DATA 8 Summer 2021 Sample Exam.

FINAL 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.

For fill-in-the-blank coding questions, you can put anything inside the blanks, including commas, parentheses, and periods.

The exam is out of 180 points.

If you encounter any logistical problems during the exam, please contact us at data8berkeley@gmail.com.

good_luck

The Exam is composed of following sections (the order on your exam may be different):

- Data Eight-Book (35 Points)
- Multiverse of Miscellaneous Questions (35 Points)
- Be Like Mike (Which One!?) (30 Points)
- GO BEARS! (17 Points)
- Data 8 Has A Pizza My Heart (30 Points)
- Data 9 (38 Points)

(a) Your name:

(b) Your @berkeley.edu email address:

(c) The Berkeley Honor Code states: "As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." Do you agree to follow the honor code on this exam?

○ Yes

○ No

1. (35 points) Data Eight-Book

Yanay is trying to create a social media platform for Data 8 staff.

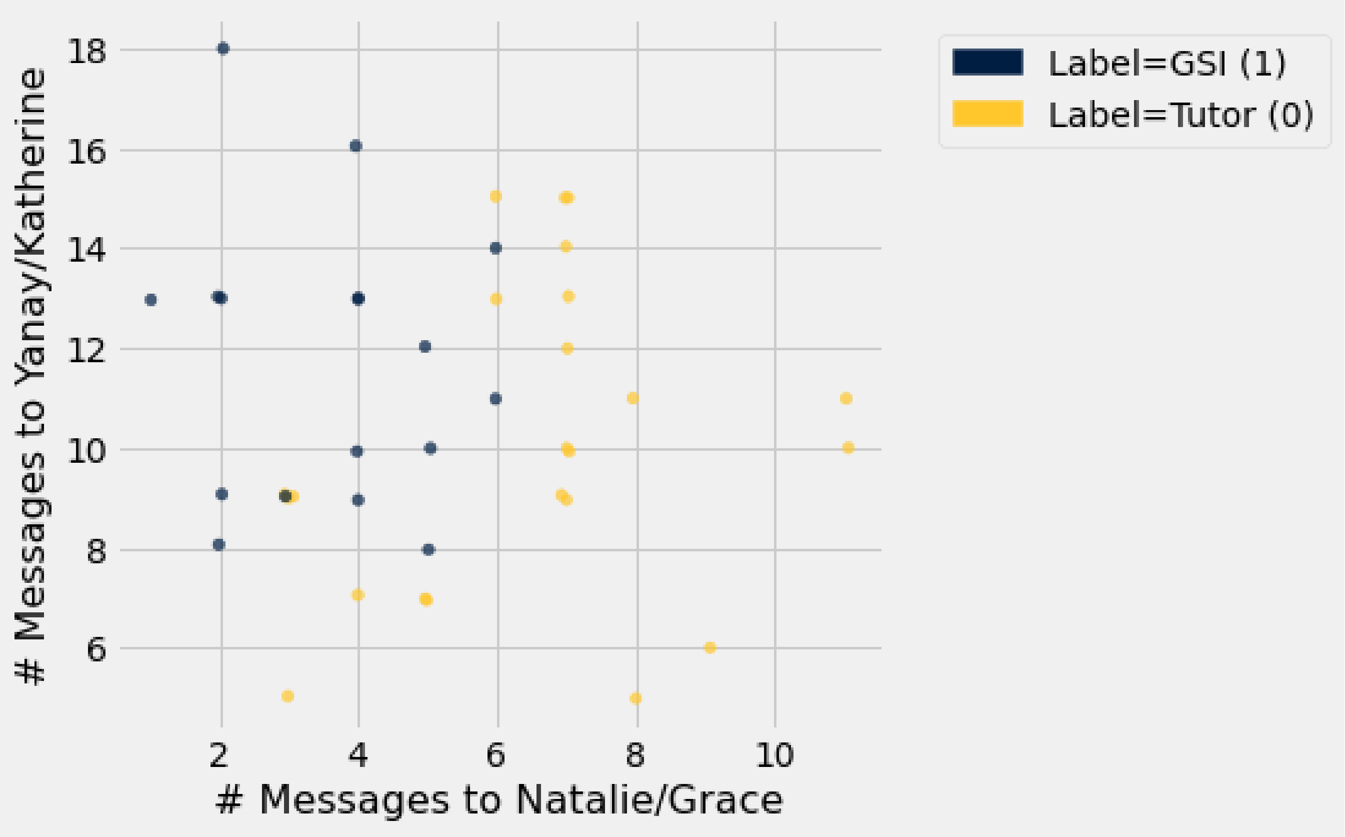
He wants to understand the connections between staff and decides to use machine learning.

First, Yanay decides to use data he collects from Slack (a messaging platform) to see if he can distinguish between tutors and GSIs. Yanay has access to Direct Messaging data since he is the administrator of the Slack group.

Note: all of the data in this question is fake

Note: The data are jittered (i.e. every point on the scatterplot has been moved by a small random amount so that the points don't all overlap).

Note: Tutors directly report to Natalie and Grace



knn

(a) i. **(3 pt)** Yanay runs a 5-Nearest Neighbors algorithm on these two features. What would his training accuracy be closest to? (The training set is visualized in the scatter plot). Training points will be excluded before they are classified.

● 90%

○ 100%

○ 50%

○ 0%

○ 34%

○ 42%

ii. **(2 pt)** Chenxi is in the test set. Chenxi has messaged Yanay/Katherine 5 times and Grace/Natalie 5 times. What would the 5-Nearest Neighbors algorithm predict for Chenxi's class?

● Tutor

○ GSI

(b) **i.** **(3 pt)** After remaking the train/test set, Yanay notices that there are 15 Tutors and 15 TAs in the new training set.

What is the smallest value of k that will always lead the classifier to always predict tutor?

● None of these values

○ 31

○ 29

○ 30

○ 14

○ 16

○ 15

○ 5

○ 17

ii. **(3 pt)** After remaking the train/test set, Yanay notices that there are 16 Tutors and 14 TAs in the new training set.

What is the smallest value of k that will always lead the classifier to always predict tutor?

○ No value

○ 31

● 29

○ 30

○ 14

○ 16

○ 15

○ 5

○ 17

(c) **(2 pt)** What are some privacy or ethical considerations that Yanay needs to account for in his machine learning model?

There are many ethical and privacy concerns to be considered. One example: it might be a violation of privacy to look at others' private messages (especially if you don't tell them).

(d) Yanay wants to design some features of his social media platform to help connect friend groups within staff.

For each member of staff, he collects information about which other members of staff they are Facebook friends with, follow on Instagram, or follow on Twitter.

A subset of some of the rows and columns from the table is shown below:

Position	Name	FB Yanay	Instagram Yanay	Twitter Yanay	FB Katherine	Instagram Katherine	Twitter Katherine
GSI	Ryan	1	0	0	1	1	0
GSI	Rita	1	0	0	1	1	1
GSI	Grace	1	1	0	1	1	1
GSI	Natalie	0	0	0	1	0	0

Note: rows represent staff members, columns represent whether or not you are friends with/follow another staff member on a specific platform, 1 means yes, 0 means no.

i. **(4 pt)** What kinds of analysis could Yanay do with this data? (Select all that apply).

