DATA 8 Final Exam, Spring 2024

Instructions
You have 2 hours and 50 minutes to complete the exam.

- The exam is closed book, closed notes, closed computer, closed calculator, except 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, and student ID to the front of the room.
- The test is designed to be completed using methods we have learned in this class. We reserve the right to deduct or not score answers using methods out of scope.

For multiple choice questions with circles, you should select exactly one choice. You should indicate your selection by **completely** filling in the circle.

◯ You must choose either this option

◯ Or this one, but not both!

For multiple choice questions with square checkboxes, you may select multiple choices. You should indicate your selection by **completely** filling in the box.

☐ You could select this choice.

☐ You could select this one, too!

Please write your initials at the top of every page as you are taking the exam.

Question	Points	Score
1	22	
2	26	
3	26	
4	22	
5	26	
6	0	
Total:	122	

Name: _____

TA's name: _____

Student ID: _____

Name of person to your left: _____

Name of person to your right: _____

1. **True or False**

(a) (2 points) All of the work on this exam is your own.

√ **True**

◯ False

(b) (2 points) The posterior probability is defined as the probability of an event before updating it with additional information.

◯ True

√ **False**

(c) (2 points) When constructing a confidence interval for the sample mean, using a 95% confidence level indicates that the confidence interval will capture the individual values of 95% of the population of interest.

◯ True

√ **False**

(d) (2 points) When running a hypothesis test using a p-value cutoff of 4%, given that the null hypothesis is true, the probability the test will reach the correct conclusion is 96%.

√ **True**

◯ False

(e) (2 points) In an A/B test, shuffling the labels and shuffling the values are equally valid methods to test for a difference between groups.

√ **True**

◯ False

(f) (2 points) The intercept of a line of best fit can be interpreted as the predicted increase in y for a zero-unit increase in x.

◯ True

√ **False**

(g) (2 points) A decision boundary for a k-nearest-neighbors classifier may fail to completely separate the 2 classes represented in the training set (i.e. not achieve perfect accuracy on the training set).

√ **True**

◯ False

(h) (2 points) In the lecture on privacy, access (with regards to fair information practices) is defined as the ability for companies to access information about individuals.

◯ True

√ **False**

(i) (2 points) In Professor Sahai's case study lecture, machine learning is differentiated from classical statistics because it involves generalized intelligence whereas classical statistics is focused on solving specific equations.

◯ True

√ **False**

(j) (2 points) By converting data into standard units, the data is transformed into a normal distribution with mean 0 and a standard deviation of 1.

◯ True

√ **False**

(k) (2 points) If A is a random event, the probability of A given no information must be smaller than the probability of A given strictly more information.

◯ True

√ **False**

2. **Psychic Police?**

Shawn Spencer, a consultant for the SBPD, claims to have psychic abilities to solve crimes. Shawn's partner, Gus, wants to analyze the probability of Shawn making a correct "prediction" based on different scenarios.

Suppose Shawn claims that he is able to use his psychic abilities to sense the type of crime. In this case, we assume there are only two possible types of crime: theft and fraud, and that all crimes occur independently of one another. Theft crimes occur with a 70% chance, and fraud has a 30% chance.

Additionally, assume the following:

- If the crime is theft, there is a 65% chance that Shawn correctly predicts the type of crime.
- If the crime is fraud, there is an 80% chance that Shawn correctly predicts the type of crime.

For all answers, please leave it as a mathematical expression.

(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.

Solution:

(b) (2 points) What is the probability that 5 crimes in a row are all thefts?

Solution: 0.7^5

(c) (3 points) For a case chosen at random, what is the probability that Shawn correctly predicts the type of crime?

Solution: $0.65 * 0.7 + 0.8 * 0.3$

(d) (3 points) Suppose Shawn predicts a crime is theft. What is the probability that the crime type is actually theft?

Solution: $(0.65 * 0.7)/(0.65 * 0.7 + 0.2 * 0.3)$

(e) (2 points) Suppose a crime is fraud. What is the probability that Shawn incorrectly predicts it as theft?

