DATA 100 Spring 2021 Midterm-Exam

INSTRUCTIONS

This is your exam. Complete it either at exam.cs61a.org or, if that doesn't work, by emailing course staff with your solutions before the exam deadline.

This exam is intended for the student with email address `<EMAILADDRESS>`. If this is not your email address, notify course staff immediately, as each exam is different. Do not distribute this exam PDF even after the exam ends, as some students may be taking the exam in a different time zone.

For questions with **circular bubbles**, you should select exactly *one* choice.

- ◯ You must choose either this option
- ◯ Or this one, but not both!

For questions with **square checkboxes**, you may select *multiple* choices.

- ☐ You could select this choice.
- ☐ You could select this one too!

You may start your exam now. Your exam is due at <DEADLINE> Pacific Time. Go to the next page to begin.

Preliminaries

You can complete and submit these questions before the exam starts.

(a) What is your full name?

(b) What is your Berkeley email?

(c) What is your student ID number?

(d) When are you taking this exam?

- ◯ Tuesday 7pm PST
- ◯ Wednesday 8am PST
- ◯ Other

(e) Honor Code: *All work on this exam is my own.*

By writing your full name below, you are agreeing to this code:

(f) Important: You must copy the following statement exactly into the box below. Failure to do so may result in points deducted on the exam.

"I certify that all work on this exam is my own. I acknowledge that collaboration of any kind is forbidden, and that I will face severe penalties if I am caught, including at minimum, harsh penalties to my grade and a letter sent to the Center for Student Conduct."

1. (16.0 points)

At the beginning of the semester, we conducted a survey to learn about the students who are taking part in Data 100 this spring. The survey was posted to Piazza and emailed to all students enrolled and on the waitlist. We learned that 400 of the 1200 survey respondents will be living in Berkeley. Assume that first full discussion section of 40 students was formed by taking a simple random sample (without replacement) of these 1200 respondents.

(a) (2.0 pt) What is the expected number of students in the discussion who will be living in Berkeley?

Note: Round your answer to the **nearest integer**.

13

(b) (3.0 pt) Is it reasonable to approximate the number of students in the discussion that will be in Berkeley as a binomial distribution? Why or why not?

- ○ **Yes**, because each student has an equal chance of being chosen from the class for the section.
- ○ **No**, because binomial distributions are defined based on sampling with replacement, so it will not approximate sampling without replacement well.
- ● **Yes**, because there are 1200 students in total, while the sample size is only 40, so the chance a student lives in Berkeley does not change much from one draw to the next.
- ○ **No**, because every time a student who lives in Berkeley is selected into the sample, the chance of having another one who also lives in Berkeley decreases.

(c) (3.0 pt) Assume for now that it is reasonable to use a binomial approximation for the previous part. What is approximately the probability that exactly 3/4 of the students in the discussion live in Berkeley?

- ● $\binom{40}{30}\left(\frac{1}{3}\right)^{30}\left(\frac{2}{3}\right)^{10}$
- ○ $1-\binom{40}{10}\left(\frac{1}{3}\right)^{10}\left(\frac{2}{3}\right)^{30}$
- ○ $\sum_{k=30}^{40}\binom{40}{k}\left(\frac{1}{3}\right)^{k}\left(\frac{2}{3}\right)^{40-k}$
- ○ $1-\sum_{k=30}^{40}\binom{40}{k}\left(\frac{1}{3}\right)^{k}\left(\frac{2}{3}\right)^{40-k}$

(d) (2.0 pt) While we observed from the survey results that 400 respondents live in Berkeley, hypothetically, if we observe _____ of 1200 survey respondents live in Berkeley, this would have maximized the probability that 3/4 of the discussion section lives in Berkeley. **Note**: Round your answer to the **nearest integer**.

900

(e) (6.0 points)

After the 4th week, to adjust discussion sections, course staff drew names not from the survey respondents, but from the list of enrolled students. As a student in this 2nd discussion section, you learn that of the 40 students, 8 of them live in Berkeley. At this point, you haven't seen the full survey results shown at the beginning of this question and know only the answers of the 40 students in your section.

You wish to use this data to estimate the proportion of the original 1200 full survey respondents who are living in Berkeley, θ. Recall that in the process of estimation, one source of error is the statistical bias of $\hat{\theta}$: $E(\hat{\theta}) - \theta$.

i. (3.0 pt) One way for statistical bias to occur is when your sample of data is not representative of your population. Which of the following features would **most likely** lead to that effect?

- ◯ The sample size for the student's discussion section is only 40. If you extend the estimation to a class that is 30 times as large, the bias in your estimation will become exponentially greater.
- ◯ There is no guarantee on whether a respondent would answer honestly about whether they live in Berkeley for the survey. This margin of error becomes 40 times as large for a discussion section.
- ● The sampling frame for the student's discussion section is all enrolled students which encompasses a different population than the population that responded to the survey.
- ◯ A respondent may have the option to submit their answers multiple times, which adds bias by increasing the observed proportion of respondents who have the same answers as they do.

ii. (3.0 pt) The other way that statistical bias can occur is for your estimator itself to be biased.

Let X_i be a Bernoulli random variable that is equal to 1 if a student from your discussion lives in Berkeley and 0 otherwise.

Use the following as your estimator for θ:

$$\hat{\theta} = \frac{1}{40-1}\sum_{i=1}^{40} X_i$$

Calculate the bias of this estimator.

Note: Round your answer to the **nearest 2 decimal places**.