■ AB Test the association between staff position and whether or not a staff member follows Yanay on Twitter.

■ Test the hypothesis that the association between following Katherine on Instagram and following Katherine on Twitter is 0.

■ Test the hypothesis that the association between being Facebook friends with Yanay and being Facebook friends with Katherine is negative.

■ AB Test the association between staff position and number of staff member friends on Facebook.

□ Test the hypothesis that staff position (GSI or Tutor) is determined by number of staff member friends on Facebook

In order to start identifying friend groups Yanay decides to implement a modified version of the K nearest neighbors algorithm. The modified algorithm works like this:

ii. A. Randomly select `m` points from the dataset without replacement, these are our starting points, also called our "center points". *Note:* `m` *will always be less than the size of the training data.*

B. Use a knn classifier with `k` = 1 on the entire dataset, with the `m` randomly selected points as the training set. The labels of the randomly selected points are 0, 1 ... `m`-1, corresponding to their order in which they were sampled.

C. Update the feature values of the m points to be equal to the average feature value of the points in the dataset which we assigned the corresponding label to.

D. Repeat steps 2 and 3 a certain number of times. Use the final `m` "center points" as the training set for future classification tasks (with `k`=1).

`m` represents the number of friend groups we are identifying. Assume it has been defined for you.

`features` is a table with only our features (the information about staff member social media).

We'll use the next few questions to implement this algorithm.

A. **(4 pt)** For the first step, fill in the following code to select `m` random points from our table of features `features`. **(Part A)**

```
center_points = features.sample(_____, ________)
center_points = center_points.with_column('Label', ___________)
```

center_points = features.sample(m, with_replacement=False) center_points = center_points.with_column('Label', np.arange(0, m))

B. **(3 pt)** Fill in the following code to perform the second step of the algorithm. Assume the function `classify(test, k, train)` has been implemented correctly.

The `classify` function will classify every row in the `test` table according to a `k`-nearest neighbor algorithm using the rows in the `train` table. The function returns a table with all the original columns of the `test` table, with the added column `Label` that contains the predicted labels.

(Part B)

```
classifications = classify(____, _____, ______)
```

classifications = classify(features, 1, center_points)

C. **(4 pt)** Fill in the following code to perform the third step of the algorithm. **(Part C)**

```
new_center_points = ___________.___________(______________)
```

new_center_points = classifications.group(label, np.mean)

D. (4 pt) Select the option that correctly implements the full algorithm: Assume the `classify` function treats columns named `"feature`’’ and ‘‘`feature mean`’’ as the same.

○
```
num_iterations = 10
(Part A)
for i in np.arange(num_iterations):
    (Part B)
(Part C)
center_points = new_center_points
```

○
```
num_iterations = 10

for i in np.arange(num_iterations):
    (Part A)
    (Part B)
(Part C)
center_points = new_center_points
```

●
```
num_iterations = 10
(Part A)
for i in np.arange(num_iterations):
    (Part B)
    (Part C)
center_points = new_center_points
```

○
```
num_iterations = 10

for i in np.arange(num_iterations):
    (Part A)
    (Part B)
    (Part C)
center_points = new_center_points
```

E. (2 pt) Yanay is having trouble choosing a value of `m`.

He decides to use numerical optimization to choose the value of m for him.

He defines the function `error` which takes a value for `m` as an argument and returns the training error value calculated for the algorithm using the value of `m`.

Write a line of code to determine the best value of `m`.

```
minimize(error)
```

F. (1 pt) What kinds of errors could this machine learning system make and will certain errors lead to worse outcomes than other kinds of error?

The system might classify different people as friends or not friends incorrectly. The difference in outcome for these two kinds of errors is a judgement call for which you think might be worse (or if they are the same). We will accept most answers that show thought!

2. (35 points) Multiverse of Miscellaneous Questions

(a) (5 pt) We mentioned in lecture that in real world data science we use 3 sets of data to evaluate classifier accuracy. One of the most popular methods to analyze classifier accuracy is called “k-Fold Cross Validation”. Here is how one implementation of the algorithm works.

i. Divide your training data into `k` equally sized, randomly chosen groups.
ii. Repeat the following process `k` times: choose one group and set it aside, then train your classifier on the remaining `k-1` groups. Evaluate the accuracy of your classifier on the `1` group that you left out of the training set.
iii. Return the average accuracies of all k classifiers you trained.

Fill in the blanks to implement this algorithm. Assume the following has been defined for you:

- The table `data` which has a column of labels called `’’Label’’` and columns of features. It also has a column of row numbers, which start at 0, called `’’Number’’`.
- The function `classify(train, test)` which returns an array of predicted labels from the classifier trained on the `train` table argument.
- The number `k` which corresponds to the number of classifiers to train. Assume that the number of rows in the `data` table is divisible by `k`.

```
accuracies = make_array()
shuffled_data = _______________________

for i in np.arange(k):
    row_numbers = _______________________
    test_set = ________________________________
    train_set = ________________________________
    predictions = classify(train_set, test_set)
    accuracy = np.mean(predictions == test_set.column(‘‘Label))
    accuracies = np.append(accuracies, accuracy)
np.mean(accuracies)
```

```
accuracies = make_array()
shuffled_data = data.sample(with_replacement=False)

for i in np.arange(k):
    row_numbers = np.arange(i, shuffled.num_rows, k)
    test_set = shuffled_data.where(‘‘Number’’, are.contained_in(row_numbers))
    train_set = shuffled_data.where(‘‘Number’’, are.not_contained_in(row_numbers))
    predictions = classify(train_set, test_set)
    accuracy = np.mean(predictions == test_set.column(‘‘Label))
    accuracies = np.append(accuracies, accuracy)
np.mean(accuracies)
```

(b) (5 pt) Kseniya has a table called `CARD_TBL` that contains information about a special deck of cards she created to cheat at poker.

Color	Suite	Rank
Red	Diamond	Ace
Black	Spade	Ace
Black	Club	Queen

{Rows Omitted}

Kseniya is going to draw 2 cards from the deck, with replacement (the first card goes back into the deck and the deck is reshuffled before the next draw).

How many different combinations of 2 card draws could Kseniya get, if the "CARD_COL" is the same for the first card and second card?

Fill in the following line of code so that it will evaluate to that number:

```
_____._____(________)._________
```

```
CARD_TBL.join(“CARD_COL”, CARD_TBL).num_rows
```

(c) **(4 pt)** One guest lecturer this summer described an experiment involving providing chlorine water purifying tablets to rural villages.

Which of the following statements are true? (Select all that apply).

■ If the experiment organizers had distributed tablets to villages only located close to main roads they would not have been able to make causal claims because they had a convenience sample.

■ If the experiment organizers had randomly distributed tablets to different schools they could have made causal claims because they had performed a RCT.

■ If the experiment organizers had randomly distributed tablets to different students within different schools they could have made causal claims because they had performed a RCT.

□ If the experiment organizers performed an AB test they could make causal claims

□ If the experiment organizers randomly shuffled their treatment labels, they could make causal claims.

■ The experiment organizers should account for ethical and moral considerations before making decisions based on data they collect from their experiment.

(d) **(4 pt)** Which of the following are true regarding experiments to determine the efficacy of vaccines like the COVID-19 vaccine? (Select all that apply).

