Fall 2023 Final Exam

Foundations of Data Science

Name

Total Score: _____ of 100 Points

Instructions

- Make sure to rip off the Table Reference and Final Exam Reference Guide attached at the end of the exam.
- Select the correct response(s) or provide a written response depending on the question type. If a prompt asks you to write code, then you can provide your own code or use the provided template. Try to provide your responses in the spaces provided. If you find that you need additional space, write your extended response(s) on one of the provided blank sheets of paper and number them, so we can connect your response to the question.
- You can assume the following code has been run, when you are writing your responses for Section B:

```
from datascience import *
import numpy as np
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plt
plt.style.use('fivethirtyeight')
```

- The multiple choice questions (◯) and multiple answer questions (☐) will be scored like they are in Canvas Quizzes.
- The open response questions will be graded as:
  - Full Points: The response is correct and may contain a very very small error.
  - Partial Points: A reasonable response was provided. The partial point value will depend on your response.
  - No Points: No reasonable attempt was provided.
- Once you are finished, turn in your exam and you are welcome to leave.

Section A

1. (2 points) A real estate company has a dataset of all their buildings, with three attributes for each building: its size (in square feet), its type (residential or commercial), and its estimated value (sale price) if sold (in dollars). The standard visualization to understand the distribution of building types is: **Choose one.**

   ◯ A bar chart ◯ A line plot ◯ A scatter plot ◯ A histogram

2. Select True or False for each of the following:

   (a) (2 points) The height of each bar in a histogram represents the proportion of data within the corresponding bin. **Choose one.**

   ◯ True ◯ False

   (b) (2 points) For any distribution, the percentage of data that lies beyond two standard deviations on either side of the mean is at most 25%. **Choose one.**

   ◯ True ◯ False

   (c) (2 points) A classifier is considered to be overfitting if it performs very well on the test set. **Choose one.**

   ◯ True ◯ False

   (d) (2 points) If we use linear regression to predict y-values based on our x-values, the median of our residuals will always be zero. **Choose one.**

   ◯ True ◯ False

   (e) (2 points) For any two events A and B, the probability $P(A \text{ and } B)$ is less than or equal to the probability $P(A \text{ or } B)$ **Choose one.**

   ◯ True ◯ False

3. Cognitive Behavioral Therapy (CBT) is a psycho-social intervention that aims to reduce symptoms of mental health conditions such as depression. As a researcher, you are tasked with designing and implementing a large study of the effect of CBT on reducing such depression symptoms.

   As part of your study, you randomly sample 1,600 individuals seeking treatment for various levels of depression. Currently, you have a table called `patients` containing the following string data:: the first name (`'First Name'`), last name (`'Last Name'`), and email (`'Email'`) for each of the sampled individuals. For a visual reference, use the first 3 columns of the `patients` table in the Table Reference section as an example.

(a) (3 points) First, you need to randomly assign 800 individuals to receive a course of well-studied antidepressant medication and the other 800 individuals to receive a sequence of CBT sessions. Write code that will update the patients table by adding a fourth column called `'CBT'` to the table that will contain `bool` values where 800 of the individuals will randomly be assigned a value of `True` and the rest a value of `False`. Your updated table should resemble the first 4 columns of the `patients` table in the Table Reference section.

Hints: The `np.random.choice` function has a parameter called `replace` that allows you to use sampling with or without replacement. The default value is `True`. Also, the code:

```
np.array([True] * 3 + [False] * 2)
```

will create the following array:

```
array([ True,  True,  True, False, False], dtype=bool).
```

(b) (2 points) With the treatment assignments, the patients are messaged with their treatment. Those receiving CBT know they are receiving several weeks of therapy sessions, and the others know that they are receiving a course of antidepressant medication that is known to reduce the symptoms of depression at all levels. All 1,600 individuals consent to and complete their assigned treatment. This study design is a randomized controlled experiment. **Choose one.**

◯ True ◯ False

(c) PHQ-9 is a 9-question patient health questionnaire that, according to the American Psychological Association, offers psychologists concise information about a patient's level of depression. PHQ-9 scores range from 0 to 27, with higher scores indicating a more severe level of depression. PHQ-9 scores are collected for every individual in the study both before and after treatment. The `patients` table is updated with columns labeled 'Pre-PHQ-9'— for the scores before treatment, and `'Post-PHQ-9'` for the scores after treatment. See a preview of the table (just the first 6 columns after this step) in the Table Reference section.

i. (2 points) Which of the following is a correct null hypothesis that could be used to test if CBT reduces depression symptoms more effectively than the antidepressant treatment? **Choose one.**

- ◯ There is no difference between the average difference in PHQ-9 scores (Pre-PHQ-9 scores minus Post-PHQ-9 score) for those receiving CBT and the average difference in PHQ-9 scores for those receiving the antidepressants.
- ◯ There is a positive difference between the average difference in PHQ-9 scores (Pre-PHQ-9 scores minus Post-PHQ-9 score) for those receiving CBT and the average difference in PHQ-9 scores for those receiving the antidepressants.
- ◯ There is a negative difference between the average difference in PHQ-9 scores (Pre-PHQ-9 scores minus Post-PHQ-9 score) for those receiving CBT and the average difference in PHQ-9 scores for those receiving the antidepressants.

ii. (2 points) Please, state a clear and complete alternative hypothesis that could be used to test if CBT reduces depression symptoms more effectively than the antidepressant treatment.

iii. (3 points) Create a python function called `PHQ_diff` that has 2 arguments, `pre` and `post` where:

- `pre` is a Pre-PHQ-9 integer score.
- `post` is a Post-PHQ-9 integer score.

