Data C100/200- Final

Summer 2025

Name: ______________________________

Email: ______________________________@berkeley.edu

Student ID: ______________________________

Name and SID of the person on your left: ______________________________

Name and SID of the person on your right: ______________________________

Exam Room: ______________________________ Seat Number: ______________________________

Instructions:

This exam consists of **58 points** spread out over **4 questions** and the **Honor Code certification**. The exam must be completed in **170 minutes** unless you have accommodations supported by a DSP letter. Note that you should:

- Each true/false question and multiple choice question has **exactly one** correct answer. Please **fully** shade in the circle to mark your answer.
- Blank answers and incorrect answers are graded identically, so it's in your best interest to answer every question.
- For all math questions, **please simplify your answer**. Please also **show your work** if a large box is provided.
- For all coding questions, you may use commas and/or one or more function calls in each blank.
- **You MUST write your Student ID number at the top of each page.**
- You should not use a calculator, scratch paper, or notes you own other than the reference sheets distributed at the beginning of the exam.

For all Python questions, you may assume `Pandas` has been imported as `pd`, `NumPy` as `np`, the Python RegEx library as `re`, `matplotlib.pyplot` as `plt`, and `seaborn` as `sns`.

Honor Code [1 Pt]:

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: ______________________________

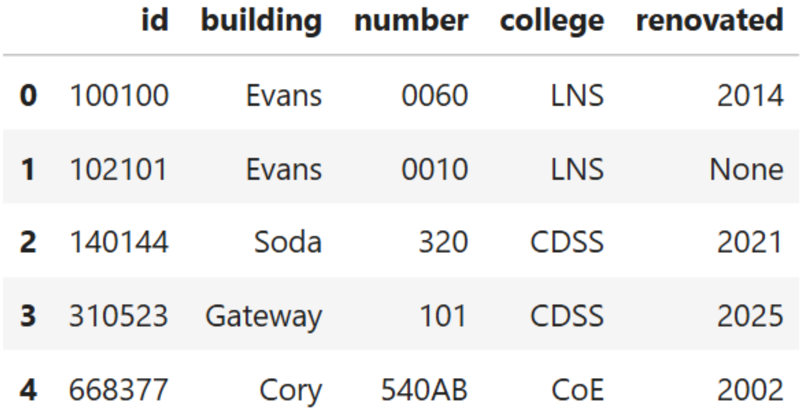
This exam was administered on a computer via PrairieLearn. The PrairieLearn software allows exam questions to differ across students. This PDF shows one possible instantiation of the exam. If you took this exam in Summer 2025, the questions you saw at the testing center may have been slightly different than the questions in this PDF.

1 Making room for pandas [17 Pts]

Josh samples information about various rooms in different buildings around Berkeley in preparation for the upcoming semester. He stores the information in a `DataFrame` called `rooms`. The columns of `rooms` are described below:

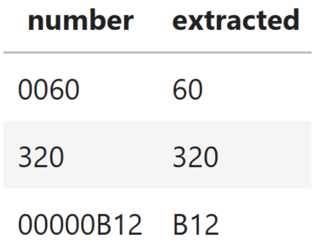
- `id`: A unique room id from Berkeley's internal database (type = `np.int64`).
- `building`: Name of the building the room is in (type = `str`).
- `number`: The room number listed on the door to the room (type = `str`).
- `college`: The Berkeley college each building is associated with from the select options of: `'CoE'`, `'CDSS'`, `'CNR'`, and `'LNS'` (type = `str`).
- `renovated`: The year the room was last renovated (type = `str`).
  Note: Certain rooms may never have been renovated, in which case `renovated` will contain a value of `None`.

The first five rows of `rooms` are shown below:



	id	building	number	college	renovated
0	100100	Evans	0060	LNS	2014
1	102101	Evans	0010	LNS	None
2	140144	Soda	320	CDSS	2021
3	310523	Gateway	101	CDSS	2025
4	668377	Cory	540AB	CoE	2002

(a) [1.5 Pts] Some of values in the `number` column have leading zeroes (e.g., "06A"). Write a regular expression to remove all leading zeroes from the entries in the `number` column. Every string in `number` starts with a non-negative count of leading zeroes, followed by a combination of numbers and/or letters. No room number consists of only zeroes. The table below provides test cases and expected outputs for your code.



number	extracted
0060	60
320	320
00000B12	B12

Fill in the `regex` pattern below to accomplish this.

```
rooms['number'] = rooms['number'].str.extract(r'_____(i)_____')
```

(i) [1.5 Pts] Fill in blank `(i)`:

(b) [6 Pts] Josh wants to determine which colleges have the most modern buildings. b A building is considered modern if at least 80% of its rooms in `rooms` with recorded renovations were renovated in the year 2003 or later.