Solution: 0.2

(f) (2 points) Gus compiles 8 random crimes for Shawn to analyze. What is the probability that at least one of these crimes is theft?

Solution: $1 - 0.3^8$

(g) (2 points) What is the probability that a crime is fraud and Shawn predicts correctly?

Solution: $0.3 * 0.8$

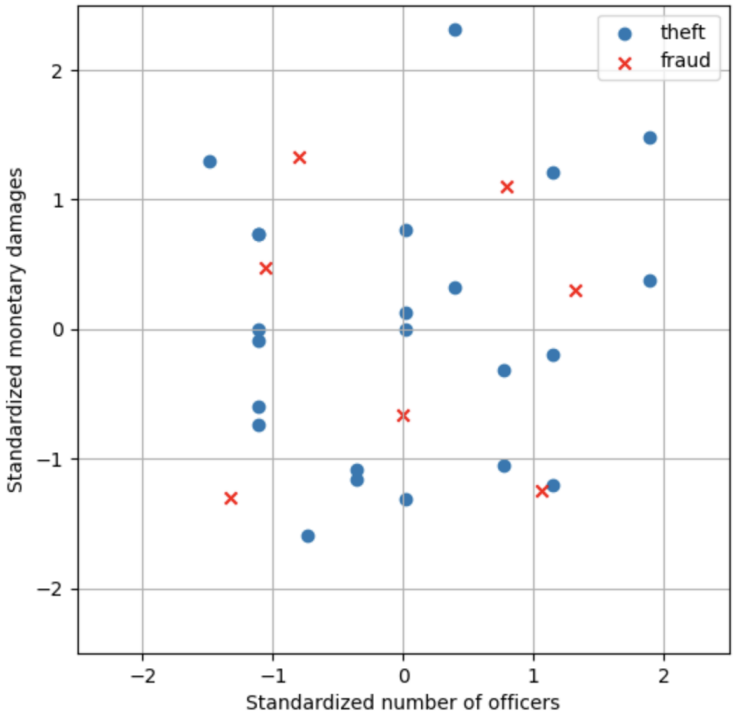
(h) Gus fires Shawn and decides to outsource his predictions to a k-nearest-neighbors classifier. He will predict the type of crime based on the geographic coordinates of the crime, the monetary value of the damages, and the amount of officers called to the scene.

i. (3 points) Which of the following are steps he should take to implement the classifier? Select all that apply.

- √ **Separate his data into a training and test set**
- ☐ Calculate the distance from all points in the training set to one another
- √ **Reserve some data to evaluate the accuracy of the classifier**
- ☐ Minimize RMSE of errors on the test set to optimize the classifier
- ☐ Repeatedly adjust the features and values of k depending on the results of rerunning the classifier on the test set
- ☐ None of the above

ii. (2 points) Gus decides to only use 2 features, the monetary value of the damages and the amount of officers at the scene. What are problems that might arise from using these two variables? Select all that apply.

- ☐ The amount of officers is not a numerical variable since it cannot take on decimal values.
- ☐ The variables can be positively correlated, which creates a confounding factor.
- √ **The variables have different scales, so one might take more importance than the other in the classifier, skewing the results.**
- √ **The variables are measured in different units, so one might take more importance than the other in the classifier, skewing the results.**
- ☐ None of the above



iii. (2 points) Above is a plot of the entire training data. Using k=3, what would a point located at (0, 0) be classified as?

- √ **Theft**
- ◯ Fraud
- ◯ Cannot determine

iv. (2 points) Using k=15, what would a point located at (2, -1) be classified as?

- √ **Theft**
- ◯ Fraud
- ◯ Cannot determine

v. (3 points) Which of the following are true statements about the value of k for Gus's classifier? Select all that apply.