The function should return the Pre-PHQ-9 score minus the Post-PHQ-9 score.

iv. (3 points) Using the `PHQ_diff` function, update the `patients` table by adding a column called
`'Diff-PHQ-9'` to the `patients` table showing the difference in PHQ-9 scores for all the patients. The table will now resemble all 7 columns in the `patients table` in the Table Reference section.

v. (3 points) Using the updated `patients` table, provide code that will generate a histogram showing the distribution of `'Diff-PHQ-9'` values for those receiving CBT overlaid with a histogram showing the distribution of of `'Diff-PHQ-9'` values for those not receiving CBT. Refer to the entire `patients` table in the Table Reference section as an example.

vi. (2 points) If there is an association between the treatment and the difference in PHQ-9 scores, then the histograms will almost perfectly overlay each other. **Choose one.**

◯ True ◯ False

vii. (3 points) Using the `patients` data, complete the following code which will calculate the observed statistic. That is, the difference between the average difference in PHQ-9 score for those patients who received CBT, and the average difference in PHQ-9 score for those who received the antidepressant.

```
averages_tbl = (patients.select('CBT', 'Diff-PHQ-9')
                        .____________(A)___________)
averages = averages_tbl.column('Diff-PHQ-9 average')
observed_diff_aves = ____________(B)___________
```

viii. (3 points) Ten thousand permutations (reshuffling the `'CBT'` column) of the `patients` table were generated and the difference in average differences was calculated each time and stored in an array called `simulated_diffs_aves`. If the observed difference is named `observed_diff_aves`, write code that will calculate the p-value for this hypothesis test using the simulated data.

ix. (3 points) If the p-value is 4%, which of the following is a valid conclusion for this hypothesis test? **Choose all that apply.**

- ☐ 4% of the test statistics simulated under the null hypothesis were as, or less extreme than the observed test statistic.
- ☐ CBT reduces the symptoms of depression equally as well as the prescribed antidepressant.
- ☐ With a p-value cutoff of 5%, our data are consistent with the null hypothesis.
- ☐ There is a statistically significant reduction in depression symptoms for those that follow a sequence of CBT sessions
- ☐ None of the above.

Section B

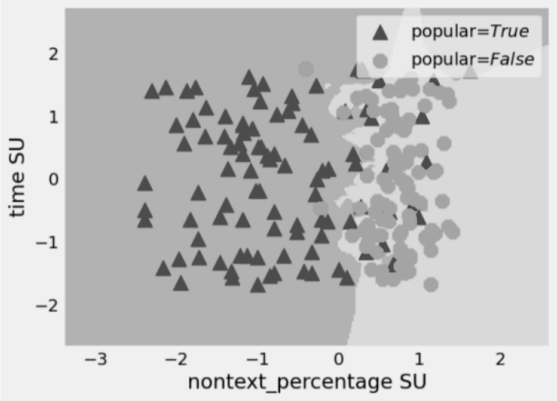
4. Meme generation is a high-profit business for the Instagram account @menime. Currently, they have 17 million followers and they post a variety of image-based or video-based memes. In an effort to create a prediction tool that will help determine if a meme will be popular before they release it, they share 200 memes with a randomly generated focus group of 1,000 Instagram users, who rate ('like') the memes. The data is stored in a table called `meme_data`. More details about this table are in the Table Reference section.

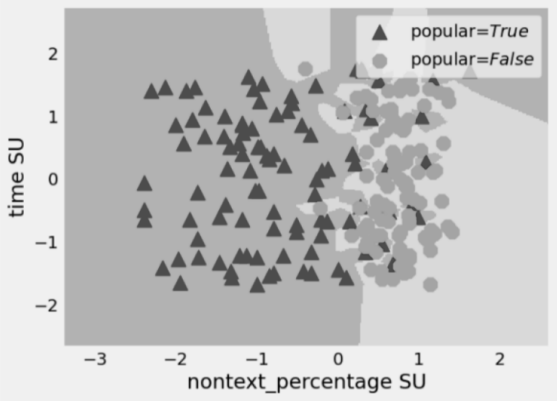
   You are interested in building a classification model that will use the numerical features in the `meme_data` table to predict whether or not a meme will be `'popular'`. A meme will be labeled `'popular'` if more than 70% of the focus group liked it.

   (a) (3 points) Complete the code below so that `meme_popular` is a copy of `meme_data` with an additional column named `'popular'`, which contain Boolean values that indicate whether a meme is popular (`True`) or not (`False`).

   ```
   pop_arr = ____________________(A)____________________
   meme_popular = ________________________(B)________________________
   ```

(b) (2 points) After converting the `time` and `nontext` percentage columns to standard units and creating two k-NN classifiers, each with a different value of `k`: `k=3` and `k=1`. Which of the following plots corresponds to the 1-NN classifier?





Choose one.

◯ The left plot ◯ The right plot

(c) (2 points) After training a 1-NN classifier, you notice that only 10% of the memes in the training data are popular, compared to 50% of memes in the testing data. Additionally, you find that this imbalance is due to an error in your code.

After correcting the error and re-distributing the data to restore the balance of popular memes, you re-train a 1-NN classifier. How would you expect the accuracy to change after re-balancing the data? **Choose one.**

☐ The accuracy increases.

☐ The accuracy remains the same.

☐ The accuracy decrease.

☐ It is not possible to determine what will happen to the accuracy.

(d) (3 points) It turns out that someone has already developed a similar classification tool called memeNN that has an accuracy of 75%. You notice that when memeNN makes a correct prediction, your 1-NN classifier's correctly predicts the popularity 82% of the time. When memeNN makes an incorrect prediction, your 1-NN classifier correctly predicts the popularity 45% of the time.

If a meme is randomly sampled from the test set and the 1-NN classifier predicts its class correctly, what is the probability that memeNN will also predict the class correctly?