-0.13

2. (22.0 points)

Agnibho and Rahul would like to analyze some statistics from all the current college first-year basketball players. The `college_ball` table contains the following basic information about each college's first-year players.

- `name` (string): the name of the player.
- `hometown` (string): the city and state of the high school that the player played at.
- `college` (string): the name of the college that player attends.
- `ppg` (double): total points score divided by total games played.

The first 5 rows of `college_ball` are provided below:

	name	hometown	college	ppg
0	Lamelo Ball	Chino Hills, CA	University of California, Los Angeles	15.6
1	James Wiseman	Nashille, TN	University of Memphis	18.3
2	Tyrese Haliburton	Oshkosh, WI	University of Memphis	25.5
3	Anthony Edwards	Atlanta, GA	University of Georgia	10.2
4	Tyler Herro	Bentonville, AR	University of Georgia	18.5

(a) (2.0 points)

i. (1.0 pt) What type of data is `hometown`?

- (●) Qualitative nominal
- ( ) Quantitative discrete
- ( ) Quantitative continuous
- ( ) Qualitative ordinal

ii. (1.0 pt) What type of data is `ppg`?

- ( ) Qualitative nominal
- ( ) Quantitative discrete
- (●) Quantitative continuous
- ( ) Qualitative ordinal

(b) (4.0 pt) You want to first make sure that `name` is a primary key in this table. Consider a function `is_primary_key` that takes in a string `column_name` and a DataFrame `df`; it returns `True` if `column_name` is a primary key of `df`, `False` otherwise. Select all of the following correct implementations of `is_primary_key`.

Implementation 1

```
def is_primary_key(df, column_name):
    return df[column_name].value_counts().max() == 1
```

Implementation 2

```
def is_primary_key(df, column_name):
    return df[column_name].is_unique
```

Implementation 3

```
def is_primary_key(df, column_name):
    return len(df[column_name].unique()) == len(df[column_name])
```

- ○ Implementation 1
- ○ Implementation 2
- ○ Implementations 1 and 2
- ○ Implementations 1 and 3
- ● Implementations 1, 2, and 3
- ○ None of the above

(c) (3.0 pt) We would like to determine the college with the greatest number of college first-year players. Which of the following would achieve that goal?

- ● `college_ball["college"].value_counts(ascending=False)[0]`
- ○ `college_ball["college"].unique()[0]`
- ○ `college_ball["college"].max()`

Note: Full credit was awarded to all students for part (c)

(d) (5.0 pt) Now we want to create a dataframe called `relative`. The code snippet to create this dataframe is given below:

```
hometown_av = college_ball[['hometown', 'ppg']].groupby('hometown').agg(np.average)
hometown_av = hometown_av.rename(columns={'ppg' : 'h_ppg'})
relative = college_ball.merge(hometown_av, how='left',
                              left_on='hometown', right_index=True)
relative['high_performance'] = np.abs(relative['h_ppg'] - relative['ppg']) > 2
```

Identify which of the following outputs is generated when we call `relative.head()`:

○

	name	hometown	college	high_performance
0	Lamelo Ball	Chino Hills, CA	University of California, Los Angeles	0.0
1	James Wiseman	Nashville, TN	University of Memphis	2.1
2	Tyrese Haliburton	Oshkosh, WI	University of Memphis	0.0
3	Anthony Edwards	Atlanta, GA	University of Georgia	0.0
4	Tyler Herro	Bentonville, AR	University of Georgia	2.4

○

	name	hometown	college	ppg	h_ppg	high_performance
0	Lamelo Ball	Chino Hills, CA	University of California, Los Angeles	15.6	15.6	0.0
1	James Wiseman	Nashville, TN	University of Memphis	18.3	20.4	-2.1
2	Tyrese Haliburton	Oshkosh, WI	University of Memphis	25.5	25.5	0.0
3	Anthony Edwards	Atlanta, GA	University of Georgia	10.2	10.2	0.0
4	Tyler Herro	Bentonville, AR	University of Georgia	18.5	20.9	-2.4

○

name	hometown	college	ppg	h_ppg	high_performance
Lamelo Ball	Chino Hills, CA	University of California, Los Angeles	15.6	15.6	False
James Wiseman	Nashville, TN	University of Memphis	18.3	20.4	True
Tyrese Haliburton	Oshkosh, WI	University of Memphis	25.5	25.5	False
Anthony Edwards	Atlanta, GA	University of Georgia	10.2	10.2	False
Tyler Herro	Bentonville, AR	University of Georgia	18.5	20.9	True

●

	name	hometown	college	ppg	h_ppg	high_performance
0	Lamelo Ball	Chino Hills, CA	University of California, Los Angeles	15.6	15.6	False
1	James Wiseman	Nashville, TN	University of Memphis	18.3	20.4	True
2	Tyrese Haliburton	Oshkosh, WI	University of Memphis	25.5	25.5	False
3	Anthony Edwards	Atlanta, GA	University of Georgia	10.2	10.2	False
4	Tyler Herro	Bentonville, AR	University of Georgia	18.5	20.9	True

(e) (8.0 pt) We want to see which colleges attract the most local players, which we denote by those first-year players whose hometown is less than $num2$ miles away from the college that they attend. Assume that we have added the column `distance` to `college_ball`, which gives the distance, in miles, between the `hometown` and `college` for each player. Please write pandas code to assign `proportion_local` below to a series that contains the proportion of local first-year players at each college. Make sure that the index of `proportion_local` is `college`.

```
proportion_local = ________
```