■ A sample size of tens of thousands is probably sufficient to create a narrow confidence interval for a population proportion (such as the proportion of extreme cases).

■ One confounding factor in experimental design might be if patients knew if they had received a vaccine or not. A way to alleviate this is to give all patients a shot, but alternate between vaccine shots and shots of a harmless substance like saline.

■ The results of a vaccine efficacy experiment might be tested for significance using an AB test.

□ Patient consent and potential side effects should not be accounted for when designing an experiment.

(e) (4 pt) Which of the following are correct examples of the listed method of learning something about an individual? Choose the best answers. Select all that apply.

■ Disclosure: Students filled out the welcome survey at the beginning of the class and identified themselves using their Berkeley email

■ Collection: The instructor sees discussion about the class on a website like Reddit and recognizes usernames.

■ Inference: The instructor predicts what the sentiment of a course evaluation a particular student might make based on their previous feedback and grade in the class.

□ Inference: Students filled out the welcome survey at the beginning of the class and identified themselves using their Berkeley email

□ Disclosure: The instructor sees discussion about the class on a website like Reddit and recognizes usernames.

□ Collection: The instructor predicts what the sentiment of a course evaluation a particular student might make based on their previous feedback and grade in the class.

(f) (4 pt) Jessie wants to visualize how the distribution of the number of "Aww" comments she observed for different lecture art changed between each unit: Programming, Inference and Machine Learning.

What is the best visualization she should choose?

● Overlaid Histogram

○ Histogram

○ Overlaid Bar Chart

○ Bar Chart

○ Scatter Plot

○ Line Plot

(g) (4 pt) One rigorous definition of outliers is as follows:

A number in a dataset is an outlier if it is less than the 25th percentile, or greater than the 75th percentile, by a value equal to 1.5 times the Interquartile Range (IQR).

The IQR is defined as the distance between the 75th percentile and the 25th percentile.

Complete the following function so that it returns a boolean array, where values are true if the value at the same index in `arr` is an outlier, and false otherwise.

```
def outlier(arr):
    iqr = _________________________
    outliers = make_array()
    for value in arr:
        if value  ___________________:
            outliers = np.append(outliers, True)
        elif ________________:
            outliers = np.append(outliers, True)
        else:
            outliers = np.append(outliers, False)
    return outliers
```

```
def outlier(arr):
    iqr = percentile(75, arr) - percentile(25, arr)
    outliers = make_array()
    for value in arr:
        if value  <= percentile(25, arr) - 1.5 * iqr:
            outliers = np.append(outliers, True)
        elif value >= percentile(75, arr) + 1.5 * iqr:
            outliers = np.append(outliers, True)
        else:
            outliers = np.append(outliers, False)
    return outliers
```

(h) (5 pt) In future probability classes you will learn about a related theorem to Chebyshev's inequality, called "Markov's Inequality".

Markov's inequality states that, for all distributions of a **non-negative** variable (a variable that will always be greater than or equal to 0), the probability that a random draw from the distribution is greater than some positive, non zero number a, is less than or equal to the mean of the distribution divided by a.

That is, for some $a > 0$, the probability that a random draw from the distribution is greater than or equal to a, which we can call $P(\text{Draw from Dist} >= a)$, will be less than or equal to $(\text{Mean of Dist})/a$

$P(\text{Draw from Dist} >= a) \leq (\text{Mean of Dist})/a$

Jackie wants to write a simulation to confirm this for herself. Which of the following code blocks successfully test this inequality?

■
```
def test_markov(distribution, a):
    checks = make_array()
    for i in np.arange(10000):
        value = distribution.sample(1).column(0).item(0) # Distribution is a one column table
        checks = np.append(checks, value >= a)
    return np.mean(checks) <= np.mean(distribution.column(0)) / a
test_markov(Table().with_column("Face", np.arange(1, 7)) , 2)
```

□
```
def test_markov(distribution, a):
    return np.mean(distribution.column(0)) <= np.mean(distribution.column(0)) / a
test_markov(Table().with_column("Face", np.arange(1, 7)) , 2)
```

□
```
def test_markov(distribution, a):
    checks = make_array()
    for i in np.arange(10000):
        value = distribution.sample(1).column(0).item(0) # Distribution is a one column table
        checks = np.append(checks, value >= a)
    return np.mean(checks) <= np.mean(distribution.column(0)) / a
test_markov(Table().with_column("Face", np.arange(1, 7)) , -2)
```

■
```
def test_markov(distribution, a):
    checks = make_array()
    for i in np.arange(10000):
        value = np.random.choice(distribution) # Distribution is an array
        checks = np.append(checks, value >= a)
    return np.mean(checks) <= np.mean(distribution) / a
test_markov(np.arange(1, 7) , 6)
```

□
```
def test_markov(distribution, a):
    return np.mean(distribution) <= np.mean(distribution) / a
test_markov(np.arange(1, 7) , 6)
```

□
```
def test_markov(distribution, a):
    checks = make_array()
    for i in np.arange(10000):
        value = np.random.choice(distribution) # Distribution is an array
        checks = np.append(checks, value >= a)
    return np.mean(checks) <= np.mean(distribution) / a
test_markov(np.arange(1, 7) , -6)
```

■
```
def test_markov(binary_proportions, a):
    checks = make_array()
    for i in np.arange(10000):
        value = sample_proportions(1, binary_proportions).item(1) # Distribution is an array of
        checks = np.append(checks, value >= a)
    return np.mean(checks) <= binary_proportions.item(1) / a
    test_markov(make_array(0.3, 0.7) , 0.6)
```

```
□ def test_markov(binary_proportions, a):
      checks = make_array()
      for i in np.arange(10000):
          value = sample_proportions(1, binary_proportions).item(1) # Distribution is an array of
          checks = np.append(checks, value >= a)
      return np.mean(checks) <= (binary_proportions.item(1) + binary_proportions.item(0) * -1) /
  test_markov(make_array(0.3, 0.7) , 0.6)
```

□ None of these options

3. (30 points) Be Like Mike (Which one!?)

Your friend Divyesh tells you that he saw Michael Jordan (MJ) on Campus and asked him for an autograph.

There are three different people Divyesh could have seen:

- Michael J. Jordan (businessman, philanthropist, former basketball player) (**J**)
- Michael B. Jordan (actor and producer) (**B**)
- Michael I. Jordan (statistics professor) (**I**)

We'll refer to each Michael Jordan using their middle initials (**J**, **B** and **I**) to avoid confusion.

(a) **(1 pt)** With no prior knowledge what is the probability that the MJ Divyesh saw was **B**?

1/3

(b) **(3 pt)** **J** will sign an autograph with a 10% chance. **B** will sign an autograph with a 65% chance. **I** will not sign an autograph. What is the probability Divyesh would have received an autograph?

● $((1/3) * (.1)) + ((1/3) * (.65)) + ((1/3) * (0))$

○ 0

○ 1

○ $0.1 + 0.65 + 0$

○ $(1/3) + (1/3) + (1/3)$

○ $((1/3) * (.1)) * ((1/3) * (.65)) * ((1/3) * (0))$

○ $((1/3) + (.1)) * ((1/3) + (.65)) * ((1/3) + (0))$

(c) **(4 pt)** If Divyesh did not get an autograph, what is the probability he saw **I**?