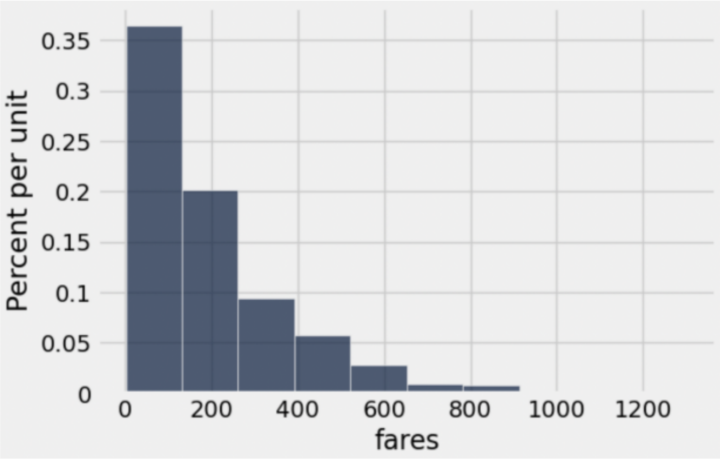
Write your answer as a mathematical/python expression.

Hint: Making a decision tree may be helpful.

Section C

5. Your data analyst team is interested in studying Bay Area public transportation, so you begin analyzing data from the widely-used BART train system and the AC Transit bus services for the year 2022. Unfortunately, due to budget cuts, your available compute power is unable to process all of the data from 2022, so instead, your team is going to work with a large random sample of 1,000 riders. That sample is loaded into a table called `transport`. The first few rows of that table are previewed in the Table Reference section.

   (a) (3 points) Given below is the distribution of the fares column from the `transport` table. Which of the following conclusions can you draw from the plot? **Select all that apply.**



   ☐ The distribution of the `'fares'` column in `transport` is right-skewed.

   ☐ The distribution of the `'fares'` column in `transport` is left-skewed.

   ☐ The median of the `'fares'` column in `transport` is less than the mean.

   ☐ The median of the `'fares'` column in `transport` is greater than the mean.

   (b) (3 points) Which of the following statements must be true? **Select all that apply.**

   ☐ The distribution of fare spending among all riders is approximately normal.

   ☐ The distribution of sample means of fare spending is approximately normal for large random samples of data.

   ☐ The distribution of sample sums of fare spending is approximately normal for large random samples of data.

   ☐ The distribution of sample medians of fare spending is approximately normal for large random samples of data.

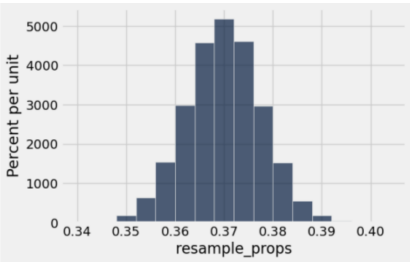
   ☐ None of the above.

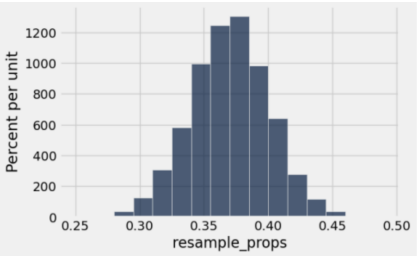
(c) Your team is interested in estimating the proportion of all riders who had transferred between a BART train and an AC Transit bus at least once. You decide to use your sample of 1,000 riders to estimate this unknown population parameter.

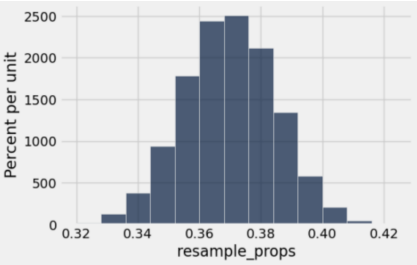
i. (4 points) Provide code that will generate a table of 10,000 bootstrapped proportions of riders who transferred between a BART train and an AC Transit bus at least once.

```
resample_props = make_array()

for i in np.arange(10_000):
    resamp = _________________________(A)___________________
    resamp_prop = ___________________(B)___________________
    ___________________________(C)___________________________

resamp_props_tbl =  Table().with_column("resample_props", resample_props)
```

ii. (2 points) You find that the mean and standard deviation of your bootstrapped proportions, resample props is 0.37 and 0.015, respectively. Which of the following most closely resembles the distribution of resample props? **Choose one.**







◯ The left graph ◯ The middle graph ◯ The right graph

iii. (3 points) Write a mathematical or Python expression that evaluates to the probability that the first row in transport is included at least once in a single bootstrap re-sample of size 1,000.

iv. (3 points) Provide code that creates the array `interval` which contains the left and right endpoints of a 95% confidence interval estimate for the proportion of riders in the population who transferred at least once.

Note: You may used variable names defined from previous sub-parts in your code.

```
left = ____________________(A)____________________
right = ____________________(B)____________________
interval = make_array(left, right)
```

v. (3 points) Which of the following conclusions can you draw using the 95% confidence interval generated in the previous part (iv)? **Select all that apply.**

- ☐ If someone takes the BART train, there is a 95% chance that they transfer to an AC Transit bus.
- ☐ If you make confidence intervals from many large random samples from the population, you can expect that roughly 95% of the intervals you create will contain the true population proportion.
- ☐ There is a 95% chance that the population's true transfer proportion is within the interval generated in part (iv).
- ☐ There is a 95% chance that the sample's true transfer proportion is within the interval generated in part (iv).
- ☐ None of the above.

vi. (3 points) Your team has one last request. They want your 95% confidence interval to be no wider than 5%. What is the smallest sample size that satisfies this requirement? **Express your answer as a whole number or a mathematical/Python expression**.

Section D

6. Farhana likes to attend a popular class at the local fitness center. She collects data about each class she attends in a table called `workouts`. The table is previewed in the Table Reference section.