- ☐ Increasing the value of k always improves accuracy as it allows more data to be used.
- √ **Odd values guarantee a decision can be made.**
- √ **k=19 is too large for the dataset above.**
- √ **k=1 is equivalent to classifying a point based on its neighbor with the smallest distance.**
- ☐ In a situation where one class has many more points than the other, the k value should not exceed the number of points in the minority class.
- ☐ None of the above

3. **Minions Merchandise**

Following the release of the latest "Minions" movie, Universal Studios noted a high demand for Minions-themed merchandise.

(a) (3 points) Kevin is analyzing the individual purchase amounts of Minions-themed merchandise made in 2023. He wants to create a 95% confidence interval for the population mean with a total width no larger than \$1. If the population standard deviation is known to be \$10, calculate the minimum sample size Kevin needs to achieve this confidence interval width. Please draw a box around your final answer.

Solution: 1600

(b) (2 points) Kevin is debating between using a 90% confidence interval and a 95% confidence interval. Which of the following are true statements about the confidence level? Select all that apply.

- √ **Given the same dataset and desired parameter, the 90% confidence interval will likely be narrower than the 95% confidence interval.**
- ☐ The 90% confidence level should not be used as it does not align with the Normal distribution, whereas the 95% confidence interval would correspond to exactly 2 SDs above and below the sample mean.
- √ **The 90% confidence interval would correspond to a hypothesis test with p-value cutoff of 0.1.**
- ☐ The 95% confidence interval would correspond to a hypothesis test with p-value cutoff of 0.95.
- ☐ Since the 95% confidence interval has a higher chance of capturing the true parameter, it should always be used instead of a 90% confidence interval.
- ☐ None of the above

(c) (2 points) Kevin decides to use the bootstrap method to create a 95% confidence interval using a sample size smaller than the one calculated in part (a). Another analyst, Bob, claims that the confidence interval obtained through bootstrapping will most likely have a width greater than \$1. Which of the following is true?

- √ **This question was not graded due to ambiguity in the wording.**
- ◯ True, because bootstrapping with a smaller sample size generally results in less precise estimates, hence a wider confidence interval.
- ◯ False, as the width of the confidence interval depends on the population standard deviation.
- ◯ True, but only if the variability in the sample is unusually high, leading to a wider interval.
- ◯ False, as the bootstrap method inherently provides narrower confidence intervals regardless of sample size.

(d) Fill in the code below such that the function `calculate_ci` returns an array containing the left and right endpoints for a confidence interval for the population mean. The function takes in a table with data (`tbl`), the name of the column containing the values of interest as a string (`values_col`), the confidence level (expressed as a percent from 0 to 100) (`level`), and the number of repetitions as an integer (`repetitions`).

```
def calculate_ci(tbl, values_col, level, repetitions):
    simulated_means = ___(a)___
    for i in np.arange(repetitions):
        resampled_table = tbl.___(b)___
        resampled_values = resampled_table.column(values_col)
        resampled_mean = np.mean(___(c)___)
        simulated_means = np.append(___(d)___, ___(e)___)

    left = percentile(___(f)___, simulated_means)
    right = percentile(___(g)___, simulated_means)
    return make_array(left, right)
```

i. (1 point) Fill in blank (a).

Solution: make_array()

ii. (2 points) Fill in blank (b).

Solution: sample()

iii. (2 points) Fill in blank (c).

Solution: resampled_values

iv. (1 point) Fill in blank (d).

Solution: simulated_means

v. (1 point) Fill in blank (e).

Solution: resampled_mean

vi. (2 points) Fill in blank (f).

Solution: (100 - level) / 2

vii. (2 points) Fill in blank (g).

Solution: 100 - (100 - level) / 2

(e) Assume Kevin used the sample size from part (a) and obtained a 95% confidence interval. For the following two questions, assume that the confidence interval he constructs for the mean purchase amount is [$25, $35].

i. (2 points) What is the probability that the actual population mean is outside this interval?

- ◯ 2.5%
- ◯ 5%
- ◯ 95%
- ◯ There is not enough information to determine because we don't know the exact distribution of the data.
- √ **None of the above**

ii. (2 points) Which of the following can be concluded from the bootstrapped confidence interval above?