● $(1 * (1/3))/((1/3 * .9) + (1/3 * 0.35) + (1/3 * 1))$

○ 0

○ 1

○ $0.1 + 0.65 + 0$

○ $0.9 + 0.35 + 1$

○ $(1/3 * 0.9) + (1/3 * 0.35) + (1/3 * 1)$

○ $(0 * (1/3))/((1/3 * 1) + (1/3 * 0.65) + (1/3 * 0))$

○ $(1 * (1/3))/((1/3) + (1/3) + (1/3))$

○ $(1 + (1/3))/((1/3 + .9) + (1/3 + 0.35) + (1/3 + 1))$

(d) **i.** **(4 pt)** We randomly guess who Divyesh saw and whether he got an autograph or not.

What is the probability we correctly guess both who he saw and whether or not he got an autograph?

There are six possible outcomes (one of **J**, **B**, **I** and either Autograph or No autograph), so we chose what to guess by rolling a fair, six sided die.

●
```
(1/3) * (.1) * (1/6) + (1/3) * (.9) * (1/6)
+ (1/3) * (.65) * (1/6) + (1/3) * (.35) * (1/6)
+ (1/3) * (0) * (1/6) + (1/3) * (1) * (1/6)
```

○ 0

○ 1

○ 1/3

○ 1/18

○ 1/2

○
```
(1/3) * (.1) + (1/3) * (.9)
+ (1/3) * (.65) + (1/3) * (.35)
+ (1/3) * (0) + (1/3) * (1)
```

○
```
(1/3) * (.5) * (1/6) + (1/3) * (.5) * (1/6)
+ (1/3) * (.5) * (1/6) + (1/3) * (.5) * (1/6)
+ (1/3) * (0.5) * (1/6) + (1/3) * (.5) * (1/6)
```

○
```
((1/3) * (.1)) ** 2 + ((1/3) * (.9)) ** 2
+ ((1/3) * (.65)) ** 2 + ((1/3) * (.35)) ** 2
+ ((1/3) * (0)) ** 2 + ((1/3) * (1)) ** 2
```

○
```
(1/3) * (1/6) * (1/6) + (1/3) * (1/6) * (1/6)
+ (1/3) * (1/6) * (1/6) + (1/3) * (1/6) * (1/6)
+ (1/3) * (1/6) * (1/6) + (1/3) * (1/6) * (1/6)
```

ii. **(4 pt)** We didn't really like that method of guessing so we come up with another method of guessing.

We'll roll a 3 sided fair die to determine whether or not to guess **J**, **B** or **I**.

Then, depending on the outcome of the die roll, we predict autograph status by randomly choosing Autograph or no Autograph based on our prior knowledge (.1 chance yes if **J**, .65 chance yes if **B**, 0 chance yes if **I**).

What is the probability we correctly guess both who he saw and whether or not he got an autograph?

○
```
(1/3) * (.1) * (1/6) + (1/3) * (.9) * (1/6)
+ (1/3) * (.65) * (1/6) + (1/3) * (.35) * (1/6)
+ (1/3) * (0) * (1/6) + (1/3) * (1) * (1/6)
```

○ 0

○ 1

○ 1/3

○ 1/18

○ 1/2

○
```
(1/3) * (.1) + (1/3) * (.9)
+ (1/3) * (.65) + (1/3) * (.35)
+ (1/3) * (0) + (1/3) * (1)
```

○
```
(1/3) * (.5) * (1/6) + (1/3) * (.5) * (1/6)
+ (1/3) * (.5) * (1/6) + (1/3) * (.5) * (1/6)
+ (1/3) * (0.5) * (1/6) + (1/3) * (.5) * (1/6)
```

●
```
((1/3) * (.1)) ** 2 + ((1/3) * (.9)) ** 2
+ ((1/3) * (.65)) ** 2 + ((1/3) * (.35)) ** 2
+ ((1/3) * (0)) ** 2 + ((1/3) * (1)) ** 2
```

○
```
(1/3) * (1/6) * (1/6) + (1/3) * (1/6) * (1/6)
+ (1/3) * (1/6) * (1/6) + (1/3) * (1/6) * (1/6)
+ (1/3) * (1/6) * (1/6) + (1/3) * (1/6) * (1/6)
```

iii. **(2 pt)** Are we more likely to make correct guesses using the first method (die roll) or the second method (prior knowledge)?

● Second option

○ First Option

○ They will be the same

(e) **i. (3 pt)** We realize that the location Divyesh saw MJ at could help us determine which MJ he saw.

- If Divyesh saw MJ at the Haas School of Business, there is a 90% chance that it was **J**.
- If Divyesh saw MJ at the Zellerbach Hall Auditorium, there is a 70% chance that it was **B**.
- If Divyesh saw MJ at Evans Hall, we believe there is a 95% chance it was **I**.

If MJ was **I**, what is the **lowest** chance that Divyesh saw him at Evans Hall?

● $(0.95 * 1)/(.1 + .3 + .95)$

○ 0

○ 1

○ $(1)/(.1 + .3 + .95)$

○ $(.95)/(.1 + .3)$

○ $(0.95 * 1) * (.1 + .3 + .95)$

○ $(0.95 * 1)/((.1 * 1/3) + (.3 * 1/3) + (.95 * 1/3))$

ii. (3 pt) What is the highest chance that Divyesh saw him at Evans Hall?

○ $(0.95 * 1)/(.1 + .3 + .95)$

○ 0

● 1

○ $(1)/(.1 + .3 + .95)$

○ $(.95)/(.1 + .3)$

○ $(0.95 * 1) * (.1 + .3 + .95)$

○ $(0.95 * 1)/((.1 * 1/3) + (.3 * 1/3) + (.95 * 1/3))$

iii. (3 pt) If Divyesh got an autograph at Zellerbach Hall, what is the probability that MJ was **B**?

For this question only, assume that if Divyesh saw MJ at the Zellerbach Hall Auditorium, there is a 15% chance that it was **J** and 15% that it was **I**.

● $(0.65 * 0.7)/((0.65 * 0.7) + (.1 * .15) + (0 * .15))$

○ 0

○ 1

○ $(0.65)/(0.65 + .1 + 0)$

○ $(0.7)/(0.7 + .15 + .15)$

○ $(0.65 * 0.7)/((0.65 * 0.7) * (.1 * .15) * (0 * .15))$

○ $(0.65 * 0.7 * 1/3)/((0.65 * 0.7) * (.1 * .15) * (0 * .15))$

iv. (3 pt) We want to get an autograph for ourselves, so we are going to guess which location MJ is at and go to it.

What would we calculate to get the probability of getting an autograph?

Reminder: P(A | B) is the probability of event A given event B

○ P(Autograph | Guess Correctly)

○ P(Autograph | Guess Correctly) + P(Autograph | Guess Incorrectly)

● P(Autograph | Guess Correctly) * P(Guess Correctly) + P(Autograph | Guess Incorrectly) * P(Guess Incorrectly)

○ P(Guess Correctly | Autograph)

○ P(Guess Correctly | Autograph) + P(Guess Incorrectly | Autograph)

4. (17 points) GO BEARS!

