DATA C8
Fall 2025

FOUNDATIONS OF DATA SCIENCE

MIDTERM
Solutions

7:10–9:00PM FRIDAY, OCTOBER 17TH 2025

PRINT Your Name: _____

PRINT Your Student ID: _____

PRINT Your Exam Room: _____

PRINT the Name of Person to your Left: _____

PRINT the Name of Person to your Right: _____

PRINT Your GSI's Name (Write N/A if in Self-Service): _____

INSTRUCTIONS

You have **110 minutes** to complete the exam. There are **5 questions** and **17 pages** on this exam, including this cover page.

Question	1	2	3	4	5	Total
Points	21	21	20	21	17	100

- This exam is closed book, closed computer and closed calculator, except the Midterm Reference Sheet provided for you.
- You may only have with you: a pencil, an eraser and your student ID, unless you have pre-approved accommodations.
- If you need to use the restroom, bring your phone, exam, reference sheet and student ID to the front of the room.
- For written questions:
  - Answers written outside the boxes provided will not be graded.
  - Failure to follow instructions will result in no credit.
  - We will grade your answers holistically. If your answer is ambiguous, do not expect to receive credit for it.
- For coding questions with blanks, you may include multiple arguments or functions per blank, but your solution must use every blank available.
- For multiple choice questions, fill in bubbles/squares completely. Read more on these question types below.
- You may assume the `datascience` and `numpy` libraries are imported, as seen in class. Use of **any code** which has not been taught in this offering of the course is prohibited and will not be graded.

Questions with **circular bubbles**: you may select only **1 choice**.

○ Unselected option (completely unfilled)

● Single option selected (completely filled)

Questions with **square boxes**: you may select **1 or more choices**.

■ You may select multiple squares

■ as long as they are completely filled

HONOR CODE: "AS A MEMBER OF THE UC BERKELEY COMMUNITY, I ACT WITH HONESTY, INTEGRITY, AND RESPECT FOR OTHERS."

SIGN Your Name: _____

This page intentionally left blank

The exam begins on the next page.

1. [21.0 points] General

Read each question carefully and answer accordingly. For **State** questions, do not explain or describe your answer.

(a) In 2020, the popular language learning app Duolingo published a blog article describing the process by which it introduces new features. Here is a brief excerpt.

"Test everything." This is one of the key operating principles that we follow at Duolingo in order to continuously improve the learning experience for our users... Think of any feature that you've come across while using Duolingo. Animated skill icons? The result of an experiment. The amount of tears that our owl mascot, Duo, cries in your inbox when you forget to do your lessons? You guessed it.

Specifically, a certain portion of learners... see the current version of the product, while [the other portion] sees the new and updated version. Whichever group seems to respond more positively indicates the version that we should move forward with for all learners. We use [this] approach for two main reasons: 1) it helps us make data-driven product decisions, and 2) if a change doesn't turn out as well as we had hoped, it gives us the opportunity to learn and iterate.

i. [4.0 pts] In order for Duolingo to definitively conclude that a new feature they implement has caused a positive change, how should they decide which learners get to see which version of the product? Explain in one sentence (**15 words max**).

Duolingo needs to assign learners randomly to one of the two groups.

ii. [3.0 pts] **State** the facet of data science which most closely aligns with the work described in the excerpt.

Inference

iii. [2.0 pts] **State** the specific name for the "approach" described in the second paragraph of the excerpt.

A/B Testing **or** Permutation Test **or** Randomized Controlled Trial (RCT)

(b) In 2025, Georgetown Law's Center on Privacy published a special briefing after analyzing statistics reported by the U.S. Customs and Border Protection's (CBP) Office of Field Operations. The Center found that, between 2020 and 2024, Customs and Border Protection agents had collected DNA from 2,000 U.S. citizens. The Center alleged that this violates the Fourth Amendment of the U.S. Constitution. Below is an excerpt from the briefing.

The spreadsheets appear to show each individual as to whom CBP collected DNA and filled out a "DOJ (Department of Justice) Request for National DNA Database Entry Form." The sheets appeared to include multiple entries for some individuals, but CBP's redactions made it difficult to discern which rows pertained to one individual... we attempted various methods to de-duplicate the data. [One such approach involved] *using Python and Stata to code a basic grouping method based on columns showing the individual's age, port name, date, custody status... to identify whether multiple entries belonged to the same person.*

i. [2.0 pts] **State** the facet of data science which most closely aligns with the work described in the excerpt.