- ◯ 95% of the purchases in the population are between $25 and $35.
- ◯ The mean purchase amount in Kevin's sample was exactly $30.
- ◯ If Bob independently repeats Kevin's process 1000 times, exactly 950 of the intervals he creates will contain the true population mean.
- √ **None of the above**

iii. (2 points) Kevin thinks that the average price of Minions-merch transactions is less than the $45 average price of Disney princess merchandise transactions. Based on his confidence interval and a p-value cut-off of 5%, what can Kevin conclude?

- ◯ The data are consistent with the hypothesis that the distribution of the transaction prices is the same for both Minions merchandise and Disney princess merchandise.
- √ **The data are consistent with the hypothesis that Minions merchandise is less expensive than Disney princess merchandise**
- ◯ The data are consistent with the hypothesis that Minions merchandise is more expensive than Disney princess merchandise.
- ◯ There is not enough information to make a conclusion of any kind.

(f) (2 points) Kevin is now interested in estimating the 75th percentile of the Minions merchandise sales for a better understanding of high-end sales performance. Which of the following methods could he use to create a 95% confidence interval for this percentile? Select all that apply.

- √ **Bootstrapping**
- ☐ Central Limit Theorem
- ☐ Randomized Control Experiment
- ☐ Linear regression prediction interval
- ☐ None of the above

(g) (0 points) OPTIONAL: Draw a picture of some Data 8 themed merchandise representing your experience this semester! Make sure to finish the rest of the exam, though.

Solution: :D

4. **Streaming Success**

Hannah and Lucas are using early streaming data to predict the chart success of pop songs. The dataset, named `pop_songs`, consists of data from 200 randomly selected pop songs released in the past year. The dataset includes the following columns:

- `Title`: a string, the name of the song.
- `Artist`: a string, the name of the artist.
- `Streams`: an integer, the number of streams a song received within the first 24 hours of its release.
- `Peak_Position`: an integer, the highest chart position the song achieved, with 1 being the highest.

(a) (2 points) Fill in blank (a) such that `compute_su` returns the input array in standard units.

```
def compute_su(array):
    mean = np.mean(array)
    sd = np.std(array)
    ___(a)___
```

Solution: return (array - mean)/sd

(b) Fill in the code to calculate the correlation coefficient between the number of streams a song received within the first 24 hours and the highest chart position the song achieved.

```
streams_in_su = compute_su(___(a)___)
chart_in_su = compute_su(___(b)___)
r = ___(c)___
```

i. (1 point) Fill in blank (a).

Solution: pop_songs.column("Streams")

ii. (1 point) Fill in blank (b).

Solution: pop_songs.column("Peak_Position")

iii. (2 points) Fill in blank (c).

Solution: np.mean(streams_in_su * chart_in_su)

(c) (2 points) Assume that Hannah calculated `r` correctly, and got -0.85. Given that the standard deviation for `Streams` is 120,000 streams and for `Peak_Position` is 250 positions, fill in blank (a) to compute the slope of the regression line in original units. (Reminder: Hannah is using the number of streams in the first 24 hours to predict the peak position.)

`slope = ____(a)____`

Solution: -0.85 * 250/120000

(d) Hannah is interested in whether there is a linear relationship between `Streams` and `Peak_Position`, so she decides to run a hypothesis test.

i. (2 points) Which of the following are valid null hypotheses? Select all that apply.