Fill in the blanks below to return a `Series` with the index as the `college` and the values as the number of its buildings that are modern.

```
# Step 1:  Remove rows that have a None value in `renovated`.

bi = _____(i)_____

# Step 2:  Remove rows associated with buildings that are not
   modern.

bii = bi.groupby("building").filter(_____(ii)_____)

# Step 3:  Count the number of modern buildings in each
   college.  Return this information as a Series.

biii = bii.groupby(["college", "building"]).size()._____(iii)
   _____
```

(i) [1 Pt] Fill in blank `(i)`:

(ii) [2 Pts] Fill in blank `(ii)`.

(iii) [2 Pts] Fill in blank `(iii)`:

(c) [2 Pts] Josh created `rooms` by repeating the following process 100 times:

1. Select one Berkeley college, where each Berkeley college has the same probability of being selected.
2. Select one building from the chosen college, where each building in the college has the same probability of being selected. *Note: Not all colleges have the same # of buildings.*
3. Select one room from the chosen building, where each room in the building has the same probability of being selected. Store this room in `rooms`. *Note: Not all buildings have the same # of rooms.*

This process results in the 100 selected rooms in `room`. Answer the true/false questions below.

○ True ○ False This is a probability sample.

○ True ○ False Every room at Berkeley has the same probability of being included in the sample.

○ True ○ False This sampling scheme is equivalent to a simple random sample of all rooms at Berkeley.

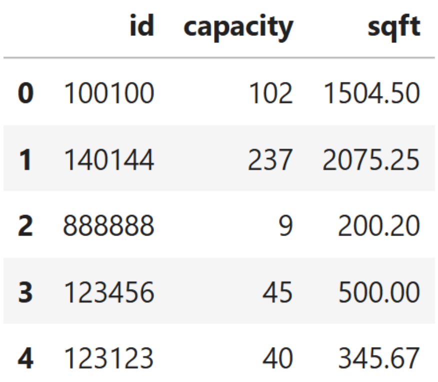
○ True ○ False This sampling scheme is equivalent to a stratified random sample of all rooms at Berkeley, where the strata are defined by every combination of building and college.

(d) [3 Pts] Michael sends Josh a `DataFrame` called `details` that contains more detailed information about the rooms in `rooms`. Unfortunately, during the data transfer process, some rooms were dropped, so `details` only contains a subset of the rooms in `rooms`.

The columns of `details` are described below:

- `id`: A unique room id from Berkeley's internal database (type = `np.int64`).
- `capacity`: The room's maximum capacity (type = `np.int64`).
- `sqft`: The room's square feet (type = `np.float`).

The first five rows of `details` are shown below:



	id	capacity	sqft
0	100100	102	1504.50
1	140144	237	2075.25
2	888888	9	200.20
3	123456	45	500.00
4	123123	40	345.67

Help Josh join `rooms` and `details` together to create a `SQL` table `detailed_rooms`.

1. Josh does not want to drop any rows of `rooms` in the merge, even if `details` doesn't contain a match.
2. `detailed_rooms` should contain a new column called `size` of the `VARCHAR` (i.e., string) type. The `size` of a room is defined as `'unknown'` if the room has no recorded `capacity`, `'small'` if the `capacity` is less than 30, `'medium'` if the `capacity` is between 30 and 99 (inclusive), and `'large'` otherwise.

Josh does not want to drop any rows of `rooms` in the merge, even if `details` doesn't contain a match.

Fill in the blanks below. Assume that `duckdb` is imported and `rooms` and `details` can be queried as SQL tables. Treat the `None` type in Python as `NULL` in SQL. Treat `DataFrame` names as SQL table names in your query.

```
SELECT r.id, _____(i)_____,
CASE
    _______________
    _______________
    _____(ii)_____
    _______________
    _______________
FROM rooms AS r
_____(iii)_____ JOIN details AS d
    ON _____(iv)_____;
```

(i) [0.5 Pts] Fill in blank `(i)`:

(ii) [1 Pt] Fill in blank `(ii)`. Use as many or as few lines as needed.