GBO (also known as Golden Bear Orientation) is UC Berkeley's orientation for new students (freshman and transfer students). The GBO program claims that the percentage of new students who attend the first GBO event is 80%. This year, Ashwin, a new student, decides that he should go to the first GBO event. He attends, but notices that there are not many new students there. After the event, Ashwin decides to randomly select 500 new students and finds that 320 of them (64%) attended the first GBO event this year. Ashwin thinks that the true percentage of new students who attend the first GBO event is lower than what the GBO program claims. Assume that each student decides to attend the event independently of the other students.

(a) (3 pt) What is a good null and alternative hypothesis that Ashwin could use for his hypothesis test?

○ **Null**: The true percentage of new students who attend the first GBO event is 80%; **Alternative**: The true percentage of new students who attend the first GBO event is not 80%.

● **Null**: The true percentage of new students who attend the first GBO event is 80%; **Alternative**: The true percentage of new students who attend the first GBO event is lower than 80%.

○ **Null**: The true percentage of new students who attend the first GBO event is lower than 80%; **Alternative**: The true percentage of new students who attend the first GBO event is 80%

○ **Null**: The true percentage of new students who attend the first GBO event is not 80%; **Alternative**: The true percentage of new students who attend the first GBO event is 80%

○ None of these

(b) (4 pt) Which of the following test statistics would be a valid test statistic to use for this hypothesis test?

□ TVD

■ percentage of new students who attend in a sample - 80

■ (percentage of new students who attend in a sample - 80) / 2

■ number of new students who attended in a sample

□ abs(percentage of new students who attend in a sample - 80)

□ abs(percentage of new students who attend in a sample - 80) / 2

(c) **(3 pt)** Fill in the following code to simulate the test statistic 10000 times. Assume `calc_test_stat` is a function that takes in a float that represents the percentage of new students that attended the first event in a sample and calculates the test statistic for you based on one of the test statistics you chose above.

```
test_statistics = make array()
for i in np.arange(10000):
    ...
```

```
○ simulated  = sample_proportions(500, make_array(0.8, 0.2)).item(0) * 100
  one_test_stat = calc_test_stat(simulated)
  np.append(test_statistics, one_test_stat)
```

```
○ simulated  = sample_proportions(500, make_array(0.8, 0.2)).item(0) * 500
  one_test_stat = calc_test_stat(simulated)
  np.append(test_statistics, one_test_stat)
```

```
○ simulated  = sample_proportions(500, make_array(0.5, 0.5)).item(1) * 10000
  one_test_stat = calc_test_stat(simulated)
  test_statistics = np.append(test_statistics, one_test_stat)
```

```
● simulated  = sample_proportions(500, make_array(0.2, 0.8)).item(1) * 100
  one_test_stat = calc_test_stat(simulated)
  test_statistics = np.append(test_statistics, one_test_stat)
```

```
○ simulated  = sample_proportions(500, make_array(0.8, 0.2)) * 500
  one_test_stat = calc_test_stat(simulated)
  test_statistics = np.append(test_statistics, one_test_stat)
```

```
○ simulated  = sample_proportions(500, make_array(0.8, 0.2)).item(0) * 500
  one_test_stat = calc_test_stat(simulated)
  test_statistics = np.append(test_statistics, one_test_stat)
```

```
○ one_test_stat = calc_test_stat()
  test_statistics = np.append(test_statistics, one_test_stat)
```

○ None of these

(d) **(4 pt)** Which of the following calculations for p-value is correct? Assume `observed_test_stat` is the observed test statistic that Ashwin observed. Assume that lower values of the test statistic will support the alternative. Select all that apply.

□ `np.count_nonzero(test_statistics >= observed_test_stat) / len(test_statistics)`

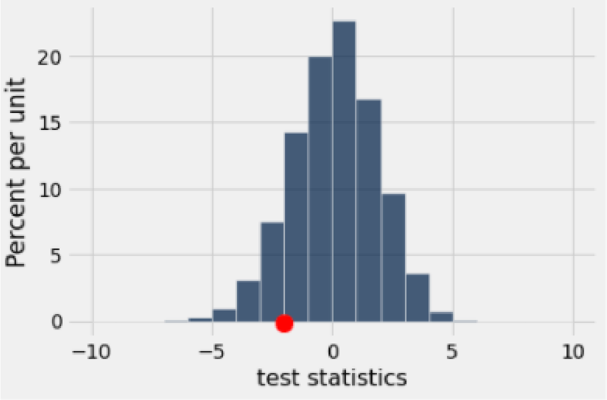
□ `sum(test_statistics >= observed_test_stat)`

■ `np.mean(test_statistics <= observed_test_stat)`

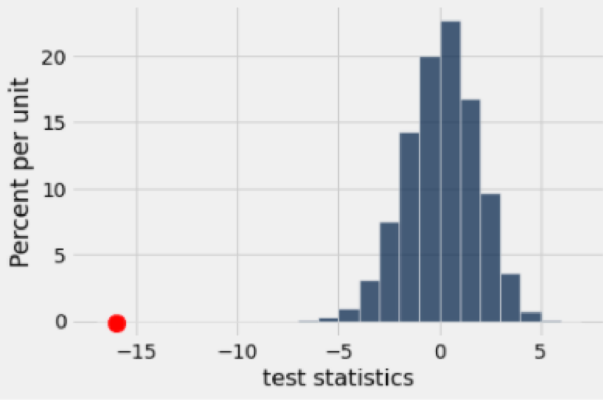
■ `sum(test_statistics <= observed_test_stat) / 10000`

□ None of these

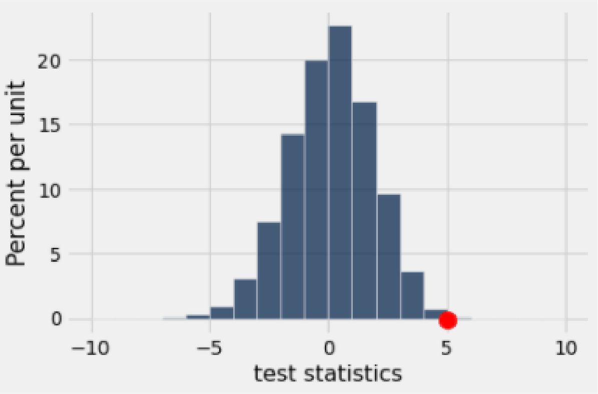
(e) **(3 pt)** Take a look at the following histograms of simulated test statistics below. Instead of using Ashwin's observed test statistic, assume that the red dot on each of the histograms represents an observed test statistic. For which of the following histograms would you reject the null hypothesis with a 5% p-value cutoff? Assume that lower values of the test statistic will support the alternative. Select all that apply.



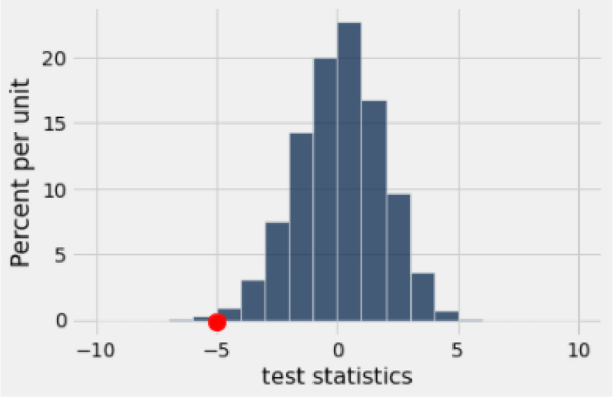
A



B



C



D

□ A

■ B

□ C

■ D

5. (25 points) Data 8 has a pizza my heart

Olivia wants to know the number of times UC Berkeley students ate at Artichoke Basille's Pizza (informally called "Artichoke's") during the fall semester of their freshman year. She takes a random sample of 2500 students and collects their responses. She finds that the average number of times students in the sample ate at Artichoke's during the fall semester of their freshman year was 7, with an SD of 2. For all questions in this section, if there's not enough information to answer the question, select or answer "Not enough information."