   (a) (2 points) Choose the best technique to answer the question "How high will Farhana's exercise heart rate be today given that 30 people attended class?" **Choose one.**

   ◯ Classification
   ◯ Linear Regression
   ◯ Hypothesis Test
   ◯ Randomized Control Experiment
   ◯ Bayes' Rule

   (b) (2 points) Choose the best technique to answer the question "Is there a difference in Farhana's heart rate between sunny and rainy days?" **Choose one.**

   ◯ Classification
   ◯ Linear Regression
   ◯ Hypothesis Test
   ◯ Randomized Control Experiment
   ◯ Bayes' Rule

   (c) (2 points) Choose the best technique to answer the question "What are the most likely weather conditions given that 12 people attended class today?" **Choose one.**

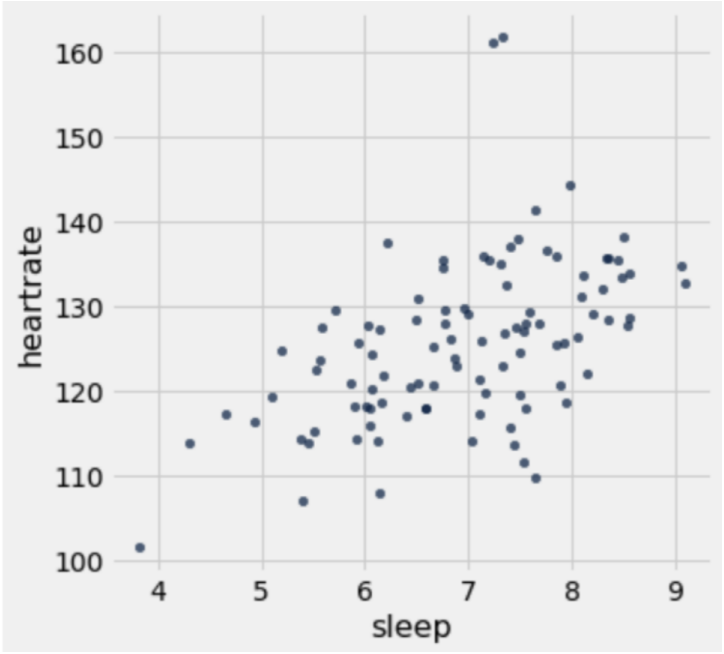
   ◯ Classification
   ◯ Linear Regression
   ◯ Hypothesis Test
   ◯ Randomized Control Experiment
   ◯ Bayes' Rule

   (d) (2 points) Choose the best test statistic for the following alternative hypothesis. **Choose one.**

   Alternative Hypothesis: "The class size is larger on sunny days than it is on rainy days."

   ◯ The total variation distance between the class size distribution of sunny days and class size distribution of rainy days
   ◯ The empirical mean of class size on sunny days
   ◯ The empirical mean of class size
   ◯ The difference of mean class size between sunny and rainy days
   ◯ The difference of mean class size between sunny and cloudy days

(e) (3 points) Farhana wants to see if there is a relationship between how much sleep she gets and her heart rate during class, so she creates the following scatter plot.



Write a line of code that would generate the scatter plot above.

(f) (3 points) Which of the following are valid conclusions from this graph? **Choose all that apply.**

- ☐ There is a positive association between her sleep and her heart rate during class
- ☐ There is a negative association between her sleep and her heart rate during class
- ☐ Getting more sleep causes Farhana to have a higher heart rate during class
- ☐ Getting less sleep causes Farhana to have a higher heart rate during class
- ☐ Fewer people attend class on rainy days

(g) (3 points) Farhana asks her friend to compute the correlation coefficient `r` between these two variables , but her friend's code has at least one mistake in it. In the code below, circle and cross out each mistake and, if applicable, write the correct code immediately above. Alternatively, you can re-write the code in the solution box. You can use the following `standard_units` function from lecture:

```
def standard_units(any_numbers):
    '''Convert any array of numbers to standard units.'''
    return (any_numbers - np.mean(any_numbers))/np.std(any_numbers)
```

Here is her friend's code:

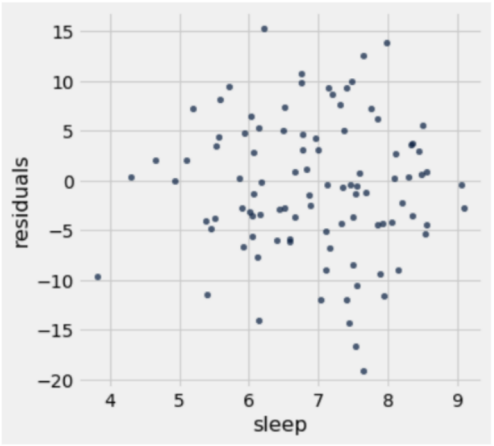
```
heartrate_in_su = standard_units('heartrate')

sleep_in_su = standard_units('sleep')

r = np.sum(heartrate_in_su * sleep_in_su)
```

(h) (3 points) Suppose we know the following:

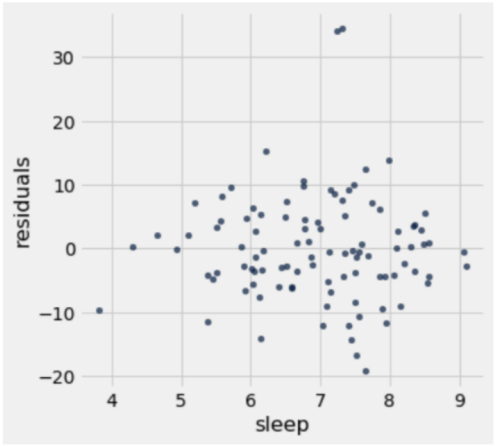
- Farhana's heart rate has an average of 125 bpm and a standard deviation of 6 bpm
- Farhana's sleep has an average of 7 hours and a standard deviation of 1 hour
- The correlation between Farhana's heart rate and sleep is 0.5

If we were to fit a regression line to the scatterplot in (e), what would the predicted heartrate be when Farhana gets 8 hours of sleep? You may leave your answer as a mathematical/Python expression.