(iii) [1 Pt] Fill in blank `(iii)`:

(iv) [0.5 Pts] Fill in blank `(iv)`:

(e) [2 Pts] For this subpart, you may assume the table detailed_rooms from Question 1d has been successfully created. Josh wants to determine which buildings he is most likely to lecture in. Write a query that results in a table with the following characteristics:

- Only includes buildings belonging to the 'CDSS' and 'LNS' colleges, and whose recorded rooms have an average sqft larger than 1000.
- Counts the total number of rooms in each building that are of size 'large'.
- Includes only two columns: the building name and the number of 'large' rooms. The column with the number of 'large' rooms should be called total.
- Ordered by total in descending order.

Complete the SQL query below to create such a table.

```
SELECT _____(i)_____
FROM detailed_rooms
WHERE _____(ii)_____
GROUP BY _____(iii)_____
HAVING _____(iv)_____
ORDER BY _____(v)_____;
```

(i) [0.5 Pts] Fill in blank (i):

(ii) [1.5 Pts] Fill in blank (ii):

(iii) [0.5 Pts] Fill in blank (iii):

(iv) [0.5 Pts] Fill in blank (iv):

(v) [0.5 Pts] Fill in blank `(v)`:

(f) [2 Pts] Josh wants to visualize the data in `detailed_rooms`. Answer the questions below.

(i) [1 Pt] Which of the following visualizations are appropriate to visualize the total number of room renovations at Berkeley by year? Assume the number of room renovations is on the y-axis, and the year is on the x-axis. You may assume there are no issues with overplotting the years in the x-axis.

□ Line plot

□ Bar plot

□ KDE Plot

□ Side-by-side box plots

(ii) [1 Pt] Josh creates a KDE plot of the maximum number of occupants of each room at Berkeley using the normal kernel. He finds that his KDE plot looks just like a normal distribution, even though he knows that the distribution of room capacities is multimodal. Which of the following choices could be the reason for this issue?

□ His bandwidth parameter is too large.

□ His bandwidth parameter is too small.

□ He didn't normalize the sum of the kernels.

□ The normal kernel is not appropriate for visualizing multimodal data.

2 A model enrollee in Data 100 [16 Pts]

CDSS wants to predict the number of students who will enroll in its courses. The CDSS course coordinators request Data 100 staff to build a variety of linear models to assist their prediction.

(a) [2 Pts] Jake receives enrollment numbers from various past courses. He is interested in creating a constant model ($\hat{y} = \theta_0$) to make predictions. Jake wants to decide between optimizing for MSE vs MAE. Select all statements below that are true.

□ Jake can use gradient descent to find a value of θ_0 that minimizes MAE.

□ It is possible for two different constant predictions to both result in the optimal MAE.

□ Since the MSE loss surface is always convex and smooth, there is always one optimal $\hat{\theta}_0$ that minimizes it.

□ Suppose Jake chooses to optimize the MSE. The median of the data can result in a lower training loss than the training loss associated with the mean.

(b) [3 Pts] Wesley has $n = 30$ unique data points and 2 features. He fits the following OLS model:

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 (x_{i1})^3 + \hat{\theta}_2 \ln x_{i2}$$

Wesley correctly calculates a **unique optimal solution** for his OLS model. Select all statements about Wesley's fitted model (which uses the optimal parameters) that must be true:

□ It is highly likely, but not guaranteed, that the sum of the residuals is 0.

□ The rank of the design matrix ($\mathbb{X}$) must be less than 3.

□ The vector of outcomes ($\mathbb{Y}$) must be in the span of $\mathbb{X}$.

□ The vector of predictions ($\hat{\mathbb{Y}}$) must be a linear combination of the outcomes ($\mathbb{Y}$).

□ If Wesley refit the same model with $(x_1)^6$ as an additional feature, a unique solution will no longer exist.

□ If Wesley refit the same model using L2 (Ridge) regularization and $\lambda = 1$, a unique solution will still exist.

(c) [5 Pts] Sammie wants to fit and assess a couple models.

(i) [2 Pts] First, Sammie wants to find the optimal solution for θ_0 with respect to the objective function below.

$$L(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left(\frac{\theta_0}{4\theta_1} - \theta_1 x_i \right) \right)^2$$

Find the value of θ_0 that minimizes $L(\theta_0, \theta_1)$.