(a) i. **(1 pt)** At least what percentage of students in the sample ate Artichoke's between [3, 11] times? Round your answer to the nearest percentage.

● 75%

○ 95%

○ 68%

○ 99.7%

○ 89%

○ 25%

○ Not enough information

ii. **(2 pt)** At least what proportion of students in the sample ate Artichoke's between [4, 10] times?

● 5/9

○ 4/9

○ 3/4

○ 1/4

○ 5/8

○ 3/8

○ Not enough information

iii. **(2 pt)** At most what percentage of students in the sample ate Artichoke's less than or equal to 1 time or greater than or equal to 13 times? Round your answer to the nearest percentage.

● 11%

○ 89%

○ 68%

○ 99.7%

○ 0.03%

○ 75%

○ 25%

○ Not enough information

(b) i. **(2 pt)** Olivia bootstraps her original sample 10,000 times, calculates the average number of times students ate Artichoke's for each resample, and plots a histogram of the results.

What is the approximate standard deviation of the distribution of sample means?

● 2/50

○ 2/2500

○ 1/50

○ 2/10000

○ 2

○ Not enough information

ii. **(1 pt)** If Olivia increased her sample size, the SD of sample means would:

○ increase

● decrease

○ stay the same

○ Not enough information

iii. **(3 pt)** What would a 95% confidence interval for the true population mean look like?

● (7 - 4/50, 7 + 4/50)

○ (7 - 2, 7 + 2)

○ (7 - 2/50, 7 + 2/50)

○ (7 - 4, 7 + 4)

○ (7 - 4/2500, 7 + 4/2500)

○ (7 - 2/2500, 7 + 2/2500)

○ Not enough information

iv. **(2 pt)** Olivia tells her friend Jackie that there's a 95% chance that the true population mean is in her confidence interval. Is Olivia right?

○ No, 95% confidence means that if Olivia were to use her sample to simulate many CIs, about 95% of those CIs would contain the true population mean.

○ Yes, this is what 95% confidence means

○ No, 95% confidence means that 95% of the students went to Artichoke's between (lower bound of the CI, upper bound of the CI) number of times

● None of these

v. **(2 pt)** Suppose Olivia's 95% confidence interval contains the true population mean. If Olivia used the same sample to generate a 99.7% confidence interval *without bootstrapping*, what is the chance that this interval also contains the true population mean?

● 100

○ 0

○ 99.7

○ 95

○ 0.997 * 0.95 * 100

○ Not enough information

vi. **(2 pt)** Olivia wants to create a 68% confidence interval for the true population mean that has a total width of 0.1 or less. In order to achieve this, what sample size should she aim for if she knows the new sample's SD would be 2.5? Feel free to leave your answer unsimplified.

(5/0.1)2 or 502 or 2500

vii. **(3 pt)** If Olivia wanted a 95% confidence interval that also has a total width of 0.1 or less, what is the ratio of her sample size for the 95% confidence interval compared to the sample size for the previous question (the 68% confidence interval)? (For example if x = 50 and y = 500, the ratio of x compared to y would be 50/500 = 1/10)

4

viii. **(1 pt)** If Olivia wanted a narrower confidence interval, she should:

● increase her sample size

○ decrease her sample size

○ not change her sample size since the width of the CI doesn't rely on the sample size

○ Not enough information

ix. **(1 pt)** If Olivia asked 1000 of her closest friends to each take a random sample and calculate a 90% confidence interval of the true population mean, how many of those 1000 intervals would you expect to contain the true population mean?

900

(c) (3 pt) Instead of a confidence interval to find the true mean number of times that students ate Artichoke's, Olivia wants to consider other parameters. Which of the following parameters can Olivia estimate using a 95% confidence interval without having to use simulation? Assume that she is able to take a large random sample from the population and calculate the corresponding statistic. Select all that apply.

□ median number of times that students ate Artichoke's.

□ maximum number of times that students ate Artichoke's.

■ total sum of the number of times that students ate Artichoke's.

■ proportion of times that students ordered the flavor, Margherita Pizza, when they ate Artichoke's (encoded as 1 if they ordered Margherita Pizza in their visit and 0 if they didn't)

■ number of times that students ordered the flavor, Margherita Pizza, when they ate Artichoke's (encoded as 1 if they ordered Margherita Pizza in their visit and 0 if they didn't)

■ average amount of money students spend at Artichoke's each visit

□ none of these

6. (38 points) Data 9

Jessie is teaching a new Data Science class at Berkeley called Data 9 which has over 1000 students. Jessie wants to see if there's any association between different study habits and exam scores. She takes a random sample of 200 Data 9 students, and generates a table with their study habits and their exam scores.

The table below contains one row for each student.

Sleep	Hours studying	Practice questions	Exam score
503	10	21	60
571	20	15	82
433	32	30	78

(197 rows omitted)

The table contains the following columns:

- *Sleep*: (int) the number of minutes of sleep the student got the night before the exam.
- *Hours studying*: (int) the number of hours the student spent studying for the exam.
- *Practice questions*: (int) the number of practice questions the student took prior to the exam.
- *Exam score*: (float) the student's exam score

For all questions in this section, if there's not enough information to answer the question, select or answer "Not enough information."

(a) **(1 pt)** What plot could you use to see if there is a linear relationship between the *Hours studying* column and the *Exam score* column?

● scatter plot

○ line plot

○ bar plot

○ overlaid histogram

Jessie visualizes the *Hours studying* and the *Exam score* columns and sees that there is a positive linear relationship between the two variables. Use the following information to complete the next 6 questions below.

(b)
- The *Exam score* column has a mean of 60 and a standard deviation of 10.
- The *Hours studying* column has a mean of 20 and a standard deviation of 5.
- The correlation between the *Exam score* and *Hours studying* columns is 0.75.

i. **(1 pt)** Jessie wants to predict a student's exam score from their number of hours studying. What is the slope of the regression line?

● 1.5

○ 0.75

○ 0.375

○ 0.25

○ 0.5

○ 2

○ Not enough information

ii. (2 pt) What exam score would Jessie predict if a student studied 30 hours?

● 75

○ 60

○ 45

○ 80

○ 70

○ 66

○ Not enough information

iii. (2 pt) Jessie decides it would be best to convert the variables into standard units first before finding the regression line. If Jessie predicts a student's exam score in standard units from their number of hours studying in standard units, what would the intercept be for the regression line?

● 0

○ -3

○ 2

○ 30

○ Not enough information

Jessie is looking at 4 different regression lines.

iv.