Exploration

ii. [1.0 pt] Imagine that the spreadsheets described in the excerpt were loaded into Python as tables. When coding the "*basic grouping method*" described in the excerpt, which of the aggregation methods taught in this class would be appropriate?

- ◯ `pivot` only
- ● `group` only
- ◯ Both `pivot` and `group`
- ◯ Neither of these methods are appropriate.

(c) [3.0 pts] What will the following Python expression output to the screen?

```
make_array(8, 24, 8) + np.arange(8, 24, 8)
```

- ◯ `array([16, 40, 32])`
- ◯ `array([16, 48, 16])`
- ◯ `array([8, 24, 8, 8, 16, 24])`
- ◯ `array([8, 24, 8, 8, 24, 8])`
- ● This expression produces an error.

(d) [3.0 pts] What will the following Python expression output to the screen?

```
make_array(False, False, True) == np.count_nonzero(make_array(True, False, False))
```

- ◯ `True`
- ● `array([False, False, True])`
- ◯ `array([False, False, False])`
- ◯ `False`
- ◯ `array([True, True, False])`
- ◯ This expression produces an error.

(e) For each of the following scenarios, choose the sampling method involved from the items below.

A Deterministic sample	**B** Convenience sample
C Random sample without replacement	**D** Random sample with replacement

i. [1.0 pt] Rolling a fair, six-sided die 100 times.

◯ **A** ◯ **B** ◯ **C** ● **D**

ii. [1.0 pt] Simulating 900 pea plant growths under the null hypothesis that each pea plant has a 75 percent chance of blossoming with purple flowers, independent of other plants.

◯ **A** ◯ **B** ◯ **C** ● **D**

iii. [0.5 pts] A UC Berkeley professor recruits participants for a research study by posting flyers around campus.

◯ **A** ● **B** ◯ **C** ◯ **D**

iv. [0.5 pts] Cal Athletics hires Qualtrics to conduct a survey on student opinions regarding the direction of the football program. Qualtrics is given a roster of the student ID and randomly selects 5,000 students to participate.

◯ **A** ◯ **B** ● **C** ◯ **D**

2. [21.0 points] Berkeley Car Crashes

The California Highway Patrol (CHP) compiles data of vehicle accident reports throughout the state. For this problem, you will be working with a table called **`berkeley`**. This table contains information on all 565 crash reports which took place in the city of Berkeley from January 2025 through September 2025. A three-row excerpt of the **`berkeley`** table lies below.

ID	Time	Type	Day of Week	Highway	Latitude	Longitude	Road 1	Road 2
4591937	Afternoon	Side Swipe	Friday	`True`	37.8821	−122.308	I-80 E/B	Buchanan
4649754	Morning	Rear End	Tuesday	`True`	37.8807	−122.296	Gilman	San Pablo
4742576	Late Night	Rear End	Saturday	`False`	37.8645	−122.302	Bolivar	Potter

(a) [5.0 pts] *Select all columns* that are numerical variables.

- ☐ ID
- ☐ Time
- ☐ Type
- ☐ Day of Week
- ☐ Highway
- ■ Latitude
- ■ Longitude
- ☐ Road 1
- ☐ Road 2

(b) [5.0 pts] Based on the excerpt and the information given, which of the following tasks are appropriate to complete using the `berkeley` table when it comes to the first nine months of 2025? *Select all that apply.*

- ■ Visualizing the distribution of accident types.
- ☐ Conducting a hypothesis test to conclude whether the proportion of highway accidents is equal to 0.5.
- ■ Finding the name of the road most commonly involved in an accident.
- ☐ Finding the exact time (hours and minutes) of each crash that occurred on a Friday.

(c) Write Python code to make a visualization which displays a rough map of the accidents.

```
berkeley.____[A]____(____[B]____)
```

i. [2.0 pts] Fill in blank `[A]`.

```
scatter
```

ii. [2.0 pts] Fill in blank `[B]`.

```
"Longitude", "Latitude" or "Latitude", "Longitude"
```