Note: Your code must be one line. Any code with more than five pandas calls (e.g. `groupby`, `agg`, `apply`, etc.) will not be accepted.

```
college_ball.groupby('college').apply(lambda group : (np.sum(group['distance']
< 2)) / group.shape[0])['distance']
```

3. (14.0 points)

Kunal really wants some boba, so he decides to do some research to pick the best boba shop in Berkeley. He collects data on all the different boba shops in Berkeley and compiles it into a table named `boba`. The first five rows are shown below:

name	avg_rating	num_reviews	avg_price
one_plus	5	69	5
plentea	3.5	128	5
tp_tea	4.5	120	4.75
asha_tea_house	4	1411	4.25
raretea_berkeley	3.5	157	5

(a) (6.0 pt) Kunal wants to find the average rating of all boba shops with an average price of num1 dollars or greater. Write a SQL query that achieves this. The final output should be a single row.

```
SELECT AVG(avg_rating) FROM boba WHERE avg_price >= 5
```

(b) (8.0 pt) As a fan of Plentea, Andrew doubts some of the data that Kunal has collected. He wants to double check that Kunal has actually counted all the reviews available. He knows Kunal used Yelp to find the average ratings, so he retrieves a SQL table called `reviews` with all the reviews submitted in Berkeley, CA to Yelp. The first five rows are shown below:

shop_name	username	rating	review
viks_chaat	spicebear	4	Great South Indian Food!
plentea	andrew_the_TA	5	I <3 Boba!!
chez_panisse	foodblog21	4	worth the long wait!
asha_tea_house	silverfox99	5	lots of great choices!
chez_panisse	stanfordstudent	2	I don't like Berkeley

Write a SQL query that outputs a table with three columns. The first column should be the shop names in `boba`, the second column should be the number of reviews based on Kunal's data. The third column should be the number of reviews based on Andrew's data. If a boba shop has no reviews in Andrew's data, its third column can be `0` or `NaN`.

```
SELECT name, num_reviews, total
        FROM boba LEFT JOIN
            ( SELECT shop_name, COUNT(*) AS total FROM reviews GROUP BY shop_name )
            ON name = shop_name


Alternate Solution:


  SELECT B.name, MAX(B.num_reviews), COUNT(R.rating) FROM boba B
    LEFT JOIN reviews R
    ON B.name = R.shop_name
    GROUP BY B.name
```

4. (8.0 points)

(a) (6.0 pt) Oh no! Kunal is writing up a new Data 100 discussion worksheet in LaTeX, but he made some mistakes. All of his summations are supposed to end at sub_n, but some of them end at sub_m instead. Here is a sample of lines in his document, with the LaTeX directly rendered next to each one.

`"$\mu = \sum_{i = 1}^{sub_m} x_i$"` $\mu = \sum_{i=1}^{sub_m} x_i$

`"$\sigma^2 = \sum_{i=1}^{sub_n} (x_i - \mu)^2$"` $\sigma^2 = \sum_{i=1}^{sub_n} (x_i - \mu)^2$

`"$L(\theta) = \prod_{i=1}^{sub_m} f(x_i | \theta)$"` $L(\theta) = \prod_{i=1}^{sub_m} f(x_i|\theta)$

`"$\sum_{j=0}^{sub_m} {sub_n \choose j} p^j (1-p)^{sub_n-j} = 1$"` $\sum_{j=0}^{sub_m} \binom{sub_n}{j} p^j (1-p)^{sub_n - j} = 1$

`"$sub_m = 100$"` $sub_m = 100$

`"The sum goes from 1 to sub_m."` The sum goes from 1 to sub_m.

Given the above strings, write a regular expression to find the offending character. That is, your regex should match the character that should be sub_n, but is not. If there are no mistakes in the LaTeX, your regex should not return a match. Your regex will be graded on whether or not the following code executes correctly, assuming all of the above strings have been loaded into a list called `strings`. You **only** need to worry about these test cases—do not worry about any other edge cases not considered in the examples.

```
import re
pattern = ...
for string in strings:
    print(re.findall(pattern, string))
```

```
['sub_m']
[]
[]
['sub_m']
[]
[]
```

`pattern =`

```
\\sum.*\{([^{i,n}]+)\}
```

(b) (2.0 pt) Assume that the regex you wrote above works correctly. How do we replace the matched character in `string` with sub_n?