- Line A: predict the *Exam score* from *Hours studying* (predict Y from X)
- Line B: predict the *Hours studying* from *Exam score* (predict X from Y)
- Line C: predict the *Exam score* in standard units from *Hours studying* in standard units (predict Y in SU from X in SU)
- Line D: predict the *Hours studying* in standard units from *Exam score* in standard units (predict X in SU from Y in SU)

A. (2 pt) Fill in the blank. The slope Line A would be ________ the slope of Line B in magnitude.

● greater than

○ less than

○ equal to

○ Not enough information to fill in the blank

B. (2 pt) Fill in the blank. The slope of Line C would be ________ the slope of Line D in magnitude.

○ greater than

○ less than

● equal to

○ Not enough information to fill in the blank

v. **(2 pt)** Fill in the blank. Suppose the correlation coefficient between the two variables was actually 0.8. The standard deviation of the residuals would _______ given the adjusted value for r?

○ increase

● decrease

○ stay the same

○ Not enough information to fill in the blank

vi. **(2 pt)** Jessie thought that linear regression was a good idea in this case. However, when she looked at the residual plot, it didn't look like a formless blob. What is the mean of the residuals?

● 0

○ 60

○ 20

○ Not enough information

vii. **(3 pt)** Now, Jessie is predicting *Exam score* from *Sleep* and has the following information about the regression line and residuals. What is the SD of the residuals? Hint: remember the formula for how you can find the standard deviation of a set of values.

r	r**2	SD of Sleep	RMSE
0.7	0.49	452	5.9

○ 452

○ (1-0.49) * 452

● 5.9

○ np.sqrt(1-0.49) * 452

○ 0.49

○ Not enough information

viii. **(2 pt)** Instead of using the RMSE, Raymond tells Jessie to use a new error function, the SMRE (squared median root error). Fill in the blank with a single line of code for this new error function.

```
def smre(actual, fitted):
    return _____________
```

np.median((actual -fitted)2) 0.5

(c) Jessie's fellow Data 9 GSI Rita suspects that there is an linear association between the *Sleep* column and the *Exam score* column.

i. (2 pt) Provide a null hypothesis Rita could use to test her suspicions.

The true correlation between *Sleep* and *Exam score* is 0.

ii. (2 pt) Provide an alternative hypothesis Rita could use to test her suspicions.

The true correlation between *Sleep* and *Exam score* is not 0.

iii. (3 pt) Which of the following statements describes a valid way for Rita to test her null hypothesis.

○ Generate a 95% confidence interval for the true slope of the regression line. If 0 is in the confidence interval, reject the null hypothesis at the 5% p-value cutoff.

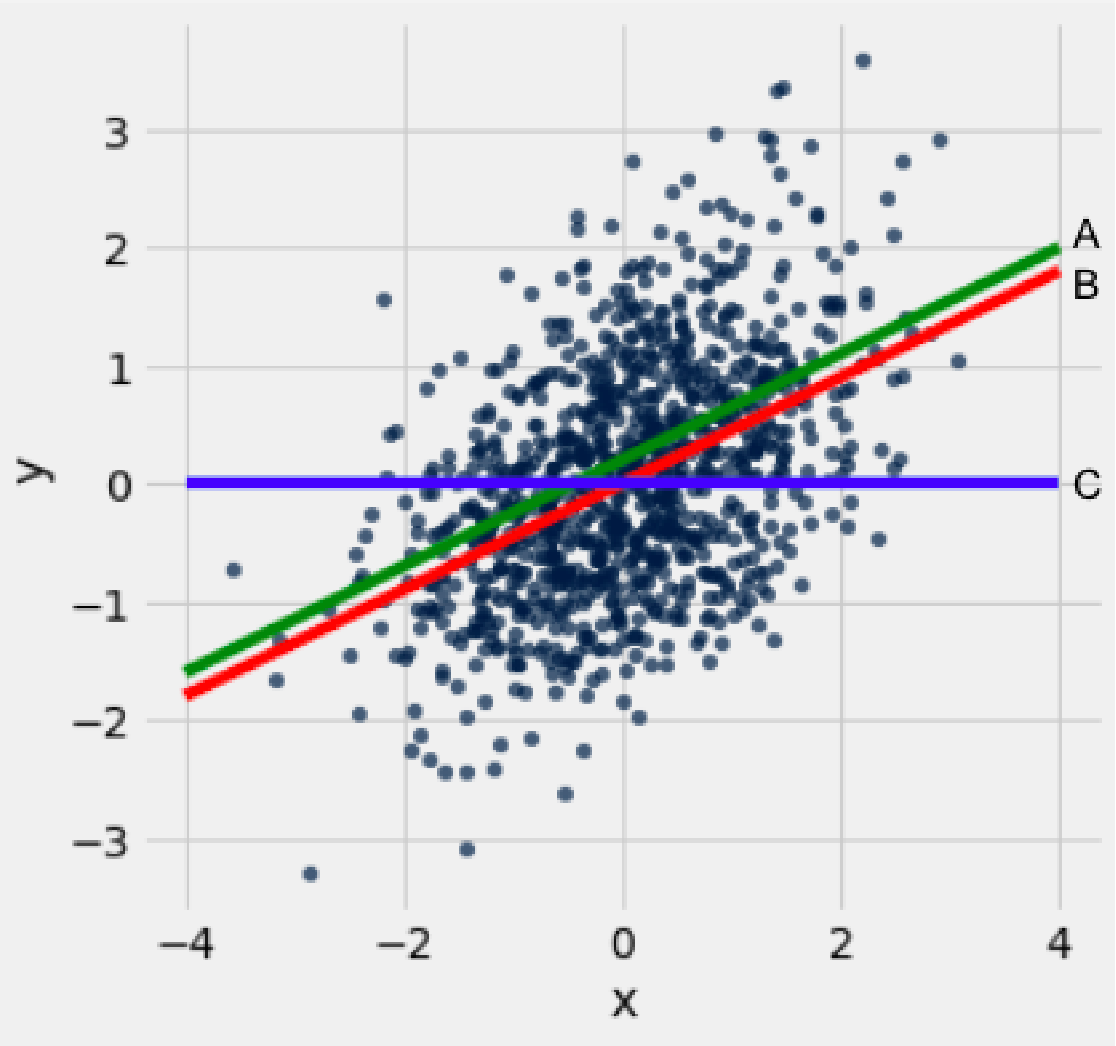
● Generate a 90% confidence interval for the true correlation between *Sleep* and *Exam score*. If 0 is not in the confidence interval, reject the null hypothesis at the 10% p-value cutoff.

○ Generate a 99% confidence interval for the true correlation between *Sleep* and *Exam score*. If 1 is not in the confidence interval, reject the null hypothesis at the 1% p-value cutoff.

○ Generate a 99% confidence interval for the true slope of the regression line. If 1 is in the confidence interval, fail to reject the null hypothesis at the 1% p-value cutoff.

○ None of the these

(d) (2 pt) Rank the following lines from the smallest MSE to the largest MSE. Note that line B is the regression line.



