DATA 8
Summer 2024

Foundations of Data Science Final Exam

INSTRUCTIONS

You have 1 hour and 50 minutes to complete the exam.

- The exam is closed book, closed notes, closed computer/calculator, except for the provided 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, exam, reference sheet, and student ID 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 clearly cross out or erase mistakes.

Preliminaries

You can complete these questions before the exam starts.

(i) What is your full name?

(ii) What is your Student ID number?

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

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

1 Potpourri, Finals Edition [20 points]

a. (2 points) In Data 8, we use `percentile(50, array)` and `np.median(array)` interchangeably to calculate the median, and they will always produce the same results.

- ◯ True
- ◯ False

b. (2 points) It is reasonable for us to estimate the maximum value in the population using the bootstrap method, given that the original sample is large enough.

- ◯ True
- ◯ False

c. (2 points) When bootstrapping, we sample from our original sample without replacement to avoid sampling the same row multiple times.

- ◯ True
- ◯ False

d. (2 points) When creating regression lines, minimizing the root mean squared error will always result in the same line as minimizing the mean squared error, assuming both lines are created using the same data.

- ◯ True
- ◯ False

e. (2 points) As a part of evaluating the accuracy of the k-nearest neighbors classifier, each point in the testing set is classified by finding its k-nearest neighbors in the testing set and picking the majority class among these neighbors.

- ◯ True
- ◯ False

f. (2 points) A histogram should be constructed such that the area of each bar is equal to the number of entries in the bin.

- ◯ True
- ◯ False

g. (2 points) Joseph has 8 hats: 4 green, 3 blue, and 1 white. Every day of the week, he picks a hat to wear at random with replacement. The hats are chosen independently of any other day. Which one of the following is the probability that on two consecutive days, he picks the same color hat?

- ◯ $(2 * \frac{4}{8}) + (2 * \frac{3}{8}) + (2 * \frac{1}{8})$
- ◯ $(\frac{4}{8} * \frac{4}{8}) + (\frac{3}{8} * \frac{3}{8}) + (\frac{1}{8} * \frac{1}{8})$
- ◯ $(\frac{4}{8} * \frac{4}{8}) * (\frac{3}{8} * \frac{3}{8}) * (\frac{1}{8} * \frac{1}{8})$
- ◯ $(\frac{4}{8} * \frac{3}{8} * \frac{1}{8}) * (\frac{4}{8} * \frac{3}{8} * \frac{1}{8})$
- ◯ $1 - (\frac{4}{8} * \frac{3}{8} * \frac{1}{8})^2$
- ◯ None of the above

h. (2 points) Mia has a bag with 10 marbles. Each of the 10 marbles has a unique color and an equal probability of being chosen. Mia draws from the bag 10 times with replacement and observes 9 draws of the red marble and 1 draw of the purple marble. After this, Mia believes that each marble does not actually have an equal probability of being chosen. She wants to run a hypothesis test. Which one of the following test statistics should she use?

- ◯ Absolute difference between the number of red and purple marbles
- ◯ Difference between the number of red and purple marbles
- ◯ Number of red marbles
- ◯ Number of purple marbles
- ◯ Number of marbles
- ◯ None of the above

i. (2 points) Ashley has told Data 8 staff that her headshot rate in the game Valorant is 45%. Fiona wants to test this claim, running a hypothesis test. Her alternative hypothesis is that Ashley has **less** than a 45% headshot rate (differences from this in a sample are **not** due to chance). Ashley plays one competitive match, and Fiona observes a headshot rate of 35%. Given the test statistic of (sample headshot rate - expected headshot rate), Fiona claims that larger, more positive values of the test statistic are **in favor** of the alternative hypothesis. Is she correct?