(d) An intersection is where two roads meet. Write Python code to construct a three-column table that contains the 5 non-highway intersections that had the most crashes, as well as how many crashes occurred at each of these intersections.

```
no_highway_intersections = berkeley.[A]([B]).[C]([D])
five_most_crashes = no_highway_intersections.[E]([F]).[G]([H])
```

i. [0.5 pts] Fill in the blank `[A]`.

```
where
```

ii. [0.5 pts] Fill in the blank `[B]`.

```
"Highway", False
```

iii. [0.5 pts] Fill in the blank `[C]`.

```
group
```

iv. [0.5 pts] Fill in the blank `[D]`.

```
make_array("Road 1", "Road 2")
```

v. [0.5 pts] Fill in the blank `[E]`.

```
sort
```

vi. [0.5 pts] Fill in the blank `[F]`.

```
"count", descending=True
```

vii. [0.5 pts] Fill in the blank `[G]`.

```
take
```

viii. [0.5 pts] Fill in the blank `[H]`.

```
np.arange(5)
```

(e) Write Python code to create a visualization that compares the number of crashes by time of day for all seven days of the week.

`berkeley.` `[A]` `(` `[B]` `).` `[C]`

i. **[1.0 pt]** Fill in the blank `[A]`.

```
pivot
```

ii. **[1.0 pt]** Fill in the blank `[B]`.

```
"Time", "Day of Week"
```

iii. **[1.0 pt]** Fill in the blank `[C]`.

```
barh("Day of Week")
```

3. [20.0 points] Homelessness

Every year, the U.S. Department of Housing and Urban Development publishes a report called AHAR (Annual Homelessness Assessment Report) and presents it to Congress. The table below, called **`homelessness`**, is taken from the 2024 report and features the total number of people experiencing homelessness in each of the fifty states. Below are the first three rows of the table.

State	Homeless Population
.Alabama	4601
.Alaska	2686
.Arizona	14737

In addition, the table **`census`** contains estimates of the total population for each of the fifty states, the District of Columbia and Puerto Rico, as calculated on July 1 of the given year by the U.S Census Bureau. Below are the first three rows of this table.

Geographic Area	2020	2021	2022	2023	2024
Alabama	5033094	5049196	5076181	5117873	5157699
Alaska	733017	734420	734442	736510	740133
Arizona	7187135	7276078	7377566	7473027	7582384

(a) The state names in the State column of the **`homelessness`** table currently having a leading period (`.`), making further analysis difficult. Give the names of the two table methods in the `datascience` library that, when used together, can help add a version of the State column to **`homelessness`** where the periods have been removed.

i. [4.0 pts] Select **only** two table methods.

- ☐ `select`
- ☐ `where`
- ☐ `take`
- ■ `apply`
- ☐ `group`
- ☐ `pivot`
- ☐ `column`
- ■ `with_column`
- ☐ `sort`
- ☐ `join`

(b) Visualize the distribution for the number of homeless per 10,000 across the 50 states in 2024 (note: 'Number of homeless per 10,000' is defined by the proportion of people experiencing homelessness, multiplied by 10,000). You can assume that the homelessness table has been modified to help you as in **part (a)**.

```
fifty_states = homelessness.____[A]____(_____[B]_____)
fifty_states = fifty_states.with_column("Homeless per 10,000",_____[C]_____)
fifty_states._____[D]_____(_____[E]_____)
```

i. [1.0 pt] Fill in the blank `[A]`.

```
join
```

ii. [1.0 pt] Fill in the blank `[B]`.

```
"State", census, "Geographic Area"
```

iii. [1.0 pt] Fill in the blank `[C]`.

```
fifty_states.column("Homeless Population") / fifty_states.column("2024") * 10000
```
or equivalent

Credit was awarded in the exam for any solution that divided by the 2024 column, and multiplied by `10000`.

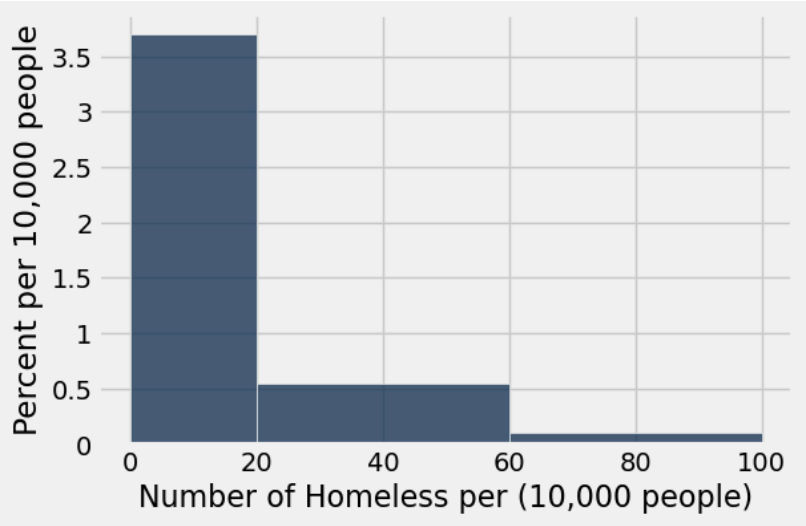
iv. [1.0 pt] Fill in the blank `[D]`.

```
hist
```