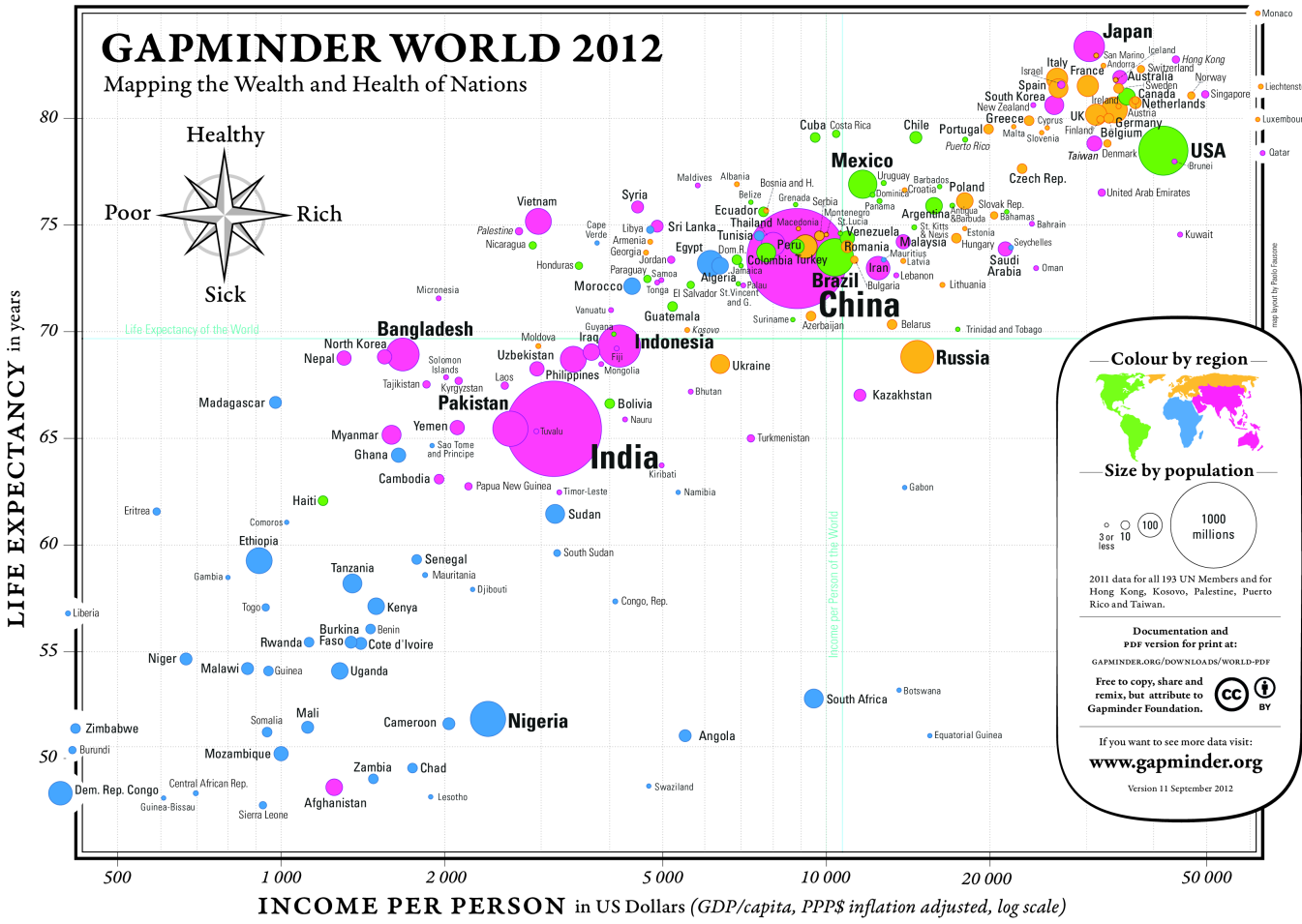
- ○ `re.replace(pattern, "sub_n", string)`
- ● `re.sub(pattern, "sub_n", string)`
- ○ `string.replace(pattern, "sub_n")`
- ◉ None of the above

Note: Both "re.sub.." and "None of the above" were accepted as valid answers

5. (11.0 points)

(a) (3.0 points)

Below is a colored graphic produced by the Gapminder Foundation.



If the image above is too small, you can find a direct link to the original PDF here.

i. (2.0 pt) One aspect of the plot that encodes data is color. Which of the following are other aspects in the plot that encode data ?

- [x] Size of the circle
- [x] Circle position along x-axis
- [x] Circle position along y-axis
- [ ] Location of text in relationship to circle
- [x] Size of text

Note: "Size of text" is optional

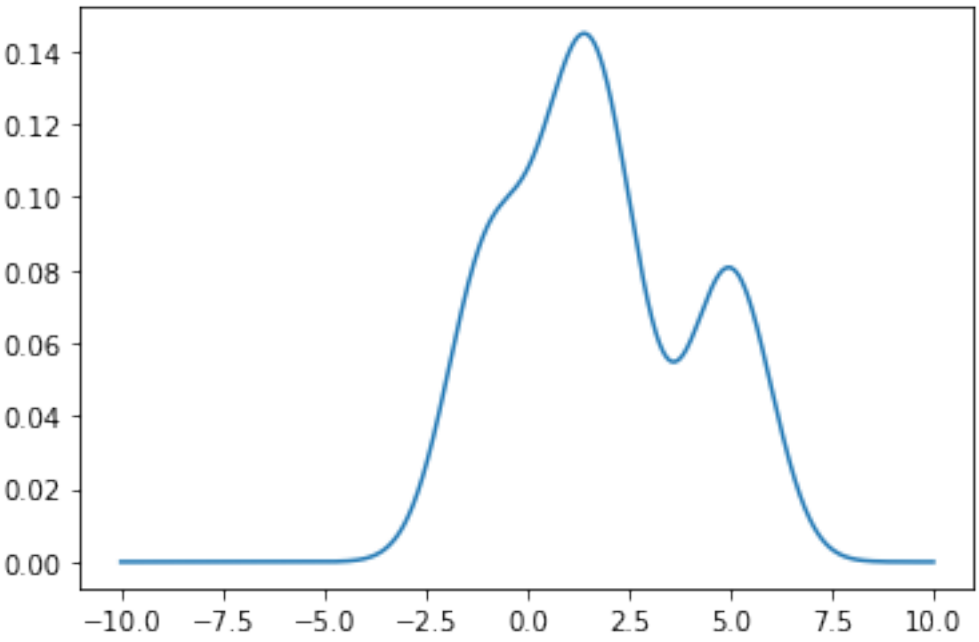
ii. (1.0 pt) True or false? The graphic above appears to show a linear relationship between a country's **income per person in US Dollars**, and its **life expectancy**.