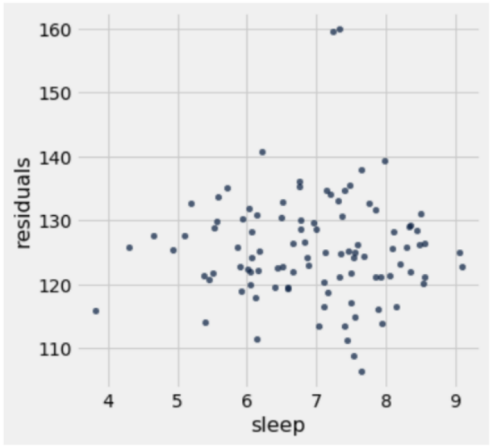
(i) (2 points) Which of the following is the residual plot for the scatter plot shown in part (e)?



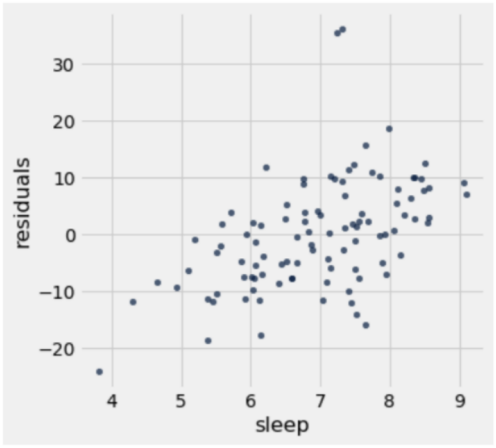
A



B



C



D

Choose one.

◯ Option A ◯ Option B ◯ Option C ◯ Option D

(j) (3 points) Farhana begins a research apprenticeship in the School of Public Health, and wants to understand whether the amount of sleep someone gets causes a change in average heart rate during exercise. Her lab starts a study with a random sample of Berkeley undergraduate and graduate students who exercise regularly. Which of the following experiments would be able to answer her causal question? **Choose all that apply.**

☐ Ask the undergraduates to sleep 7 hours per night and the graduate students to sleep 9 hours per night. Then, collect heart rate data during exercise.

☐ Randomly assign the subjects to two groups. Have the first group exercise for 1 hour per day, and have the second (control) group not exercise at all. Then, measure how much sleep they get before and after each exercise session.

☐ Randomly assign the subjects to two groups. Have the first group sleep for 7 hours or less per night, and the second group sleep 9 hours or more per night. Then, collect heart rate data during exercise.

☐ It is impossible to determine a causal link between these two variables.

This page was intentionally left blank.

Table Reference

patients

Here is a preview of the `patients` table:

First Name	Last Name	Email	CBT	Pre-PHQ-9	Post-PHQ-9	Diff-PHQ-9
Lily	Smith	lily.smith@email.com	False	23	20	3
Ethan	Garcia	ethan.garcia@email.com	True	18	16	2
Sophia	Lee	sophia.lee@email.com	False	20	19	1
Zoe	Nguyen	zoe.nguyen@email.com	True	15	16	-1
Lily	Smith	lily.smith@email.com	False	23	20	3

... (1595 rows omitted)

- When starting Question 3, the table initially resembles the first 3 columns of this table.
- After successfully completing Question 3 (a), the table resembles the first 4 columns of this table.
- In Question 3 (c), the table resembles the first 6 columns of this table.
- After successfully completing Question 3 (c) iv, the table resembles the first 7 columns of `patients`.

meme_data

Each row of the table `meme_data` represents a meme. The columns are:

- `category` (string): the category of the meme, which is either an "image" or a "video".
- `time` (integer): the duration of the meme, in seconds. Images will have a time value of 0.
- `nontext_percentage` (float): the percentage of the meme that is non-textual content (scale: [0.0 - 100.0]).
- `rating` (float): the percentage of the focus group who liked the meme (scale: [0.0 - 100.0]).

A preview of this table is not provided and should not be needed.

transport

Here is a preview of the `transport` table and some information about the data values:

- `id` (**integer**): identification (id) of the rider.
- `transfer` (**boolean**): whether that particular rider transferred between a BART train and an AC Transit bus at least once during 2022.
- `fares` (**float**): total amount that particular rider spent on fares in 2022, measured in dollars.

id	transfer	fares
32849	True	12.5
29490	False	62
81305	False	131.75
70654	False	43

... (996 rows omitted)

workouts

Here is a preview of the `workouts` table and some information about the data values:

size	heartrate	weather	sleep
33	145.1	sunny	7.3
28	100.7	sunny	6.5
23	124	sunny	5
10	137.8	rainy	9

The table contains four columns:

- **size**: an int, the number of people who attended the class
- **heartrate**: a float, her average heart rate during the class in beats per minute (bpm)
- **weather**: a string, the weather conditions for that day
- **sleep**: a float, the number of hours of sleep she got the night before

MATH 108 Final Reference Guide — Page 1

Statements

`more_than_1` = `2 + 3`

Name — Any expression

- Statements don't have a value; they perform an action
- An assignment statement changes the meaning of the name to the left of the = symbol
- The name is bound to a value (not an equation).

Comparisons

- **<** and **>** mean what you expect (less than, greater than)
- **<=** means "less than or equal"; likewise for **>=**
- **==** means "equal"; **!=** means "not equal"
- Comparing strings compares their alphabetical order

Arrays - sequences of the same type that can be manipulated

- Arithmetic and comparisons are applied to each element of an array individually
  - `make_array(1,2,3) ** 2 # array([1, 4, 9])`
- Elementwise operations can be done on arrays of the same size
  - `make_array(3,2) * make_array(5,4) # array([15,8])`

Defining a Function

```
def function_name(arg1, arg2, ...):
    # Body can contain anything inside of it
    return # a value (the output of the function call)
```

Defining a Function with no arguments

```
def function_name():
    # Body can contain anything inside of it
    return # a value (the output of the function call)
```

- Functions with no arguments can be called by `function_name()`

For Statements

```
total = 0
for i in np.arange(12):
    total = total + i
```

- The body is executed **for** every item in a sequence
- The body of the statement can have multiple lines
- The body should do something: assign, sample, print, etc.

