressions to compute the described values. You must use only the lines provided. *The last line of the answer should evaluate to the value described.*

a. (5 points) A three-column table where the third column lists the number of tea shops having a particular combination of neighborhood and late-night availability.

```
businesses._______A_______(_______B_______)
```

(i) (3 points) Blank `A`:

group

(ii) (2 points) Blank `B`:

['Neighborhood', 'Late Night']

b. (5 points) The number of tea shops with a rating above 4.0 stars and an iced black tea priced less than 7 dollars.

```
drinks.____A____(_______B_______).____C____(_______D_______)._______E_______
```

(i) (1 point) Blank `A`:

where

(ii) (1 point) Blank `B`:

'Rating', are above (4.0)

(iii) (1 point) Blank `C`:

where

(iv) (1 point) Blank `D`:

'Price', are below (7.0)

(v) (1 point) Blank `E`:

num_rows

c. (3 points) A two column table containing the average price of iced black teas in tea shops for each neighborhood.

```
drinks.____A____(_______B_______).select(_______C_______).____D____(_______E_______)
```

(i) (0.6 points) Blank `A`:

join

(ii) (0.6 points) Blank `B`:

'Shop', businesses, 'Business'

(iii) (0.6 points) Blank `C`:

'Price', 'Neighborhood'

(iv) (0.6 points) Blank `D`:

group

(v) (0.6 points) Blank `E`:

'Neighborhood', np.average

d. (1 point) The name of the restaurant which offers the best value for an iced black tea, given the amount of time it takes to walk to the restaurant from Evans Hall.

```
d_and_b = drinks.___A___(_______B_______)

d_and_b = d_and_b.___C___(_______D_______).sort('Price per minute')

d_and_b.___E___('Shop').item(____F____)
```

(i) (0.2 points) Blank `A`:

join

(ii) (0.1 points) Blank `B`:

'Shop', businesses, 'Business'

(iii) (0.1 points) Blank `C`:

with_column

(iv) (0.4 points) Blank `D`:

'Price per minute', d_and_b.column('Price')/d_and_b.column('Minutes')

(v) (0.1 points) Blank `E`:

column

(vi) (0.1 points) Blank `F`:

0

e. (4 points) Now, Richard would like to visualize the relationship between the time it takes to walk to a tea shop from Evans Hall and the price of the shop's iced black tea. What kind of visualization would be most appropriate for this task?

◯ Bar chart ◯ Histogram ◉ Scatter plot ◯ Line plot

3 Rock, Paper, Scissors [15 points]

When Bing isn't teaching data science, he is honing his rock-paper-scissors skills. **Rock-paper-scissors games are considered to last four rounds**. Each round is contested between two players. In a given round, a player can choose to show either `"rock"`, `"paper"` or `"scissors"`. Both players show their choice at the same time. Bing's strategy is simple; he always chooses `"rock"`, regardless of what happens in previous rounds of the game. In each round of the game (regardless of the results of previous rounds), Bing's opponents always choose at random between `"rock"`, for which the game ends in a tie; `"paper"`, for which Bing loses; or `"scissors"`, for which Bing wins.

a. (4 points) What is the probability that Bing wins at least one round in a rock-paper-scissors game? **Show your work** and write your answer in the box below. You may leave your answer as an arithmetic expression.

P(at least one round) = 1 - P(no rounds) COMPLEMENT RULE

P(no rounds) = P(lose all 4 rounds) = $2/3 \cdot 2/3 \cdot 2/3 \cdot 2/3 = (2/3)^4$

$= 1 - (2/3)^4$

b. (3 points) The function `one_game` simulates one game of rock-paper-scissors and returns *the number of rounds in the game won by Bing*. Complete the skeleton code for `one_game` below.

```
def one_game():

    options = make_array("rock", "paper", "scissors")

    number_of_wins = __________A__________

    for __________B__________:

        opponent_choice = np.random.choice(__________C__________)

        if __________D__________:

            number_of_wins = __________E__________

    return __________F__________
```

(i) (0.5 points) Blank `A`:

0

(ii) (0.5 points) Blank `B`:

i in np.arange(4)

(iii) (0.5 points) Blank `C`:

```
options
```

(iv) (0.5 points) Blank `D`:

```
opponent_choice == "scissors"
```

(v) (0.5 points) Blank `E`:

```
number_of_wins + 1
```

(vi) (0.5 points) Blank `F`:

```
number_of_wins
```

c. (5 points) Bing has been a student at Berkeley for three years (nearly 1,000 days)! On each day, he plays *one* rock-paper-scissors game with a student. The function `one_thousand_days` simulates 1,000 days of rock-paper-scissors games and returns *an array containing the number of rounds Bing won on each day*. Write the function in the space provided below. You may use functions that have previously been defined. You may also assume that Bing's performance on any given day is independent from his performance on other days.

```
def one_thousand_days():
```

```
number_of_rounds = make_array()
for i in np.arange(1000):
    one_day = one_game()
    number_of_rounds = np.append(number_of_rounds, one_day)

return number_of_rounds
```