- ○ True
- ● False

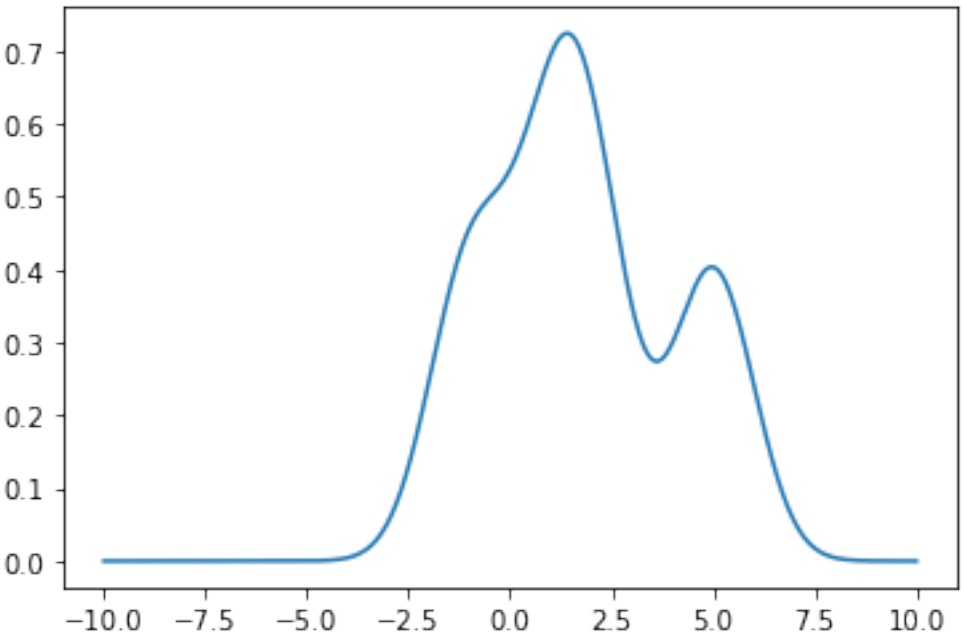
(b) **(2.0 pt)** Suppose I am a data scientist for Reddit who wants to find out more about the r/wallstreetbets community. What is the best way for me to see the distribution of Reddit users by education level (e.g. high school, bachelors, masters, PhD)?

- ● A Bar Chart
- ○ A Scatter Plot
- ○ A Pie Chart

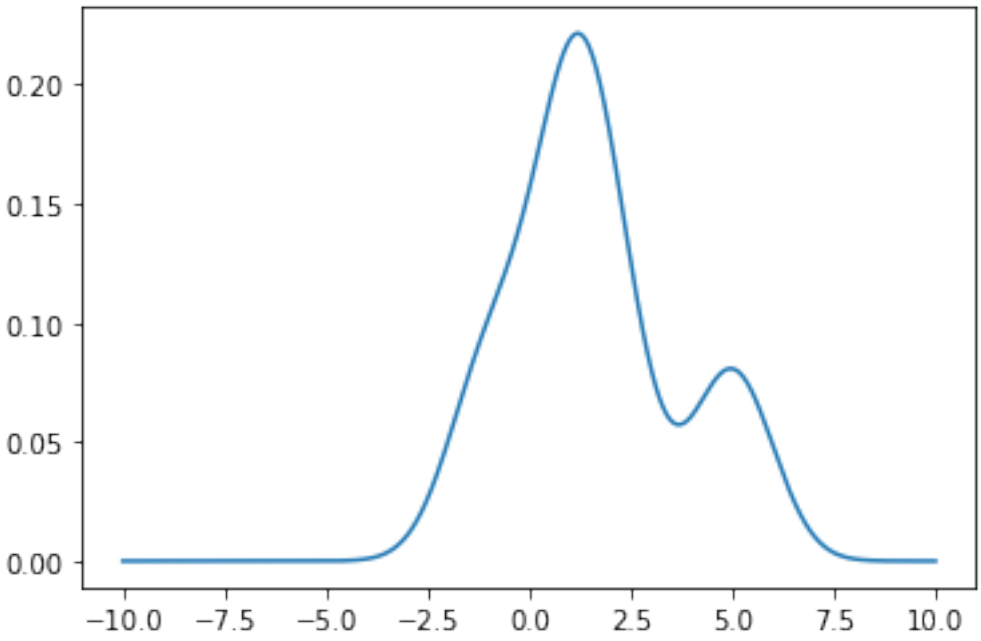
(c) (4.0 pt) Suppose we have a dataset $X = [1, 1, -1, 5, 2]$. Which of the following is the KDE plot of X with $\alpha = 1$ (smoothing factor)?