- √ **The true correlation coefficient between `Streams` and `Peak_Position` is 0.**
- ☐ The true correlation coefficient between `Streams` and `Peak_Position` is not 0.
- √ **The true slope between `Streams` and `Peak_Position` is 0.**
- ☐ The true slope between `Streams` and `Peak_Position` is not 0.
- ☐ None of the above

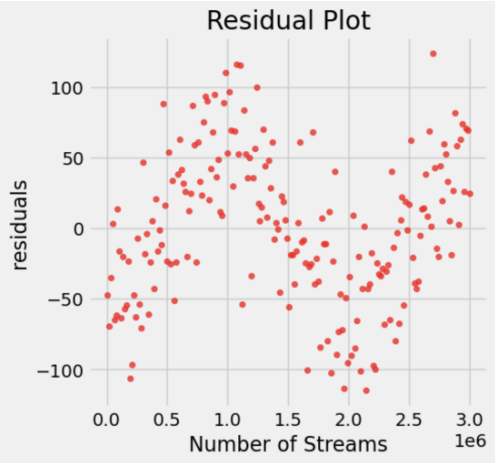
ii. (2 points) Which of the following are valid alternative hypotheses for this hypothesis test? Select all that apply.

- ☐ There is a positive linear relationship between `Streams` and `Peak_Position`.
- ☐ There is a negative linear relationship between `Streams` and `Peak_Position`.
- ☐ The correlation coefficient between `Streams` and `Peak_Position` is zero.
- √ **The correlation coefficient between `Streams` and `Peak_Position` is not zero.**

iii. (2 points) Hannah decides to use bootstrapping to test her hypothesis. She creates a 90% confidence interval for the true correlation coefficient between `Streams` and `Peak_Position`. The 90% confidence interval is: [-0.15, -0.55]. Using a p-value cutoff of 10%, Which of the following is true? Select all that apply.

- ☐ 90% of the data we observe can be explained by the regression line.
- ☐ The data supports the null hypothesis.
- √ **The data supports the alternative hypothesis.**
- ☐ There's a 90% chance that the true correlation coefficient is between -0.15 and -0.55.
- √ **If we run this process of collecting data and running a hypothesis test 100 times, we would expect around 90 of the confidence intervals to contain the true correlation coefficient.**

(e) (2 points) Hannah fits a regression line to predict `Peak_Position` from `Streams` and also creates a residual plot, as shown below. Based on the residual plot, would linear regression be a good choice for making predictions from this data?



- ◯ Yes, linear regression is always a good approach, regardless of the data.
- ◯ Yes, linear regression is a good model here because there is a strong negative correlation between `Streams` and `Peak_Position`.
- ◯ Yes, linear regression is a good model here because the residual plot does not show an upward trend or a downward trend.
- ◯ Yes, linear regression is a good model because the residual plot does not show a linear trend.
- √ **No, linear regression is not a good model here because the residual plot shows a curved pattern.**
- ◯ No, linear regression is not a good model here because this was not a controlled experiment, and an association does not imply causation.

(f) (4 points) Fill in the code below to complete a function that creates the scatterplot above when called. The function takes in the following 3 arguments: `tbl`, a table containing two columns named `x_vals` (string) and `resids` (string) corresponding to the name of the x-values' column and the name of the residuals' column. You do not need to generate the title of the graph.

```
def residual_plot(tbl, x_vals, resids):
    ________(a)________
```

Solution: tbl.scatter(x_vals, resids)

(g) (2 points) Lucas suggests using a different type of technique, such as quadratic regression or polynomial regression, instead of a straight line to improve predictions. But Hannah believes that the regression line gives you the smallest possible RMSE, so there's no possible way to get a prediction with a lower average error. Is Hannah right or wrong? Select the correct statement below.

- ◯ Hannah is correct, as this is the definition of the regression line.
- √ **Hannah is wrong, as the regression line gives the smallest possible RMSE among all straight lines, but a different type of regression might give an even lower RMSE.**
- ◯ Hannah is correct, as the regression line is exactly the same as minimizing the RMSE.
- ◯ Hannah is wrong, as the regression line may give a different result depending on whether you minimize RMSE or use linear regression equations.

5. **Restaurant Reviews**

Amber, Noah, Haley, and Vivian have a shared database where they leave reviews of restaurants. The table is called `reviews` and contains the following columns:

- `reviewer`: string, the name of the reviewer
- `name`: string, the name of the restaurant
- `location`: string, containing the city and the state of the restaurant
- `rating`: integer, rating of the restaurant on a scale of 1 to 10
- `cuisine`: string, the style/method of cooking

(a) (3 points) Amber wants to create a table only containing restaurants with a rating of 10. Write a line of code to do this.