(ii) [3 Pts] Sammie chooses a new constant model specification and objective function. She calculates the gradient of her new objective function below with respect to θ_0:

$$L(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta_0)^2$$

$$\nabla L(\theta_0) = \left[-\frac{2}{n} \sum_{i=1}^{n} y_i - \theta_0 \right]$$

Sammie asserts that $\theta_0^{(1)}$ = 5, $\alpha = 1$, and $\vec{y} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Using the values above, calculate $\theta_0^{(0)}$ (i.e., the initializing θ_0 value).

(d) [6 Pts] Milena chooses a new model specification and decides to fit her chosen model with regularization.

(i) [2 Pts] Select all statements below that are true.

- □ For a fixed model specification, training dataset, and validation dataset, the optimal choice of λ hyperparameter is always the same for L1 and L2 regularization.
- □ L1 regularization typically increases model bias.
- □ L2 regularization typically increases model variance.
- □ Unlike L2 regularization, L1 regression can be used to identify the most predictive features. But, L1 regularization will generally result in higher average loss.

(ii) [1 Pt] Milena chooses a special form of regularization that takes in 2 hyperparameters, λ and ρ. To pick the values for these parameters, she decides to use 7-fold cross validation. Milena tests 2 values of λ and 5 values of ρ. To estimate the optimal λ value and ρ value, how many times is each data point used to fit a model?

(iii) [2 Pt] Milena uses gradient descent to find the optimal parameters. Select all statements below that are true..

- □ With an **infinite** number of updates and a **fixed** learning rate, gradient descent will always converge to a **global** minimum.
- □ With an **infinite** number of updates and a **fixed** learning rate, gradient descent will always converge to a **local** minimum.

□ The gradient descent algorithm can compute gradient updates using both convex and non-convex functions.

□ Consider the average of the estimated gradients computed over one epoch of stochastic gradient descent. This value is always equal to the true gradient of the loss surface at the initial values of the model parameters.

(iv) [1 Pt] Milena decides to use mini-batch gradient descent on a dataset with 500 data points. She records that a total of 125 gradients were calculated after 5-epochs. How many data points were in each mini-batch?

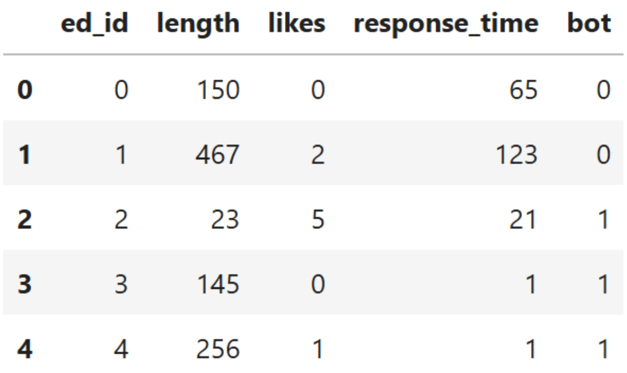
3 Edison takeover [15.5 Pts]

The DataFrame ed_replies is created for an Ed reply predictor.

The columns of ed_replies are described below:

- ed_id: The Ed reply's unique id (type = np.int64).
- length: The number of characters in the reply (type = np.int64).
- likes: The number of likes the reply received (type = np.int64).
- response_time: The number of minutes between the time of the reply and the time of its original post (type = np.int64).
- bot: Indicates if the reply was made by an Ed bot. 0 for a human and 1 for an Ed bot (type = np.int64).

The five rows of ed_replies are shown below:



	ed_id	length	likes	response_time	bot
0	0	150	0	65	0
1	1	467	2	123	0
2	2	23	5	21	1
3	3	145	0	1	1
4	4	256	1	1	1

(a) [2 Pts] Justin wants to fit a logistic regression model to predict whether a particular Ed reply is from a bot or a real person. Select all statements below that are true.

□ Let $\mathbb{X}$ be the design matrix. The quantity $\mathbb{X}^\top\mathbb{X}$ must be invertible in order to compute the optimal parameter vector.

□ It is possible for the sigmoid function σ to output exactly $P(y_i = 1|x_i) = 1$ for some data point x_i with real-valued features.

□ The model predictions on the probability scale are a nonlinear function of inputted features.

□ If the training dataset is linearly separable, it is impossible for gradient descent to converge on a unique optimal $\hat{\vec{\theta}}$.