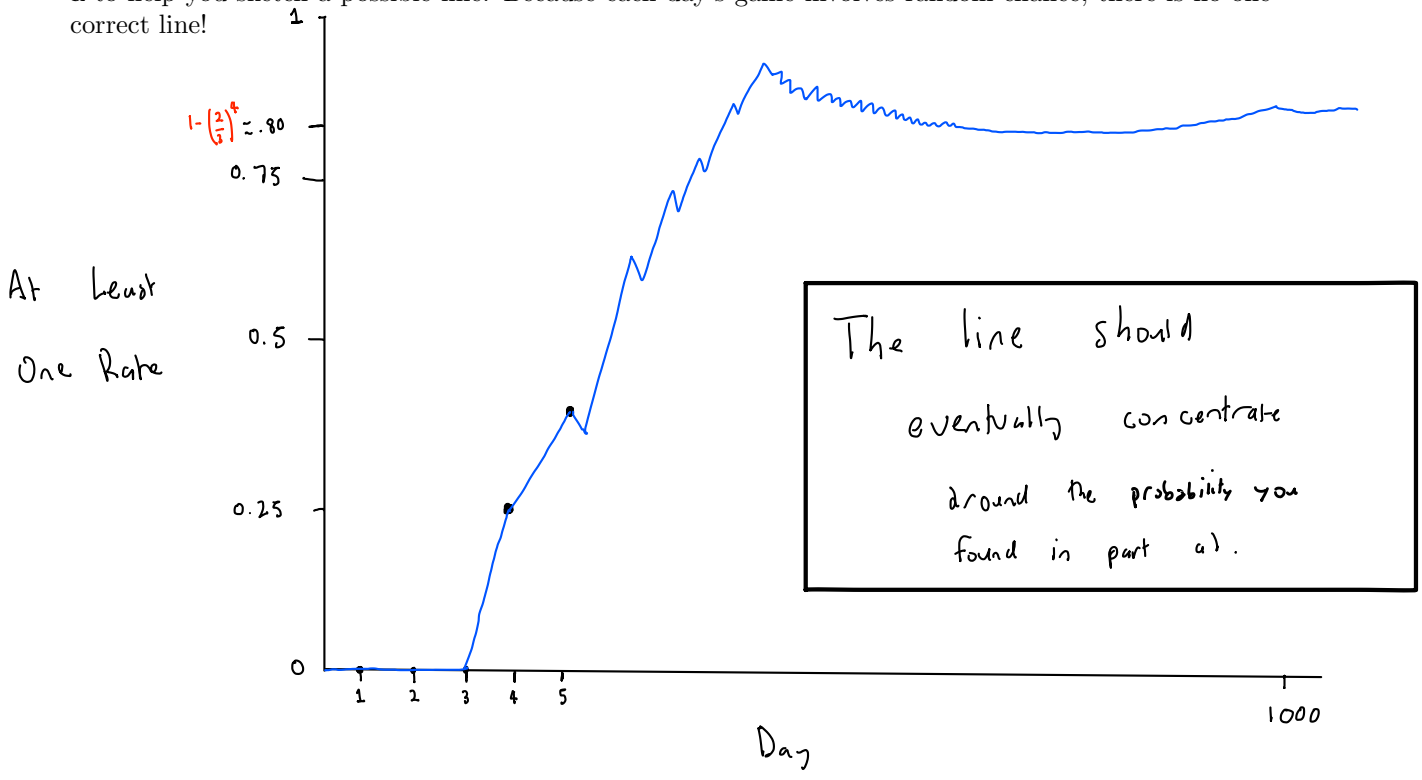
Using the completed `one_thousand_days` function, Bing creates a table called `rps` which contains three columns:

- **Day** (integer): Helps keep track of how many days Bing has been on campus in his simulation; ranges from 1 to 1000.
- **Rounds Won** (integer): How many rounds in Bing's simulated, daily, four-round rock-paper-scissors game that he wins.
- **At Least One Rate** (float): The cumulative proportion of simulated days on which Bing won at least one round. In other words: on Day 4, Bing played his first game in which he won at least one round. Therefore, at the end of Day 4, the number of days on which he had won at least one round was $\frac{1}{4} = 0.25$.