- ◯ True
- ◯ False

j. (2 points) What conditions need to be met in order to invoke the Central Limit Theorem to create a confidence interval for a population parameter? **Select all that apply.**

- ☐ The original population is normally distributed
- ☐ The statistic of interest is the mean or sum
- ☐ The data collected comes from a large and random sample with replacement
- ☐ All of the above
- ☐ None of the above

2 Mini-Crossword Mystery [19 points]

The Data 8 staff have an addiction to *New York Times* mini games, in particular, the daily mini crossword.

a. (2 points) Marissa and Mia want to figure out the ***average time*** it takes Data 8 staff to complete the mini crossword. They only have access to a large, random sample of the staff's past times. Marissa and Mia store their sample into a table named `crosswords`. They want to bootstrap from this original sample. Which of the following will produce valid bootstrap samples? **Select all that apply.**

- ☐ `crosswords.sample()`
- ☐ `crosswords.sample(10)`
- ☐ `crosswords.sample(crosswords.num_rows)`
- ☐ `crosswords.sample(with_replacement = True)`
- ☐ `crosswords.sample(crosswords.num_rows, with_replacement = False)`
- ☐ None of the above

b. (9 points) Marissa and Mia want to create a function named `crosswords_ci` that will conduct 20,000 bootstrap resamples of `crosswords`. The function calculates the average number of seconds it takes to solve a mini crossword for each resample, and returns an array of the left and right endpoints of a 97% confidence interval for the mean time **(in seconds)** that it takes Data 8 staff to solve the mini crossword.

Assume that the `crosswords` table has one column named **Time** with the time **(in minutes)** that it took to solve each crossword. Also, assume that the function `bootstrap_crosswords()` will perform one bootstrap resample on the `crosswords` table.

```
def crosswords_ci():
    cw_times = ________A________
    for i in _______B_______:
        resampled_cw = bootstrap_crosswords()
        stat = _______C_______ # This must be in seconds
        cw_times = _______D_______
    left = _______E_______
    right = _______F_______
    return _______G_______
```

(i) (1 point) Fill in blank (A)

(ii) (1 point) Fill in blank (B)

(iii) (3 points) Fill in blank (C)

(iv) (1 point) Fill in blank (D)

(v) (1 point) Fill in blank (E)

(vi) (1 point) Fill in blank (F)

(vii) (1 point) Fill in blank (G)

c. (2 points) After calling the function, Marissa and Mia generate a 97% confidence interval of [202, 204]. Which one of the following is an appropriate estimate of the probability that the true population mean is in their interval?

- ◯ 0%
- ◯ 3%
- ◯ 97%
- ◯ 98.5%
- ◯ 100%
- ◯ None of the above

d. (2 points) Which of the following can be concluded from Marissa's and Mia's confidence interval in part (c)? **Select all that apply.**

- ☐ Marissa's and Mia's mean completion time was exactly 203 seconds.
- ☐ The mean completion time in their original sample was exactly 203 seconds.
- ☐ 95% of the completion times in the population are between 202 and 204 seconds.
- ☐ 95% of the completion times in the original sample are between 202 and 204 seconds.
- ☐ If another data scientist independently repeats the bootstrap process 1000 times, exactly 950 of the intervals created will contain the true population mean time.
- ☐ None of the above.

e. (2 points) Marissa and Mia are considering changing their confidence level from 97% to 99% for their confidence interval calculation. How would this change affect the width of their confidence interval for the average completion time? **Select all that apply.**

- ☐ The new confidence interval's width would increase.
- ☐ The new confidence interval's width would decrease.
- ☐ The new confidence interval's width would remain the same.
- ☐ The new confidence interval would contain 202 seconds.
- ☐ The new confidence interval would contain 206 seconds.
- ☐ The effect on the new confidence interval width cannot be determined from the given information.

f. (2 points) Now, Marissa and Mia want to figure out the *proportion of times that the Data 8 staff complete the mini crossword in under one minute*, but only find it practical to take a large, random sample of the staff's past times. They want to construct a **95%** confidence interval for this proportion with a total width of only 2 percent. What is the smallest sample size they should use? **Please show all of your work and put a box around your final answer.**