(b) [1.5 Pts] Justin fits the following logistic regression model to predict the probability that a reply was made by an Ed bot:

$$P(y = 1|\vec{x}) = \sigma(\hat{\theta}_1 \times \texttt{likes} + \hat{\theta}_2 \times \texttt{response\_time});\ \hat{\theta}_1 = 1,\ \hat{\theta}_2 = 2$$

Based on the fitted model above, what are the estimated **odds** that a reply with `likes = 3` and `response_time = 34` was made by a **human**? *If preferred, feel free to use σ in your solution. You do not need to simplify algebra.*

(c) [3.5 Pts] Justin fits a different classification model on the training data. He applies this model to a test set and gets the following results.

True Label	1	1	1	0	0
Prediction	1	0	1	0	0
Outputted probability (y = 1)	0.88	0.4	0.75	0.3	0.54

(i) [1 Pts] Given the values in the table above, what are the minimum and maximum possible values of Justin's chosen threshold? *For full credit, your answer does not need to address whether the bounds are inclusive or exclusive.*

(ii) [1 Pts] Calculate the recall of the model on the test set. *You do not need to simplify algebra.*

(iii) [1.5 Pts] Select all statements below that are true.

○ True ○ False After reducing the classification threshold from 0.9 to 0.1, it is possible for the true positive rate to increase.

○ True ○ False It is possible for the F_1 score of a classifier to change after changing the threshold.

○ True ○ False It is possible for AUC-ROC to change after changing the threshold.

(d) [6 Pts] Justin would like to perform PCA analysis on the `ed_replies` dataset. He asks Hannah to provide him the necessary data. She first subsets the `DataFrame` to only the features using the following code:

```
ed_subset = ed_replies[['length', 'likes','response_time']]
```

Then, Hannah randomly samples a number of rows of `ed_subset` and converts the resulting `DataFrame` into a matrix denoted by X. Hannah performs SVD on X. Recall that in SVD, X is decomposed into the product of 3 matrices, U, S and $V^\top$ (using the same naming convention as lecture). Hannah only sends Justin the complete matrix S (the singular values).

$$S = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 88 & 0 \\ 0 & 0 & 8 \end{bmatrix},$$

(i) [1.5 Pts] Select all statements below that are true. *Note: Each option is independent of the other options.*

○ True ○ False If Justin also received the first column of U, he could compute the values of the first latent feature for all observations in the sample.

○ True ○ False If Justin also received X and the first two rows of V, he could compute the values of the first two latent feature for all observations in the sample.

○ True ○ False If Justin received the entire SVD of X and utilized 3 principal components to reconstruct a latent vector representation of X, the resultant matrix would have 0 reconstruction loss on the sample data.

(ii) [1 Pt] Justin now wants to produce a scree plot using S. Calculate the proportion of variance captured by the third principal component. *You do not need to simplify algebra.*

(iii) [2 Pts] Hannah sends four mystery vectors, v_1, v_2, v_3, and v_4 to Justin, two of which are columns of V. She also sends over two tables containing information about the vectors.

In the first table, she reports the L2-norm of each vector.

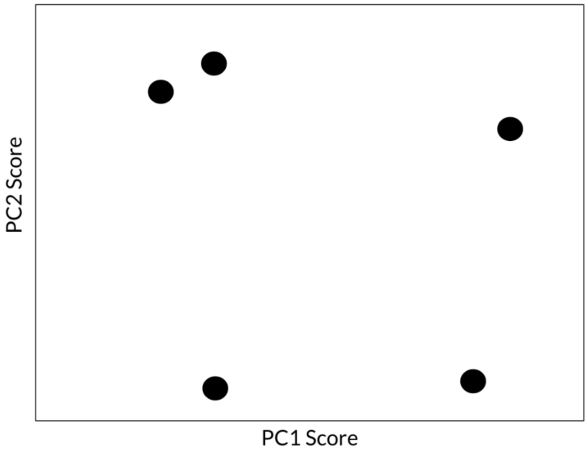
Vector	v_1	v_2	v_3	v_4
Norm	1	1	2	1

In the second table, she reports the pairwise dot product between the vectors.

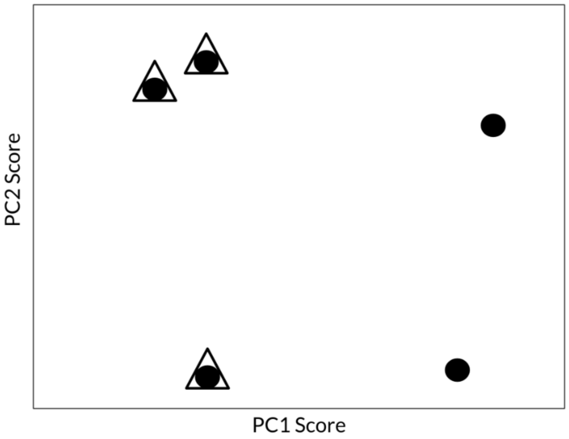
Vector Pairing	(v_1, v_2)	(v_1, v_3)	(v_1, v_4)	(v_2, v_3)	(v_2, v_4)	(v_3, v_4)
Dot Product	1	0	0	0	8	100

Which two of the four vectors are the columns of V? Additionally, justify your answer. Correct answers with missing or fully incorrect justifications will receive no credit.