Conditional Statements

```
if <if expression>:
    <if body>
elif <elif expression 0>:
    <elif body 0>
elif <elif expression 1>:
    <elif body 1>
...
else:
    <else body>
```

Total Variation Distance

Total variation distance is a statistic that represents the difference between two distributions

```
TVD = 0.5*(np.sum(np.abs(dist1-dist2)))
```

Operations: addition `2+3=5;` subtraction `4-2=2;` division `9/2=4.5` multiplication `2*3=6;` division remainder `11%3=2;` exponentiation `2**3=8`

Data Types: string "`hello`"; **boolean** `True, False;` **int** `1, -5;` **float** `- 2.3, -52.52, 7.9, 8.0`

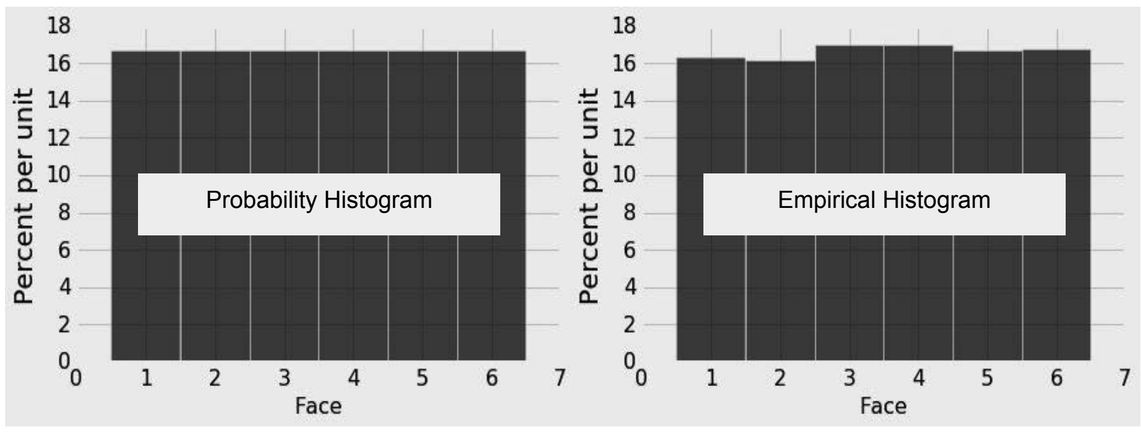
`Table.where` predicates: Any of these predicates can be negated by adding "`not_`" in front of them, e.g. `are.not_equal_to(x)`

- `are.equal_to(x) # val == x`
- `are.above(x) # val > x`
- `are.above_or_equal_to(x) # val >= x`
- `are.below(x) # val < x`
- `are.between(x, y) # x <= val < y`
- `are.containing(s) # contains the string s`

A **histogram** has a few defining properties:

- The bins are continuous (though some might be empty) and are drawn to scale
- The **area** of each bar is equal to the percent of entries in the bin
- The total area is 100%

- The histogram on the left represents the theoretical probabilities in the distribution of the face that appears on one roll of a fair die
- The histogram on the right represents the observed distribution of the faces after rolling the die many times
- If we keep rolling, the right hand histogram is likely to look more like the one on the left



Calculating Probabilities

Complement Rule: P(event does not happen) = 1 - P(event happens)

Multiplication Rule: P(two events both happen) =
P(one happens) * P(the other happens, given that the first happened)

Addition Rule: If an event can happen in ONLY one of two ways:
P(event happens) =
P(first way it can happen) + P(second way it can happen)

Bayes' Rule: P(event A happened given event B happened) =
P(both event A and event B happened) / P(event B happened)

For Bayes' rule, if the probabilities are displayed on a tree diagram, the denominator is the chance of the branches in which B happens, and the numerator is the chance of the branches in which both A and B happen.

Simulating a Statistic:

- Create an empty array in which to collect the simulated values
- For each repetition of the process
  - Simulate one value of the statistic
  - Append this value to the collection array
- At the end, all simulated values will be in the collection array

MATH 108 Final Reference Guide — Page 2

In the examples in the left column, `np` refers to the NumPy module, as usual. Everything else is a function, a method, an example of an argument to a function or method, or an example of an object we might call the method on. For example, `tbl` refers to a table, `array` refers to an array, and `num` refers to a number. `array.item(0)` is an example call for the method `item`, and in that example, `array` is the name previously given to some array.

`max(array); min(array)`	Maximum or minimum of an array
`sum(array)`	Sum of all elements in an array; The sum of an array of boolean values is the number of values that are True
`len(array)`	Length (num elements) in an array
`round(num); np.round(array)`	The nearest integer to a single number or each number in an array
`abs(num); np.abs(array)`	The absolute value of a single number or each number in an array
`np.average(array), np.mean(array)`	The average of the values in an array

`np.arange(start, stop, step)`
`np.arange(start, stop)`
`np.arange(stop)`
	An array of numbers starting with `start`, going up in increments of `step`, and going up to but excluding `stop`. When `start` and/or `step` are left out, default values are used in their places. Default `step` is 1; default `start` is 0.