Day	Rounds Won	At Least One Rate
1	0	0
2	0	0
3	0	0
4	3	0.25
5	2	0.4

...(995 rows omitted)

d. (2 points) Bing is interested in visualizing the rate at which he wins at at least one round in a game as the number of days he spends on campus increases. In the space below, *sketch (do not write code!)* a line plot with **Day** on the horizontal axis and **At Least One Rate** on the vertical axis. Make sure you label your axes. Use what we have learned about the Law of Large Numbers and your result in **part a** to help you sketch a possible line. Because each day's game involves random chance, there is no one correct line!



e. (1 point) The skeleton of the code Bing used to add the **At Least One Rate** column to the `rps` table is below. Complete the skeleton.

```
def cumulative_proportion(row_number):

    numer = __________A__________(__________B_________.column('Rounds Won'))
    denom = __________C__________

    return numer/denom

rps = rps.with_column('At Least One Rate',__________D__________)
```

(i) (0.25 points) Blank `A`:

np. count_nonzero

(ii) (0.25 points) Blank `B`:

rps.take(np.arange(row_number))

(iii) (0.25 points) Blank `C`:

row_number

(iv) (0.25 points) Rlank `D`:

rps.apply(cumulative_proportion, 'Day')

4 There's a lot of potential for...*aggressive expansion*! [18 points]

A certain company wants to roll out a new recruiting system which uses AI. An AI recruiter browses the resumes of new applicants and, if the resumes are satisfactory, approves the applicant for a first-round interview. To evaluate the equity of the recruiter along gender lines, the company statisticians collect two random samples, each of size 24, that have been submitted to a particular position over the past year. The first sample contains resumes submitted by applicants who identify as Female; the second sample contains resumes submitted by applications who identify as Male. They run these resumes through the AI recruiting system and use their results to help construct the following table, called `resumes`.

- **Gender** (string): The identified gender of the applicant who submitted the resume.
- **Approved** (string): Whether or not the applicant was approved by the AI recruiter to receive a first-round interview.

Gender	Approved
Male	Yes
Female	Yes
Male	No
Female	No

...(44 rows omitted)

In order to determine whether the AI discriminates along gender lines, the company statisticians decide to perform a hypothesis test with a p-value cutoff of 0.05.

a. (5 points) Write appropriate hypotheses for this test.

(i) (3 points) Null hypothesis:

- ● In the population, the distribution of AI approvals among Male applicants is the same as the distribution of AI approvals among Female applicants.
- ○ In the population, the distribution of AI approvals among Male applicants is different from the distribution of AI approvals among Female applicants.
- ○ The distribution of AI approvals among Male applicants in the `resumes` table is different from the distribution of AI approvals among Female applicants in the `resumes` table.
- ○ The distribution of AI approvals among Male applicants in the `resumes` table is the same as the distribution of AI approvals among Female applicants in the `resumes` table.

(ii) (2 points) Alternative hypothesis:

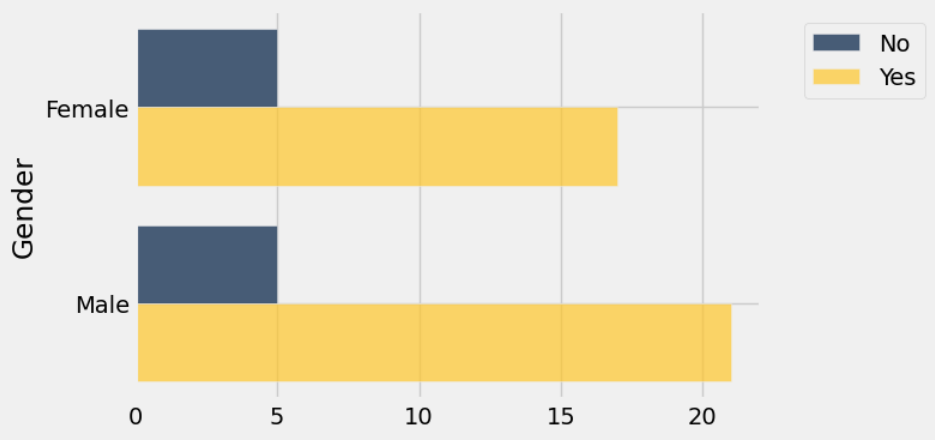
- ○ In the population, the distribution of AI approvals among Male applicants is the same as the distribution of AI approvals among Female applicants.
- ● In the population, the distribution of AI approvals among Male applicants is different from the distribution of AI approvals among Female applicants.
- ○ The distribution of AI approvals among Male applicants in the `resumes` table is different from the distribution of AI approvals among Female applicants in the `resumes` table.
- ○ The distribution of AI approvals among Male applicants in the `resumes` table is the same as the distribution of AI approvals among Female applicants in the `resumes` table.

b. (3 points) Identify an appropriate test statistic for this test.