○ A, C, B

○ A, B, C

○ B, C, A,

● B, A, C

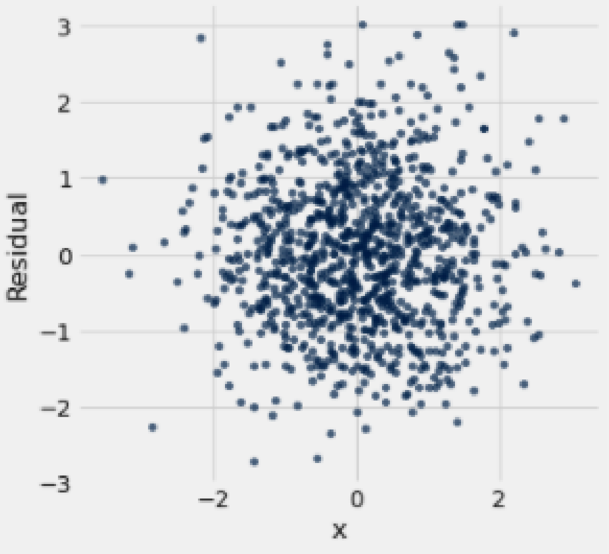
○ C, A, B

○ C, B, A

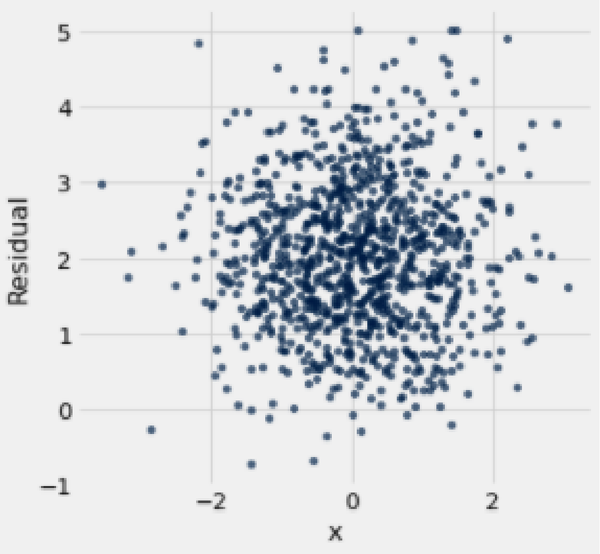
○ None of these

○ Not enough information

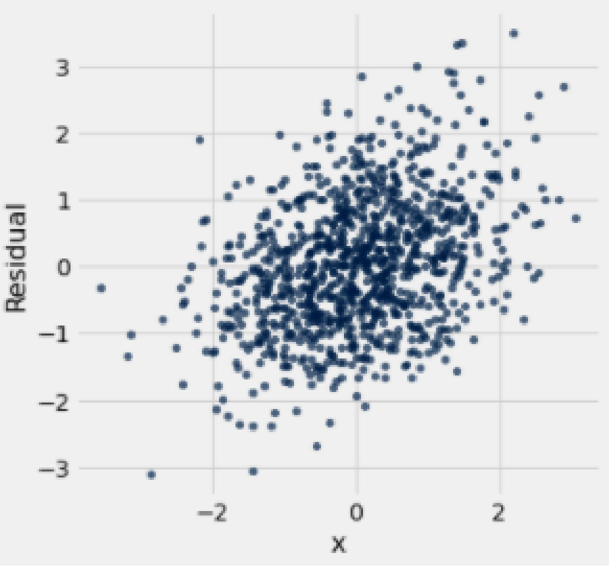
(e) **(3 pt)** Which of the following statements about the 4 plots below are true? Select all that apply.



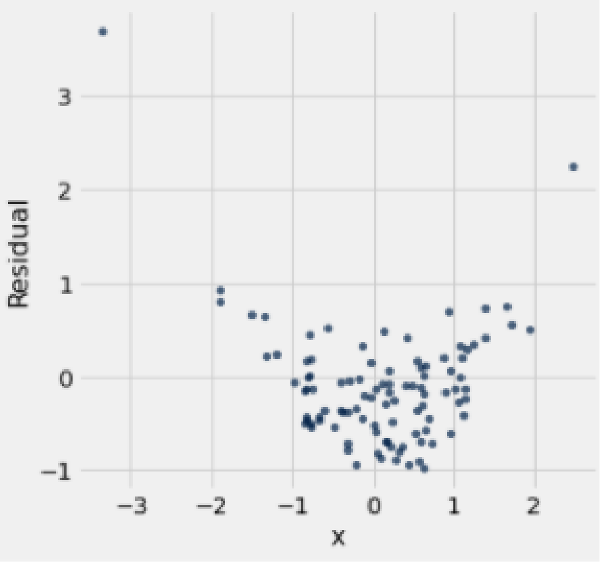
A



B



C



D

■ Plot A indicates that linear regression is a good idea

□ Plot B indicates that linear regression is a good idea

□ Plot C indicates that linear regression is a good idea

□ Plot D indicates that linear regression is a good idea

□ Plot A is an impossible residual plot

■ Plot B is an impossible residual plot

■ Plot C is an impossible residual plot

(f) (4 pt) Jessie finds that there is a strong positive correlation between *Practice questions* and *Exam score* and wants to predict *Exam score* from *Practice questions* using linear regression. The mean of *Practice questions* is 25. Which of the following statements are true? Select all that apply.

■ The 95% prediction interval at **Practice questions** = **27** is narrower than the 95% prediction interval at **Practice questions** = **37**

□ The 95% prediction interval at **Practice questions** = **25** is more variable than the 95% prediction interval at **Practice questions** = **10**

■ If *Practice questions* and *Exam score* are both in standard units, the 95% prediction interval at **Practice questions SU** = **1** would have the same theoretical bounds as the 95% CI for the true correlation

□ If *Practice questions* and *Exam score* are both in standard units, the 95% prediction interval at **Practice questions SU** = **0** would have the same theoretical bounds as the 95% CI for the true slope of the regression line

■ If *Practice questions* and *Exam score* are both in standard units, the 95% prediction interval at **Practice questions SU** = **0** would have the same theoretical bounds as the 95% CI for the true intercept of the regression line in standard units

□ If we generated a 95% prediction interval at **Practice questions** = **30** to be (60, 80), this means that there's a 95% chance that the true prediction for *Exam score* at **Practice questions** = **30** falls within (60, 80).

■ If *Practice questions* and *Exam score* are both in standard units, the 95% prediction interval at **Practice questions SU**= **0** will have a width very close to 0

□ None of these are true because it's impossible to calculate a prediction interval since the CLT doesn't apply

(g) (3 pt) Jessie finds that there is a strong positive correlation between *Practice questions* and *Exam score* and wants to predict *Exam score* from *Practice questions* using linear regression. Which of the following statements are true? Select all that apply.

■ If our regression model is correct then *Exam score* will deviate from its mean less than or equal to the amount that *Practice questions* deviates from its mean when both variables are in standard units

□ Our *Exam score* value will deviate from its mean less than or equal to the amount that *Practice questions* deviates from its mean when both variables are in standard units

□ If we see many *Exam score* values that deviate further from its mean than their *Practice questions* values deviate from its mean, we should expect future *Exam score* values we collect to regress to the mean of *Exam score*

7. (0 points) Last Words

(a) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question (including the title of the section, e.g., Be Like Mike) and state your assumptions. Be warned: We only plan to consider this information if we agree that the question was erroneous or ambiguous and we consider your assumption reasonable.

8. (0 points) Just for fun

(a) (0 pt) Choose one:

● Yanay

○ Katherine

(b) (0 pt) Please help us name the two bears from our lecture art.

No more questions.