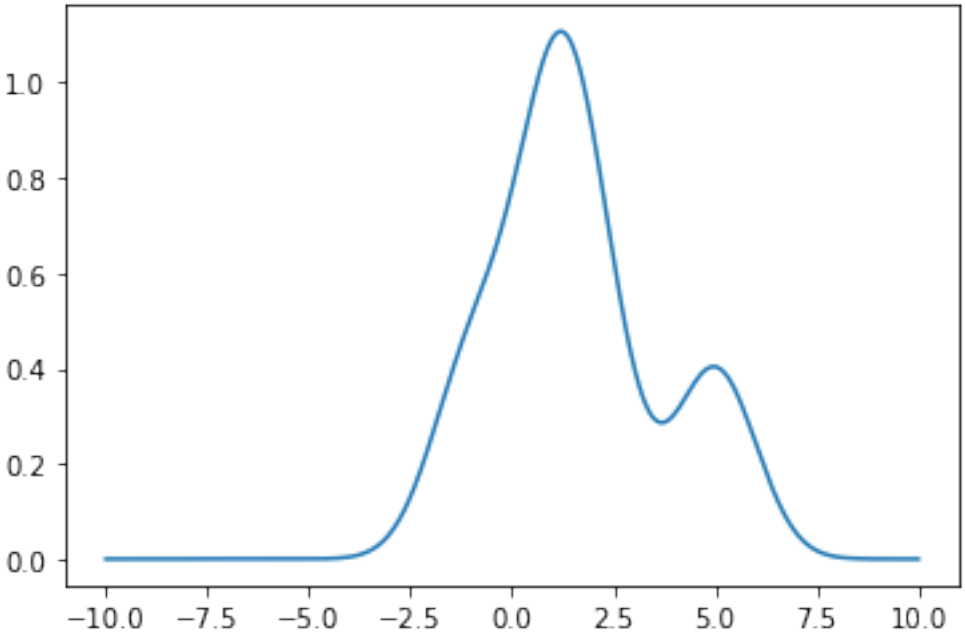
○



○

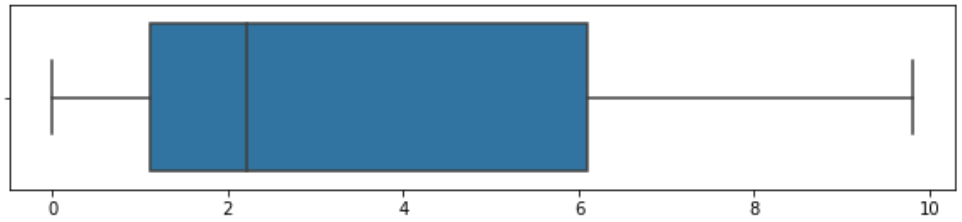

0.20
0.15
0.10
0.05
0.00
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

●


1.0
0.8
0.6
0.4
0.2
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

○

(d) (2.0 pt) Which of the following is guaranteed to be true about the boxplot below?



- ☐ The distribution is left skewed
- ☐ The distribution is unimodal
- ☐ The distribution is bimodal
- ■ The distribution is right skewed
- ☐ The distribution is symmetric

6. (13.0 points)

(a) (4.0 pt) Suppose we have scalar data points $x_1, \ldots, x_n$ and we use the following loss to fit the data: $R(\theta) = -n \log \theta + \theta \sum_{i=1}^{n} x_i$.

Compute $\underset{\theta}{\operatorname{argmin}} R(\theta)$. Note we denote $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$ in the options below.

- ○ $n\mu$
- ● $\frac{1}{\mu}$
- ○ $\frac{n}{\mu}$
- ○ μ

(b) (3.0 pt) Select all correct statements about MAE (mean absolute error) and MSE (mean squared error) below:

- ☐ Both MAE and MSE are smooth functions.
- ☐ MAE is more sensitive to outliers than MSE.
- ☐ Both MAE and MSE may have multiple distinct minimizers.
- ☐ The minimizers of MAE and MSE on the same dataset are always different.
- ■ None of the above.

(c) (6.0 pt) Suppose we have a dataset $\{(x_i, y_i)\}_{i=1}^{n}$ and we model y as $\hat{y} = f_\theta(x) = \theta \sin(x)$. Assume we use the MSE (mean squared error) loss:

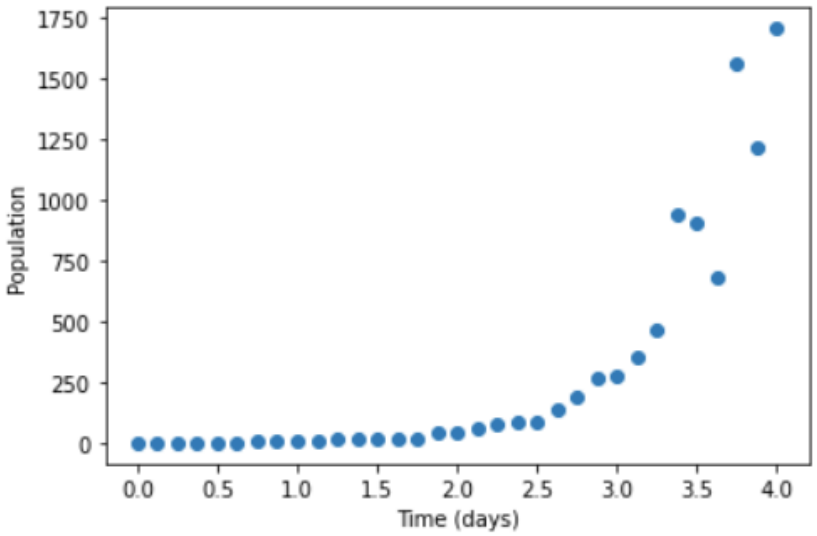
$$\frac{1}{n} \sum_{i=1}^{n} (f_\theta(x) - y_i)^2$$

What is the minimizer $\hat{\theta}$ of this loss:?