`array.item(index)`	The item in the array at some index. `array.item(0)` is the first item of `array`.
`np.append(array, item)`	A copy of the array with item appended to the end. If item is another array, all of its elements are appended.
`np.exp(array)`	Calculate the exponential fo all the elements in the array.
`np.random.choice(array)` `np.random.choice(array, n)`	An item selected at random from an array. If `n` is specified, an array of `n` items selected at random with replacement is returned. Default `n` is 1.
`np.ones(n)`	An array of length `n` which consists of all ones.
`np.diff(array)`	An array of length `len(array)-1` which contains the difference between adjacent elements.
`np.count_nonzero(array)`	An integer corresponding to the number of non-zero (or `True`) elements in an array.
`sample_proportions(sample_size, model_proportions)`	An array of proportions that add up to 1. The result of sampling `sample_size` elements from a distribution specified by `model_proportions`, and keeping track of the proportion of each element sampled.
`Table()`	An empty table.
`Table.read_table(filename)`	A table with data from a file.
`tbl.num_rows`	The number of rows in a table.
`tbl.num_columns`	The number of columns in a table.
`tbl.labels`	A list of the column labels of a table.
`tbl.with_column(name, values)` `tbl.with_columns(n1, v1, n2, v2...)`	A table with an additional or replaced column or columns. `name` is a string for the name of a column, `values` is an array.
`tbl.column(column_name_or_index)`	An array containing the values of a column
`tbl.select(col1, col2, ...)`	A table with only the selected columns. (Each argument is the label of a column, or a column index.)
`tbl.drop(col1, col2, ...)`	A table without the dropped columns. (Each argument is the label of a column, or a column index.)
`tbl.relabeled(old_label, new_label)`	A new table with a label changed.
`tbl.take(row_index)` `tbl.take(row_indices)`	A table with only the row(s) at the given index or multiple indices. `row_indices` must be an array of indices.
`tbl.exclude(row_index)` `tbl.exclude(row_indices)`	A table without the row(s) at the index or multiple indices. `row_indices` must be an array of indices.
`tbl.sort(column_name_or_index)`	A table of rows sorted according to the values in a column (specified by name/index). Default order is ascending. For descending order, use argument `descending=True`. For unique values, use `distinct=True`.
`tbl.where(column, predicate)`	A table of the rows for which the column satisfies some predicate. See "`Table.where predicates`" on Page 1.
`tbl.apply(function, column_or_columns)`	An array of results when a `function` is applied to each item in a `column`.
`tbl.group(column_or_columns)`	A table with the counts of rows grouped by unique values or combinations of values in a column or columns.
`tbl.group(column_or_columns, func)`	A table that groups rows by unique values or combinations of values in a column or columns. The other values are aggregated by `func`. All column names (except the one(s) we group by) will now be `original_name func`. If a column is named 'price', and we group using the `min` function, our new column name will be 'price min'.
`tblA.join(colA, tblB, colB)` `tblA.join(colA, tblB)`	A table with the columns of `tblA` and `tblB`, containing rows for all values of a column that appear in both tables. Default value of `colB` is `colA`. `colA` is a string specifying a column name, as is `colB`.
`tbl.pivot(col1, col2)` `tbl.pivot(col1, col2, vals, collect)`	A pivot table where each unique value in `col1` has its own column and each unique value in `col2` has its own row. The cells of the grid contain row counts (two arguments) or the values from a third column, aggregated by the collect function (four arguments) .
`tbl.sample(n)` `tbl.sample(n, with_replacement)`	A new table where `n` rows are randomly sampled from the original table. Default is with replacement. For sampling without replacement, use argument `with_replacement=False`. If sample size `n` is not specified, the default is the number of rows in the original table.
`tbl.scatter(x_column, y_column)`	Draws a scatter plot consisting of one point for each row of the table.
`tbl.barh(categories)` `tbl.barh(categories, values)`	Displays a bar chart with bars for each category in a column, with length proportional to the corresponding frequency. If `values` is not specified, overlaid bar charts of all the remaining columns are drawn.
`tbl.bin(column, bins)`	A table of how many values in a column fall into each bin. Bins include lower bounds & exclude upper bounds.
`tbl.hist(column, unit, bins, group)`	Displays a histogram of the values in a column. `unit` and `bins` are optional arguments, used to label the axes and group the values into intervals (bins), respectively. Bins include lower bounds & exclude upper bounds. If `group` is specified, the rows are grouped by the values in the column, and histograms for all the groups are overlaid.

- **P-Value**: The chance, **under the null hypothesis**, that the test statistic comes out equal to the one in the sample, or more in the direction of the alternative:
  - If the p-value is small and the null is true, something very unlikely has happened.
  - Conclude that the data support the alternative hypothesis more than they support the null.

- Even if the null is true, your random sample might indicate the alternative, just by chance
- The **cutoff** for P is the chance that your test makes the wrong conclusion when the null hypothesis is true
- Using a small cutoff limits the probability of this kind of error

---

A/B test for comparing two samples

- **Example**: Among babies born at some hospital, is there an association between birth weight and whether the mother smokes?
- **Null hypothesis**: The distribution of birth weights is the same for babies with smoking mothers and non-smoking mothers.
- **Inferential Idea:** If maternal smoking and birth weight were not associated, then we could simulate new samples by replacing each baby's birth weight by a randomly picked value from among all the birth weights.
- **Simulating the test statistic under the null:**
  - Permute (shuffle) the outcome column many times. Each time:
    - Create a shuffled table that pairs each individual with a random outcome.
    - Compute a sampled test statistic that compares the two groups, such as the difference in mean birth weights.