3 Rocket League Regression [40 points]

In his free time, Conan likes to play Rocket League, an online multiplayer game where players play soccer but with rocket-powered cars that can jump, flip, and even fly! Conan aspires to be in the top 0.1% of players, so he plays 50 games of Rocket League and records his game statistics. Each row in the `rocket` table represents one game that Conan played, and the table includes the following four columns:

- **Score**: (int) The number of points Conan had in the game. Players earn points in various ways, including scoring a goal, making a save, taking a shot on net, etc.
- **Touches**: (int) The number of times Conan's car touched the ball.
- **Boost**: (int) The amount of boost Conan's car used in the game.
- **Won**: (Boolean) A value indicating whether Conan's team won (True) or lost (False) the game.

Score	Touches	Boost	Won
256	24	1870	False
280	38	2000	True
472	44	2276	False
284	32	2140	True
299	28	1749	False

(... 45 rows omitted)

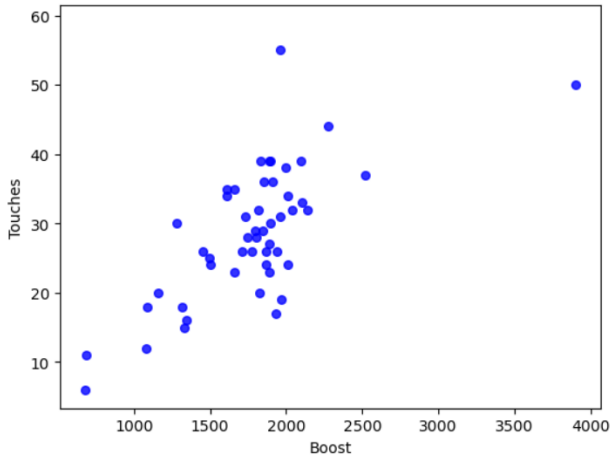
a. (2 points) Conan wants to examine how the distribution of touches varies by whether Conan's team won the game. Write one line of code that creates the most appropriate visualization for these data.

```
rocket._____A_____(_____B_____)
```

(i) (1 point) Fill in blank (A)

(ii) (1 point) Fill in blank (B)

b. (4 points) An important part of the game involves collecting boost that is scattered around the field, as boost fuels the rocket-powered cars, allowing them to go faster and propel them into the air. Conan wants to increase the number of ball touches he has in a given game, as touching the ball frequently can increase offensive pressure, ultimately increasing his chances of winning the game. He suspects that the amount of boost he uses in a game affects the number of ball touches he makes. In order to visualize his data, he creates the following scatter plot.



Use the scatter plot to answer the following questions:

(i) (2 points) Which one of the following is the best estimate for the correlation coefficient between Boost and Touches?

- ◯ Less than 0
- ◯ Exactly 0
- ◯ Greater than 0
- ◯ We cannot determine this from the scatter plot alone

(ii) (2 points) The standard deviation of **Boost** is greater than the standard deviation of **Touches**.

- ◯ True
- ◯ False

c. (7 points) Conan wants to construct a regression line to predict the number of ball touches he makes from the amount of boost he used in a given game. He finds calculating the correlation coefficient tedious, and wants to explore new ways that are potentially faster.

(i) (4 points) First, help Conan add two new columns to the rocket table, **Touches_su** and **Boost_su**, which are the Touches and Boost columns in standard units, respectively.

```
for col_name in ______A______:
    arr = rocket.column(col_name)
    avg = np.mean(arr)
    sd = ______B______(arr)
    rocket = rocket.with_columns(______C______, ______D______)
```

A. (1 point) Fill in blank (A)

B. (1 point) Fill in blank (B)

C. (1 point) Fill in blank (C)

D. (1 point) Fill in blank (D)

(ii) (3 points) While you help Conan, he creates a function named `weird_multiply` which takes in a row object and returns the product of the elements in index 0 and index 1.

```
def weird_multiply(row):
      return row.item(0) * row.item(1)
```

Use the function above to calculate the correlation coefficient between Touches and Boost.

```
rocket_su = rocket._____A_____("Touches_su", "Boost_su")
r = _____B_____(rocket_su._____C_____)
```

A. (1 point) Fill in blank (A)

B. (1 point) Fill in blank (B)

C. (1 point) Fill in blank (C)

d. (8 points) We find the correlation coefficient between Touches and Boost to be approximately 0.705. We also find that across the 50 games:

- The average number of Touches was 28.54 with a standard deviation of 9.51.
- The average of Boost used was 1773.4 with a standard deviation of 471.7.