- ○ $\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$
- ● $\frac{\sum_{i=1}^{n} \sin(x_i) y_i}{\sum_{i=1}^{n} \sin(x_i)^2}$
- ○ $\frac{\sum_{i=1}^{n} \sin(x_i) y_i}{\sum_{i=1}^{n} \sin(x_i^2)}$
- ○ $\frac{\sum_{i=1}^{n} \cos(x_i) y_i}{\sum_{i=1}^{n} x_i^2}$

7. (14.0 points)

(a) (2.0 pt) Scientists are trying to experimentally model bacterial growth over the span of 4 days. The plotted data can be seen below.



In order to further understand the relationship present in the data, the scientists would like to linearize the relationship. Which of the following transformations would create a linear relationship.

- (●) x and $\log(y)$
- ( ) $\log(x)$ and $\log(y)$
- ( ) x^2 and y
- ( ) $\log(x)$ and y
- ( ) None of the above

(b) (4.0 pt) The scientists get a new batch of data. After running linear regression, the scientists obtained a correlation coefficient of 0.969 between x (time) and y (population) denoting a strong linear relationship, and the regression model $y = 0.04 + 2.01x$.

Using the linearized data, the scientists would like to calculate the ratio of standard deviations between the transformed variables (in other words, they'd like to obtain an estimate for σ_x/σ_y). Which of the following values is the closest to this ratio?

- ◯ 2
- ● 1/2
- ◯ 1
- ◯ 1/4
- ◯ None of the above

(c) (8.0 points)

The scientists get a new batch of data again, but the relationship is not as strong anymore. We are experimenting with different loss functions to use to fit the data to the model $y = \theta x$, and have produced four graphs. Each of the graphs correspond to EXACTLY ONE of the following loss functions that was minimized to generate the line over the data:

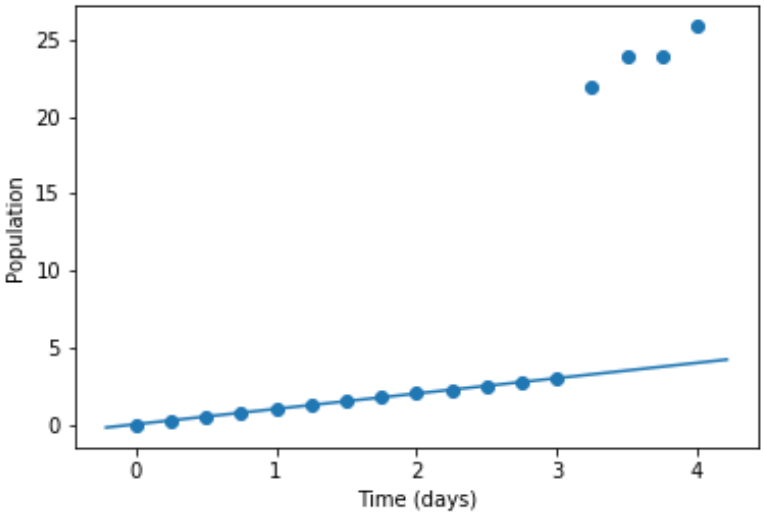
$L_1 = L(x_i, y_i; \theta) = e^{x_i}(y_i - \theta x_i)^2$

$L_2 = L(x_i, y_i; \theta) = e^{-x_i}(y_i - \theta x_i)^2$

$L_3 = L(x_i, y_i; \theta) = (y_i - \theta x_i)^2$

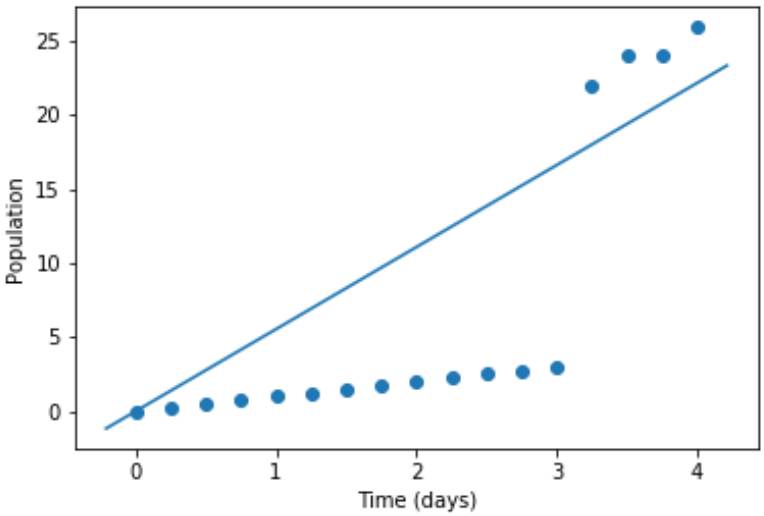
$L_4 = L(x_i, y_i; \theta) = (y_i - \theta x_i)^2$ if $y_i < 10$, otherwise 0

We have lost our code and are only left with the plots. We want to select the loss function that corresponds to each of the graphs:



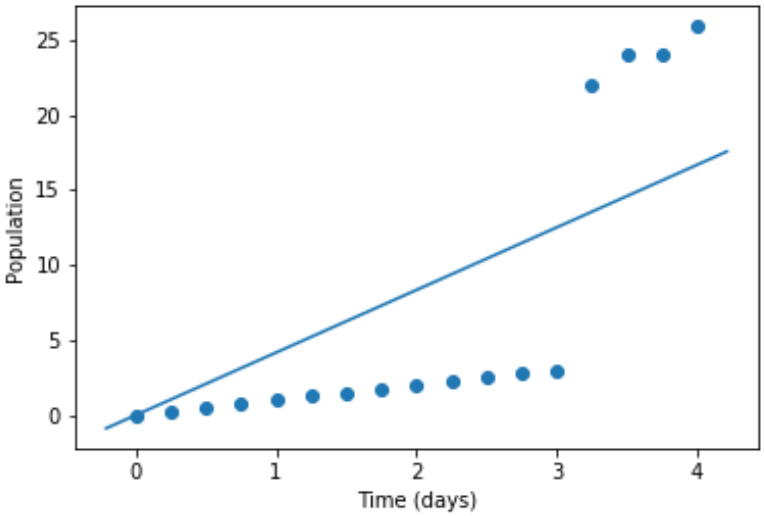
i. (2.0 pt)

- ◯ L_1
- ◯ L_2
- ◯ L_3
- ● L_4



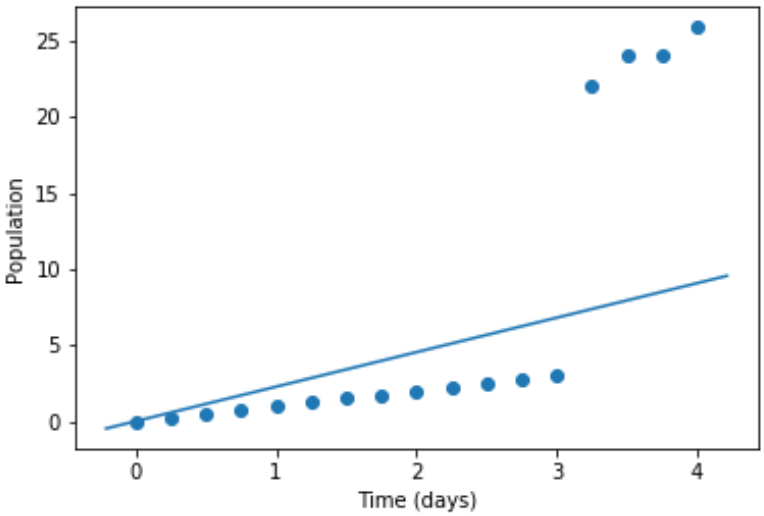
ii. (2.0 pt)

- ● L_1
- ○ L_2
- ○ L_3
- ○ L_4



iii. (2.0 pt)

- ○ L_1
- ○ L_2
- ● L_3
- ○ L_4



iv. (2.0 pt)

- ○ L_1
- ● L_2
- ○ L_3
- ○ L_4

8. (9.0 points)

(a) (6.0 pt) Suppose we have an $n \times d$ design matrix $\mathbf{X}$ and an outcome vector $\mathbf{y}$. Note that the first column of $\mathbf{X}$ is a column of ones. We want to regress $\mathbf{y}$ on $\mathbf{X}$ and have chosen the model

$$\mathbf{y} = \mathbf{X}\theta$$

For a single observation,

$$\mathbf{y}_i = \mathbf{X}_i \cdot \theta = \sum_{j=1}^{d} \mathbf{X}_{ij}\theta_j$$

We are performing OLS to regress $\mathbf{y}$ on $\mathbf{X}$, and we will denote the optimal model weights as $\hat{\theta}$. Let $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ be the vector of residuals of our model.

Recall the the first column of $\mathbf{X}$ is a column of ones, a bias (intercept) term. If we did *not* have a bias term, which of the following relationships would **no longer be true**?

- ☐ $\mathbf{X}^\top \mathbf{X}\hat{\theta} = \mathbf{X}^\top \mathbf{y}$
- ☐ $\hat{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$
- ☐ $\mathbf{X}^T e = 0$
- ■ $\mathbf{1}^T e = 0$, where $\mathbf{1}$ is a vector of ones
- ☐ $\mathbf{e} \perp \text{span}\{\mathbf{X}\}$
- ☐ $\mathbf{e} \perp \hat{\mathbf{y}}$
- ■ $\sum_{i=1}^{n} \mathbf{e}_i = 0$
- ■ $\frac{1}{n}\sum_{i=1}^{n} \mathbf{e}_i = 0$
- ■ $\frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{y}}_i$

(b) (3.0 pt) We have a DataFrame named for a small store named `store_data`, which only has 4 entries, and is given below:

	apples	pears	total_price
0	2	2	14
1	0	3	9
2	1	3	13
3	2	1	11

It contains customer data on the number of apples and pears that each customer purchases. The price of apples is fixed at \$4 and the price of pears is fixed at \$3, and the total price is computed as

```
store_data['total_price'] = store_data['apples']*4 + store_data['pears']*3
```

We convert this dataframe into a numpy matrix called `X` of dimension $\mathbb{R}^{4\times 3}$, where the rows and columns are the same as the ones in `store_data`. If we run:

```
theta_hat = np.linalg.inv(X.T @ X) @ X.T @ y
```

What error would we expect on the above line, if any?

- ◯ `ValueError: dimension mismatch`
- ● `LinAlgError: Singular matrix` (Singular is another word for non-invertible)
- ◯ `ZeroDivisionError: division by zero`
- ◯ No Error

No more questions.