The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set

For `s = [1, 7, 3, 9, 5]`, `percentile(80, s)` is 7

The 80th percentile is ordered element 4: `(80/100) * 5`

For a percentile that does not exactly correspond to an element, take the next greater element instead

`percentile(10, s)` is 1 `percentile(20, s)` is 1
`percentile(21, s)` is 3 `percentile(40, s)` is 3

- `minimize` must take in a function whose arguments are numerical, and returns an array of those numerical arguments
- If the function `rmse(a, b)` returns the root mean squared error of estimation using the line "estimate = a*x* + b",
  - then `minimize(rmse)` returns array [a_0, b_0]
  - a_0 is the slope and b_0 the intercept of the line that minimizes the rmse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

Population (fixed) → Sample (random) → Statistic (random)

A 95% **Confidence Interval** is an interval constructed so that it will contain the true population parameter for approximately 95% of samples

For a particular sample, the generated interval either contains the true parameter or it doesn't; the process works 95% of the time

Bootstrap: When we wish we could sample again from the population, instead sample from the *original large random sample the same number of times as there are data-points in the sample*

Using a confidence interval to test a hypothesis about a numerical parameter:

- Null hypothesis: **Population parameter = *x***
- Alternative hypothesis: **Population parameter ≠ *x***
- Cutoff for P-value: *p*%
- Method:
  - Construct a (100-*p*)% confidence interval for the population parameter
  - If *x* is not in the interval, reject the null
  - If *x* is in the interval, fail to reject the null

The Central Limit Theorem (CLT)

If the sample is large, and drawn at random with replacement,

Then, *regardless of the distribution of the population,*

the probability distribution of the sample average (or sample sum) is roughly bell-shaped

- Fix a large sample size
- Draw all possible random samples of that size
- Compute the mean of each sample
- You'll end up with a lot of means
- The distribution of those is the *probability distribution of the sample mean*
- It's roughly normal, centered at the population mean
- The SD of this distribution is the (population SD) /√sample size

---

Choosing sample size so that the 95% confidence interval is small

- CLT says the distribution of a sample proportion is roughly normal, centered at the true population proportion
- **95% confidence interval:**
  - Sample proportion ± 2 SDs of the sample proportion
- **CI Width** = 4 SDs of the sample proportion

  = 4 x (SD of 0/1 population)/√sample size
- The SD of a 0/1 population is less than or equal to 0.5

Expression	Description
`percentile(n, arr)`	Returns the n-th percentile of array `arr`
`np.std(arr)`	Return the standard deviation of an array `arr` of numbers
`minimize(fn)`	Return an array of arguments that minimize the function `fn`
`tbl1.append(tbl2)` `tbl1.append(row)`	Append a row or all rows of `tbl2`, mutating `tbl1`. Appended object and `tbl1` must have identical columns.
`table.rows`	All rows of a table; used in `for row in table.rows:`
`table.row(i)`	Return the row of a table at index *i*
`row.item(j)`	Returns item `j` from some row

Mean (or average: Balance point of the histogram

Standard deviation (SD) =

root	mean	square of	deviations from	average
5	4	3	2	1

Measures roughly how far off the values are from average

Most values are within the range "average ± *z* SDs"

- *z* measures "how many SDs above average"
- If *z* is negative, the value is below average
- *z* is a value in **standard units**
- Chebyshev: At most $1/z^2$ are z or more SDs from the mean
- Almost all standard unit values are in the range (-5, 5)
- Convert a value to standard units: (value - average) / SD
- *z* * SD + average is the original value

Percent in Range	All Distributions	Normal Distribution
average ± 1 SD	at least 0%	about 68%
average ± 2 SDs	at least 75%	about 95%
average ± 3 SDs	at least 88.888...%	about 99.73%

Correlation Coefficient (*r*) =

average of	product of	x in standard units	and	y in standard units

Measures how clustered the scatter is around a straight line

- $-1 \le r \le 1$; r = 1 (or -1) if the scatter is a perfect straight line
- *r* is a pure number, with no units

Regression for y and x in standard units: $y_{predicted,su} = r * x_{su}$

The regression line minimizes the root mean square error among all lines used to predict y from x.
The slope and intercept found by linear regression are unique.
Fitted value: height of the regression line at some x: a*x + b
Residual: difference between y and regression line height at x

$$y_{predicted} = slope * x + intercept$$

$$\text{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept of the regression line} = \text{mean of } y - \text{slope} * \text{mean of } x$$

Properties of fitted values, residuals, and the correlation r:

- mean of fitted values = mean of y
- SD of fitted values = |r| * (SD of y)
- mean of residuals = 0
- SD of residuals = $\sqrt{1 - r^2}$ * (SD of y)

The following functions were defined in lecture, but will **not** be available for use during the final exam. If you would like to use one of the functions, you must define it yourself.

- `standard_units`
- `correlation`
- `slope`
- `intercept`
- `fitted_values`
- `residuals`
- `prediction_at`

- **Regression Model**: y is a linear function of x + normal "noise"
- The errors are randomly sampled from a normal distribution that has mean 0
- Under this model, residual plot looks like a formless cloud

---

Prediction Intervals (assuming the regression model)

- Creating an interval of predictions of the true value of *y* based on a specified value of *x*
- Steps for creating an approximate 95% prediction interval:
  - Bootstrap your original sample
  - Calculate the slope and intercept of the regression line based on the new sample
  - Calculate slope * x + intercept, for the given *x*
  - Repeat the above steps many times and keep track of all of your fitted values
  - Create the prediction interval by taking the middle 95% of all the fitted values

---

Distance between two points

- Two numerical attributes *x* and *y*: $D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}$.
- Three numerical attributes *x, y,* and *z:*

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$$

***k*-Nearest Neighbors Classifier**

Choose *k* to be odd. To find the *k* nearest neighbors of an example:

- Find the distance between the example and each example in the training set
- Augment the training data table with a column containing all the distances
- Sort the augmented table in increasing order of the distances
- Take the top *k* rows of the sorted table

To classify an example into one of two classes:

- Find its *k* nearest neighbors
- Take a majority vote of the *k* nearest neighbors to see which of the two classes appears more often
- Assign the example the class that wins the majority vote

Accuracy of a classifier: The proportion of examples in the data set that are classified correctly

---