(i) (2 points) A correlation coefficient of 0.705 suggests that there must be a linear relationship between Touches and Boost.

◯ True
◯ False

(ii) (2 points) If we fit a regression line to the scatter plot, the sum of the residuals will be _____**A**_____ and the sum of the squared residuals will be _____**B**_____.

◯ A: zero, B: zero
◯ A: non-zero, B: zero
◯ A: zero, B: non-zero
◯ A: non-zero, B: non-zero
◯ We do not have enough information to answer this.

(iii) (2 points) The prediction interval for the number of touches made at Boost = 2000 is likely to be _____ the prediction interval made at Boost = 1000. **Select all that apply.**

☐ The same as
☐ Wider than
☐ The same in width as
☐ Narrower than
☐ None of the above

(iv) (2 points) The regression line that minimizes the RMSE for predicting Touches from Boost is always the same as the regression line that minimizes the RMSE for predicting Boost from Touches.

◯ True
◯ False

e. (7 points) Conan wants to construct a regression line. However, he refuses to use the regression equations because he does not believe that it produces a good regression line. Instead, he decides to use his computer to find a regression line that minimizes the RMSE.

(i) (3 points) First, help Conan define the function `rmse` that takes in a slope and intercept, and returns the RMSE between the predictions made by the regression line and the actual y values.

```
x = rocket.column("Boost")
y = rocket.column("Touches")
def rmse(slope, intercept):
    pred = ______A______
    residuals = ______B______
    squared_resid = residuals ** 2
    return ______C______
```

A. (1 point) Fill in blank (A):

B. (1 point) Fill in blank (B)

C. (1 point) Fill in blank (C)

(ii) (2 points) Write a line of code that assigns `best_params` to a two-element array containing the slope and intercept for the regression line that minimizes the RMSE.

```
best_params = ______A______(______B______)
```

A. (1 point) Fill in blank (A)

B. (1 point) Fill in blank (B)

(iii) (2 points) What is your best estimate of `best_params.item(0)`? If you believe you lack the information needed, write "Not enough information". You may incorporate any numbers and statistics introduced in previous subparts. **You do not need to simplify your answer.**

f. (12 points) Mia has watched Conan prioritize grabbing boost over hitting the ball on multiple occasions. Therefore, she believes that there is no linear relationship between **Boost** and **Touches**. On the other hand, Conan believes that there is a linear relationship between **Boost** and **Touches**. In order to settle this dispute, they decide to conduct a hypothesis test.

(i) (2 points) Formulate a valid null hypothesis.

(ii) (2 points) Formulate a valid alternative hypothesis.

(iii) (1 point) Choose all valid test statistics for the hypothesis test. **Select all that apply.**

- ☐ Total variation distance
- ☐ Difference in means
- ☐ Intercept
- ☐ Correlation coefficient
- ☐ None of the above

(iv) (1 point) Choose all valid simulation methods for the hypothesis test. **Select all that apply.**

- ☐ Conduct a bootstrap resample and compute the test statistic.
- ☐ Sample from the `rocket` table without replacement, and compute the test statistic.
- ☐ Shuffle the Boost column and compute the test statistic.
- ☐ Shuffle the Touches column and compute the test statistic.
- ☐ None of the above.