Solution: reviews.where('rating', are.equal_to(10))

(b) Fill in the code such that `top_reviewer` is equal to the person with the most reviews.

```
number_reviews_per_person = reviews.___(a)___(___(b)___)
sorting_reviews = number_reviews_per_person.sort(___(c)___)
top_reviewer = sorting_reviews.column(___(d)___).item(___(e)___)
```

i. (1 point) Fill in blank (a).

Solution: group

ii. (1 point) Fill in blank (b).

Solution: 'reviewer'

iii. (2 points) Fill in blank (c).

Solution: 'count', descending=True)

iv. (1 point) Fill in blank (d).

Solution: 'reviewer'

v. (1 point) Fill in blank (e).

Solution: 0

(c) Haley is interested in each reviewer's average rating by cuisine. Fill in the code so that `cuisine_by_reviewer` is a table where every cuisine gets its own row, and every reviewer gets its own column.

```
cuisine_by_reviewer = reviews.___(a)___(___(b)___, ___(c)___, ___(d)___, ___(e)___)
```

As a reminder, here are the columns in `reviews`:

- `reviewer`: string, the name of the reviewer
- `name`: string, the name of the restaurant
- `location`: string, containing the city and the state of the restaurant
- `rating`: integer, rating of the restaurant on a scale of 1 to 10
- `cuisine`: string, the style/method of cooking

i. (2 points) Fill in blank (a).

Solution: pivot

ii. (1 point) Fill in blank (b).

Solution: 'reviewer'

iii. (1 point) Fill in blank (c).

Solution: 'cuisine'

iv. (1 point) Fill in blank (d).

Solution: 'rating'

v. (1 point) Fill in blank (e).

Solution: np.mean

(d) (2 points) Amber has calculated the average rating for each Ethiopian restaurant in the dataset. She is interested in visualizing the distribution of these averages. Which of the following table functions and methods would help her do this? Select all that apply.

☐ .scatter
☐ .plot
☐ .barh
√ **.hist**
☐ None of the above

(e) Haley is suspicious that Amber's ratings are consistently higher than hers. She decides to run a hypothesis test to verify her suspicion. Fill in the code to perform one simulation of an A/B test. Assume that `calculate_test_stat` is a function that calculates the difference in means between Haley and Amber's ratings. `haley_amber` is a table containing only Haley and Amber's reviews and contains the same columns as `reviews`.

```
shuffled_labels = haley_amber.___(a)___(___(b)___).column(___(c)___)
shuffled_table = haley_amber.___(d)___("reviewer", ___(e)___)
shuffled_group_means = shuffled_table.___(f)___(___(g)___, np.average)
simulated_test_stat = calculate_test_stat(shuffled_group_means)
```

i. (1 point) Fill in blank (a).

Solution: sample

ii. (1 point) Fill in blank (b).

Solution: with_replacement = False

iii. (1 point) Fill in blank (c).

Solution: "reviewer"

iv. (1 point) Fill in blank (d).

Solution: with_column

v. (1 point) Fill in blank (e).

Solution: shuffled_labels

vi. (1 point) Fill in blank (f).

Solution: group

vii. (1 point) Fill in blank (g).

Solution: "reviewer"

viii. (2 points) Besides the difference in means, what are other valid test statistics Haley could have used? Select all that apply.

- ☐ Absolute difference in means between Haley and Amber's ratings
- √ **Difference in 50th percentile ratings between Haley and Amber's ratings**
- ☐ Absolute difference in 50th percentile ratings between Haley and Amber's ratings
- ☐ Total variation distance between Amber and Haley's rating distribution

6. **Assumptions**

(a) (0 points) If you felt any question required additional assumptions, please write them here. Be warned: We will only consider these assumptions if the question indeed required additional information.

Solution: wahoo!