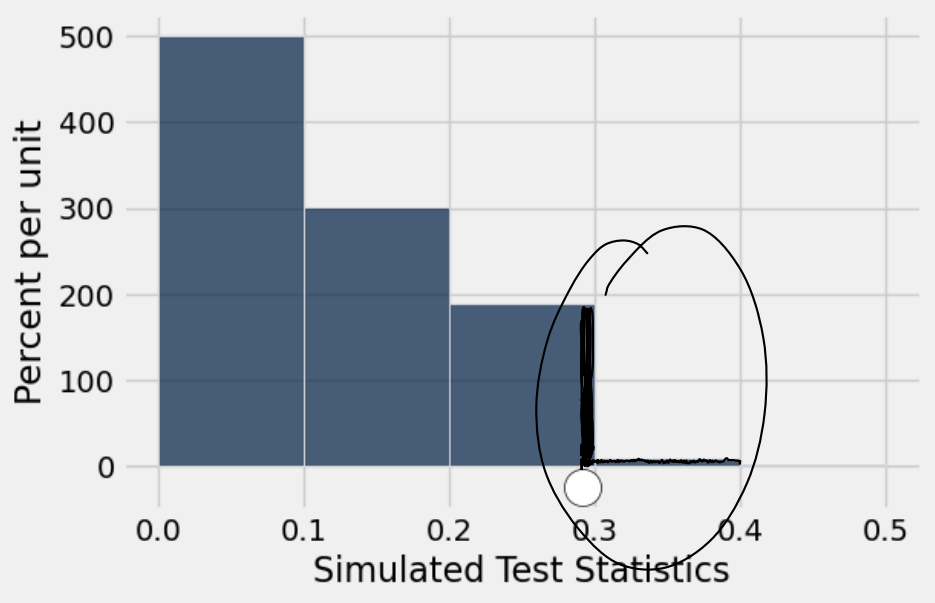
- ◯ The proportion of AI approved applicants
- ◯ The difference in the proportions of AI approved Male applicants and of AI approved Female applicants
- ◯ The proportion of Female applicants
- ● The absolute difference in the proportions of AI approved Male applicants and of AI approved Female applicants

c. (2 points) Below is a visualization of the **Gender** and **Approved** variables belonging to a table which was randomly generated from the original `resumes` table. *True or False*: The generated table has been correctly simulated under the null hypothesis.

◯ True ● False



d. (3 points) Below is a simulated distribution of test statistics under the null hypothesis, along with the value of the observed test statistic, 0.29, which is indicated with a white dot. Shade the area on the distribution corresponding to the p-value.



e. (2 points) The p-value obtained is 0.049. Interpret this value.

- ◯ Given our observed test statistic, the probability that the alternative hypothesis is true is 0.049.
- ◯ Assuming that the alternative hypothesis is true, the probability of seeing the observed test statistic or something larger is 0.049.
- ◯ Given our observed test statistic, the probability that the null hypothesis is true is 0.049.
- ● Assuming that the null hypothesis is true, the probability of seeing the observed test statistic or something larger is 0.049.
- ◯ Assuming that the null hypothesis is true, the probability of seeing the observed test statistic or something smaller is 0.049.

f. (3 points) Should the company statisticians be required to make a definitive conclusion about the results of the hypothesis test, select a conclusion which is appropriate.

- ● We reject the null hypothesis. The evidence is consistent with the idea that in the population, the distributions of AI approvals among Male and Female applicants differ.
- ◯ We retain the null hypothesis. The evidence is consistent with the idea that in the population, the distributions of AI approvals among Male and Female applicants are the same.
- ◯ We reject the null hypothesis. The evidence is consistent with the idea that in the `resumes` table, the distributions of AI approvals among Male and Female applicants differ.
- ◯ We retain the null hypothesis. The evidence is consistent with the idea that in the `resumes` table, the distributions of AI approvals among Male and Female applicants are the same.
- ◯ We reject the null hypothesis and confirm that in the population, the distributions of AI approvals among Male and Female applicants differ.
- ◯ We retain the null hypothesis and confirm that in the population, the distributions of AI approvals among Male and Female applicants are the same.
- ◯ We reject the null hypothesis and confirm that in the `resumes` table, the distributions of AI approvals among Male and Female applicants differ.
- ◯ We retain the null hypothesis and confirm that in the `resumes` table, the distributions of AI approvals among Male and Female applicants are the same.

5 Congratulations! [0 points]

You have completed the Midterm Exam. If you have not been told otherwise, you may bring all of your testing materials (reference sheet and this test paper), as well as your student ID, to the front of the room. Once you have been checked off, you may leave quietly.

- Please make sure that you have written your initials on each page of the exam. **You may lose points on pages where you have not done so.**
- Please make sure you have filled in circles and squares completely rather than having used a check mark, cross or any other mark.
- Double check that you have not skipped over any questions!

Below, you may draw and caption your favorite Data C8 experience or staff member!