(v) (3 points) Conan and Mia decide to use a p-value cutoff of 10% for their hypothesis test. They perform the first part of their hypothesis test, and obtain 10,000 simulated statistics stored in an array called `simulated_stats`. Fill in the blanks so that the code prints the correct conclusion for their hypothesis test.

```
left = percentile(_____A_____, simulated_stats)
right = percentile(_____B_____, simulated_stats)
if _____C_____:
     print("Fail to Reject the Null Hypothesis")
else:
     print("Reject the Null Hypothesis")
```

A. (1 point) Fill in blank (A)

B. (1 point) Fill in blank (B)

C. (1 point) Fill in blank (C)

(vi) (3 points) After running the code from part (v), Conan and Mia find that left and right values are 0.578 and 0.814, respectively. **Select all that apply.**

- ☐ Using a p-value cutoff of 10%, Conan and Mia should reject the null hypothesis.
- ☐ Using a p-value cutoff of 5%, Conan and Mia should reject the null hypothesis.
- ☐ There is approximately a 10% chance that the next simulated statistic Conan and Mia calculate is between 0.578 and 0.814.
- ☐ There is approximately a 90% chance that the next simulated statistic Conan and Mia calculate is between 0.578 and 0.814.
- ☐ There is a 10% chance that Conan and Mia falsely reject the null hypothesis when it is actually true.
- ☐ None of the above.

4 Lila? An Evil Lila?! [11 points]

As self-nominated Protector of the Garden, Cynthia's dog Lila has been in a long standing battle with the enemy Squirrels, who outnumber her greatly. To bolster her numbers, Cynthia decides to clone Lila, but the machine is not perfect. 17% of the time, it creates an identical but Evil Lila clone.

Confronted with a mixed army of Good and Evil Lilas, Cynthia develops a Lila scanner to identify the Evil clones. If the Lila clone is Good, the scanner will return an accurate result 96% of the time. If the Lila clone is Evil, the scanner will return an accurate result 93% of the time.

For this section, you may leave any of your answers unsimplified or as mathematical expressions. Please also put a box around your final answer for each of the questions below.

a. (0 points) SCRATCH WORK: You can use this space to write any extra calculations or diagrams that may be helpful. Anything written in this box will not be graded. Alternatively, use this space to draw your interpretation of an Evil Lila (she is your typical small fluffy white dog)!

b. (2 points) What is the probability that 6 Good Lilas are created in a row?

c. (3 points) If Cynthia scans a Lila clone at random, what is the probability that the scanner says the Lila clone is Good?

d. (3 points) Suppose the scanner says a Lila clone is Evil. What is the probability that the Lila clone is actually Evil?

e. (3 points) Aha! Cynthia stumbles upon clone #8, who seems to be asleep in the perfect position to be scanned. Prior to scanning, the position of the clone #8 leads Cynthia to believe that there's a 30% chance the clone is Evil. Upon scanning, the scanner reads "Good".

Given the information in this question and assuming the conditional probabilities in the problem statement are still valid, what is Cynthia's subjective probability that the clone is actually Good?

- ◯ $0.30 * 0.07 \ + \ 0.70 * 0.96$
- ◯ $0.70 * 0.04 \ + \ 0.30 * 0.93$
- ◯ $\frac{0.83*0.96}{0.83*0.96 \ + \ 0.17*0.07}$
- ◯ $\frac{0.70*0.96}{0.70*0.96 \ + \ 0.30*0.07}$
- ◯ $\frac{0.30*0.70}{0.70*0.96 \ + \ 0.30*0.07}$
- ◯ None of the above

5 The Olympi*k*s [25 points]

The Paris 2024 Olympics are here, and you are tasked with predicting whether a sports event is a Track event or not based on past data. You have a table named `olympics` containing information about various sports events from previous Olympic Games. The table contains the following columns:

- **event**: (string) The name of the sports event.
- **duration**: (float) The duration of the event in minutes.
- **participants**: (int) The number of participants in the event.
- **spectators**: (int) The number of spectators watching the event.
- **is_track**: (Boolean) Whether the event is a Track event (True) or not (False).