v. [1.0 pt] Fill in the blank `[E]`.

```
"Homeless per 10,000"
```

(c) The following is the completed visual from **part (b)**, with labels edited for better readability. The data are separated into three bins: [0, 20), [20, 60) and [60, 100). Answer the items below based on this visual.



i. [3.0 pts] Which bin is most dense?

- (●) $[0, 20)$
- ( ) $[20, 60)$
- ( ) $[60, 100)$
- ( ) An answer cannot be determined.

ii. [3.0 pts] Which bin has the most states in it?

- (●) $[0, 20)$
- ( ) $[20, 60)$
- ( ) $[60, 100)$
- ( ) An answer cannot be determined.

iii. [2.0 pts] Roughly how many states have between 20 and 60 people experiencing homelessness per $10,000$ people?

- ( ) 5
- (●) 10
- ( ) 20
- ( ) 40
- ( ) An answer cannot be determined.

iv. [2.0 pts] Roughly what percentage of states have between 20 and 40 people experiencing homelessness per $10,000$ people?

- ( ) 0.5
- ( ) 5
- ( ) 10
- ( ) 20
- (●) An answer cannot be determined.

v. [1.0 pt] Consider changing the visual so that the [0, 20) bin is split into [0,10) and [10, 20) bins. Which of the following statements are true? *Select all that apply.*

- [■] The combined area of the [0,10) and [10, 20) bins will be equal to the area of the original [0, 20) bin.
- [■] The height of the [0,10) bin may be 0 percent per 10,000 people.
- [■] The height of the [10,20) bin may be greater than the height of the original [0,20) bin.

4. [21.0 points] A Gr8t Game Night

Jaina and Toby are taking a break from studying by organizing a game night for the Data 8 staff. They have come up with a two-player game (Game 1) using four blank index cards. Jaina writes the number 1 on the first card, the number 2 on the second card, the number 3 on the third card, and the number 4 on the fourth card. Below are the rules for one round of Game 1.

> Game 1 Rules
>
> 1. The cards are turned face down and the four-card deck is shuffled.
> 2. Jaina draws a card from the pile and holds onto it.
> 3. Toby draws a card from the pile.
> 4. Jaina and Toby compare the numbers on their cards. Whoever has the highest card wins the game.

(a) [4.0 pts] What is the probability that in five rounds of this game, Jaina never draws the 3 or the 4? Show your work in the box below, and leave your answer as a math expression.

By addition and complement rule,

$$\mathbb{P}[\text{no 3 or 4}] = \frac{1}{2}$$

By multiplication rule,

$$\mathbb{P}[\text{no 3 or 4 in 5 rounds}] = \left(\frac{1}{2}\right)^5$$

(b) Jaina and Toby have compiled a one-column table called results, which contains the names of the winning player ("Jaina" or "Toby") for 1,000 rounds of their game.

i. [4.0 pts] Which type of plot is most appropriate to visualize the data in `results`?

◯ Line plot	● Bar chart	◯ Histogram	◯ Scatter plot
◯ Overlaid line plot	◯ Overlaid bar chart	◯ Overlaid histogram	◯ Overlaid scatter plot

(c) Fill in the blanks of the following statement using the options below. It is possible to use an item more than once.

"By the [1], [2] that Jaina wins the game should be [3] that Jaina wins the game, which is equal to [4].

A Inference facet of data science	**B** Central Limit Theorem	**C** Law of Large Numbers
D the proportion of times	**E** the number of times	**F** the theoretical probability
G close to the proportion of times	**H** equal to the empirical probability	**I** close to the empirical probability
J equal to the theoretical probability	**K** close to the theoretical probability	**L** $\frac{4}{12}$
M $\frac{6}{12}$	**N** $\frac{6}{16}$	**O** $\frac{10}{16}$

i. [1.0 pt] Fill in blank [1].

○ A	○ B	● C	○ D	○ E	○ F
○ G	○ H	○ I	○ J	○ K	○ L
○ M	○ N	○ O			

ii. [1.0 pt] Fill in blank [2].