Here is a sample of the dataset:

event	duration	participants	spectators	is_track
100m Dash	0.17	8	30,000	True
Long Jump	0.5	12	15,000	True
200m Freestyle	1.83	8	20,000	False
Basketball	38	24	40,000	False

(... 996 rows omitted)

Answer the following questions based on this dataset using the k-Nearest Neighbors algorithm.

a. (5 points) **Data Preparation**

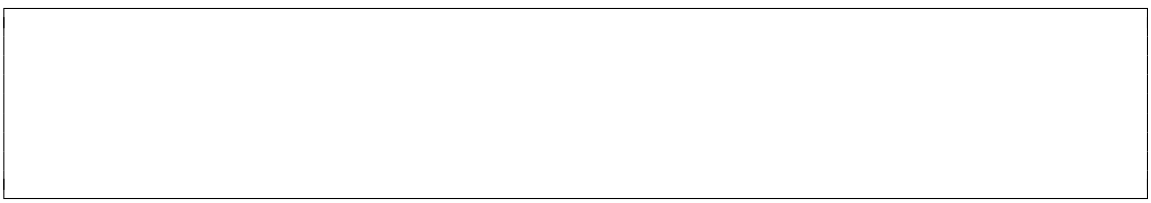
To implement the k-NN algorithm, we first need to standardize the **duration**, **participants**, and **spectators** columns. Fill in the blanks to create a new table `standardized_olympics` where the columns are standardized to have a mean of 0 and a standard deviation of 1.

```
import numpy as np
from datascience import *

def standardize_column(column):
    return ______A______

standardized_olympics = Table()
for label in olympics.______B______:
    standardized_olympics = standardized_olympics.with_column(label, ______C______)
```

(i) (2 points) Fill in blank (A)



(ii) (1 point) Fill in blank (B)

(iii) (2 points) Fill in blank (C)

b. (11 points) ***k*-Nearest Neighbors Classification**

Implement a function named `knn_classify` that takes in the standardized table, a new event (represented as a standardized array of **duration, participants,** and **spectators**), and a value for k, and returns the predicted type of the event. Fill in the blanks to complete the function.

```
def distance(row):
    return ______A______

def knn_classify(table, new_event, k):
    table_with_distances = table.with_column("distance", ______B______)
    k_rows = table_with_distances.sort("distance").______C______
    next_step = k_rows.group("is_track")
    common_neighbor = next_step.______D______.row(0).______E______
    return common_neighbor
```

(i) (3 points) Fill in blank (A)

(ii) (2 points) Fill in blank (B)

(iii) (2 points) Fill in blank (C)

(iv) (2 points) Fill in blank (D)

(v) (2 points) Fill in blank (E)

c. (6 points) **Prediction Example**

(i) (2 points) In Data 8, when using k-nearest neighbors, we pick an even k value, so each class has an equal chance of being selected.

- ◯ True
- ◯ False

(ii) (2 points) When using k-nearest neighbors with the same dataset, a prediction model based on standard units will always produce the same results as a prediction model based on non-standardized units.

- ◯ True
- ◯ False

(iii) (2 points) If we want to build a model to predict 3 classes instead of 2, we can use the same method of picking k that we discussed in class to avoid ties.

- ◯ True
- ◯ False

d. (3 points) **Class Imbalance**

(i) (2 points) There are 329 total events in the Paris 2024 Olympics, with 48 being track events. What is the smallest value of k (greater than 0) that we can pick that will always give us the same label regardless of our input row?

(ii) (1 point) If n is the total number of rows, w is the number of elements in the majority class, and m is the number of elements of the minority class (in a binary classification task), what is the general formula for k that we can pick that will always give us the same label regardless of our input row? You must use k, a comparison operator (such as $>$, $>=$, $=$, $<=$, $<$), and a mathematical expression in terms of n and/or m and/or w.

6 Optional [0 points]

a. (0 points) **Assumptions**
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. **We will only consider assumptions that are written inside the box below.**

b. (0 points) **Fun Drawing**
Draw and caption your favorite Data 8 experience or staff member!