○ A	○ B	○ C	● D	○ E	○ F
○ G	○ H	○ I	○ J	○ K	○ L
○ M	○ N	○ O			

iii. [1.0 pt] Fill in blank [3].

○ A	○ B	○ C	○ D	○ E	○ F
○ G	○ H	○ I	○ J	● K	○ L
○ M	○ N	○ O			

iv. [1.0 pt] Fill in blank [4].

○ A	○ B	○ C	○ D	○ E	○ F
○ G	○ H	○ I	○ J	○ K	○ L
● M	○ N	○ O			

Toby wants to try something new. He comes up with another two-player game (Game 2) that has ten index cards, numbered 1 through 10. The rules for one round of this game are below.

Game 2 Rules

1. The ten-card deck is shuffled.
2. Three cards are picked out, one by one.
   - The first card goes to Toby.
   - The second card goes to Jaina.
   - The third card is set aside.
3. Toby privately looks at the number on his card.
   - If the number is 5 or less, he takes Jaina's card as his new card.
     - Jaina then takes the third card (that was previously set aside) as her new card.
4. Toby and Jaina flip over their cards and compare them. The player with the highest numbered card wins.

(d) The function `game_2_round()` simulates one round of GAME 2 and returns the name of the winning player. Complete the skeleton code below.

```
def game_2_round():
    deck = np.arange(_____[A]_____)
    cards_drawn = _____[B]_____(deck, 3, replace =_____[C]_____)
    tobys_card = ____[D]____
    jainas_card = ____[E]____
    _____[F]_____:
        tobys_card = ____[G]____
        jainas_card = ____[H]____
    _____[I]_____:
        return "Toby"
    _____[J]_____:
        return "Jaina"
```

i. **[0.5 pts]** Fill in the blank `[A]`.

```
1, 11
```

ii. **[0.5 pts]** Fill in the blank `[B]`.

```
np.random.choice
```

iii. **[0.5 pts]** Fill in the blank `[C]`.

```
False
```

iv. **[0.5 pts]** Fill in the blank `[D]`.

```
cards_drawn.item(0)
```

v. **[0.25 pts]** Fill in the blank `[E]`.

```
cards_drawn.item(1)
```

vi. **[0.25 pts]** Fill in the blank `[F]`.

```
if tobys_card <= 5
```

vii. **[0.25 pts]** Fill in the blank `[G]`.

```
jainas_card or cards_drawn.item(1)
```

viii. **[0.25 pts]** Fill in the blank `[H]`.

```
cards_drawn.item(2)
```

ix. **[0.5 pts]** Fill in the blank `[I]`.

```
if tobys_card > jainas_card
```

x. **[0.5 pts]** Fill in the blank `[J]`.

```
else
```
or equivalent

(e) **[5.0 pts]** Write a function called `game_2_simulation()` that simulates 1000 rounds of Game 2 and returns an array containing the name of the winning player for each round. You may use any functions that have been previously defined and assume they work as intended.

```
def game_2_simulation():
    res = make_array()

    for i in np.arange(1000):
        sim = game_2_round()
        res = np.append(res, sim)

    return res
```

5. [17.0 points] Performative Squirrel Contests

Marissa, Cyrus, and Lena recently spectated Berkeley's Performative Squirrel Contest! During the competition, squirrel contestants were given a score of 1 through 10 based on how performative they were. Looking across a vast sea of matcha lattes, tote bags, wired headphones, our three Data C8 staff members all noticed one thing: contestants with Labubu dolls (Labubus) seemed to be getting higher scores than those without them!

Marissa would like to test, using a **p-value cutoff of 0.05**, whether contestants with Labubus get higher scores than contestants without Labubus in general for all such contests. To collect data for the test, she obtains a random sample of 180 contestants from the Worldwide Performative Squirrel Contest Database; 80 of which carried Labubus and 100 of which did not.

(a) [4.0 pts] Which of the statements below are valid null hypotheses for Marissa's test?

- ◯ In our sample, the distribution of scores for contestants is the same for contestants with Labubus as for contestants without Labubus. Any observed difference in our sample is due to chance.
- ◯ In our sample, the contestants with Labubus have a higher score, on average, than the contestants without Labubus. Any observed difference in our sample is not due to chance.
- ◯ In our sample, the contestants with Labubus have a different average score than the contestants without Labubus. Any observed difference in our sample is not due to chance.
- ● In the population, the distribution of scores for contestants is the same for contestants with Labubus as for contestants without Labubus. Any observed difference in our sample is due to chance.
- ◯ In the population, the contestants with Labubus have a higher score, on average, than the contestants without Labubus. Any observed difference in our sample is not due to chance.
- ◯ None of these statements represent correct null hypotheses.

(b) [3.0 pts] Which of the statements below are valid alternative hypotheses for Marissa's test?

- ◯ In our sample, the distribution of scores for contestants is the same for contestants with Labubus as for contestants without Labubus. Any observed difference in our sample is due to chance.
- ◯ In our sample, the contestants with Labubus have a higher score, on average, than the contestants without Labubus. Any observed difference in our sample is not due to chance.
- ◯ In our sample, the contestants with Labubus have a different average score than the contestants without Labubus. Any observed difference in our sample is not due to chance.
- ◯ In the population, the distribution of scores for contestants is the same for contestants with Labubus as for contestants without Labubus. Any observed difference in our sample is due to chance.
- ● In the population, the contestants with Labubus have a higher score, on average, than the contestants without Labubus. Any observed difference in our sample is not due to chance.
- ◯ None of these statements represent correct alternative hypotheses.

(c) [3.0 pts] Read the incomplete statement below.

"*An appropriate test statistic for this hypothesis test is the* [1] . [2] *values of this statistic support the* ***null*** *hypothesis, while* [3] *values of this statistic support the* ***alternative*** *hypothesis.*"

A	absolute difference between the average score of contestants with Labubus and average score of contestants without Labubus
B	average score of contestants with Labubus minus average score of contestants without Labubus
C	average score of contestants without Labubus minus average score of contestants with Labubus
D	Total Variation Distance between the observed distribution of contestants with and without Labubus and the distribution where half of the contestants have Labubus and the other half do not
E	Low
F	High

Choose from the items above a set of choices which can fill in blanks [1], [2] and [3], respectively. *Select all options that apply.*

- ☐ **A, E, F**
- ☐ **A, F, E**
- ■ **B, E, F**
- ☐ **B, F, E**
- ☐ **C, E, F**
- ■ **C, F, E**
- ☐ **D, F, E**
- ☐ **D, E, F**

(d) [3.0 pts] How should the Data 8 Staff simulate one test statistic under the null hypothesis?

- ○ Create a set of 90 'Labubu' labels and 90 'no Labubu' labels. Sample with replacement from these labels 180 times to get a final set of labels. Shuffle these among the contestants randomly, and then, calculate the test statistic of choice.
- ○ Create a set of 90 'Labubu' labels and 90 'no Labubu' labels. Shuffle these new labels among the contestants randomly, and calculate the test statistic of choice.
- ○ There are 80 'Labubu' labels and 100 'no Labubu' labels in the original table of 180 rows. Sample with replacement from these labels 180 times to get a new set of labels. Shuffle these labels among the contestants randomly, and then, calculate the test statistic of choice.
- ● A correct method for simulating one test statistic under the null hypothesis is not listed here.

(e) [2.0 pts] Marissa obtains a p-value of 0.06. Which of the following statements are true? *Select all that apply.*

- ☐ The probability that the null hypothesis is true, given the observed test statistic, is 0.06.
- ■ The probability of observing a test statistic as rare as or even rarer than the one that Marissa observed, given that the null hypothesis is true, is 0.06.
- ☐ At this stage of the hypothesis test, Marissa is free to set a new p-value cutoff which she thinks is appropriate for this scenario. She can use this p-value cutoff to determine whether she supports the null hypothesis or alternative hypothesis.
- ☐ Cyrus would like to enter his pet squirrel into an upcoming Performative Squirrel Contest. Based on Marissa's original p-value cutoff, Cyrus knows that incorporating a Labubu doll into his squirrel's outfit will directly lead to a higher score.

(f) [2.0 pts] What is the probability that Marissa concludes that the distribution of scores among the two squirrel contestant groups are different, when in reality they are the same?

- ○ 0
- ○ 0.01
- ● 0.05
- ○ 0.06
- ○ 0.10
- ○ An answer cannot be determined.

Congratulations!

You have now 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.

- Make sure you have written your initials on **each page** of the exam, otherwise you may lose points.
- Make sure you have filled in bubbles and squares completely, and that you have not used a checkmark or cross.
- *Double check that you have not skipped over any questions!*

Below, you may draw and caption your favorite Data 8 experience or staff member!