(e) [2 Pts] Justin uses a sampling of the points from `ed_replies` and projects the points onto the first two principal components, where there are five data points represented by dots.

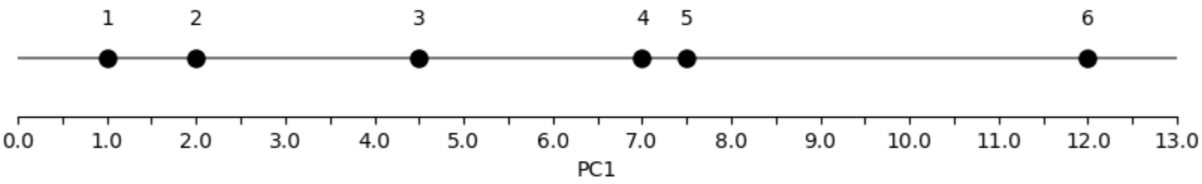


Justin instantiates K-Means clustering with K=3, where a triangle shape indicates a centroid. Note that the K-Means algorithm is run using the two principal components, not the original features.



Using the frames on your scratch paper, illustrate the process of running the K-means algorithm until the cluster centers (i.e., centroids) no longer change. You are not required to use all the frames.

(f) [1 Pt] Justin projects a different sampling of points onto the first principal component.



Provide the state of the clusters after 3 steps of agglomerative clustering using **complete linkage**. Use curly braces to denote the clusters. For example, if all points except points `5` and `6` are in one cluster, with `5` and `6` in their own cluster, respond with `{1, 2, 3, 4}{5,6}`. *Note: The clustering algorithm is run using the first principal component, not the original features.*

(g) [1 Pt] Which of the following statements are true?

○ True ○ False A negative silhouette suggests that a point is, on average, closer to the points in the nearest cluster compared to the points in its own cluster.

○ True ○ False Agglomerative clustering will always converge to the same solution if re-run with the same parameters. *For this option, you can assume there are never exact ties in distances between clusters.*

4 Diet P(EPSI) [9.5 Pts]

A team of researchers is studying the consumption of sugary drinks of adults living in Berkeley. In the study, the team surveys 1000 adults from Berkeley at random **with replacement** and collects the follow data.

- `drink_soda`: Whether the surveyed adult consumes at least 1 soda every day. The value is `True` if the surveyed adult consumes at least 1 soda every day, and `False` otherwise. (type: boolean)

The team decides to estimate the proportion of adults living in Berkeley that consume at least 1 soda every day. In the problems below, you can assume that the proportion of adults living in Berkeley that consume at least 1 soda every day is p, a fixed but unknown proportion strictly between 0 and 1.

Note: In probability notation, the `drink_soda` data collected can be viewed as i.i.d. random variables $X_1, X_2, \dots, X_{1000}$, where each $X_i \sim \text{Bernoulli}(p)$.

(a) [3 Pts] The team uses an estimator $\hat{\alpha}$ that is just the sample proportion. That is,

$$\hat{\alpha} = \frac{1}{1000} \sum_{i=1}^{1000} X_i$$

Find Bias($\hat{\alpha}$) and Var($\hat{\alpha}$). You answer can be in terms of p.

The team then uses the bootstrap to analyze the estimator $\hat{\alpha}$. Recall that "a bootstrap sample" under this context means "a sample of size 1000 drawn uniformly at random with replacement from the original sample". Suppose Michael appears once in the original sample.

(b) [1 Pt] The team first draws one bootstrap sample. Find the probability that Michael is drawn as the first individual of the bootstrap sample.

○ 0
○ $\frac{1}{1000}$
○ $\frac{1}{2}$
○ $\frac{999}{1000}$
○ 1

(c) [1 Pt] Let m be the correct answer to (d). Find the probability that Michael does not appear in the bootstrap sample at all. Your answer can be in terms of m.

○ 0

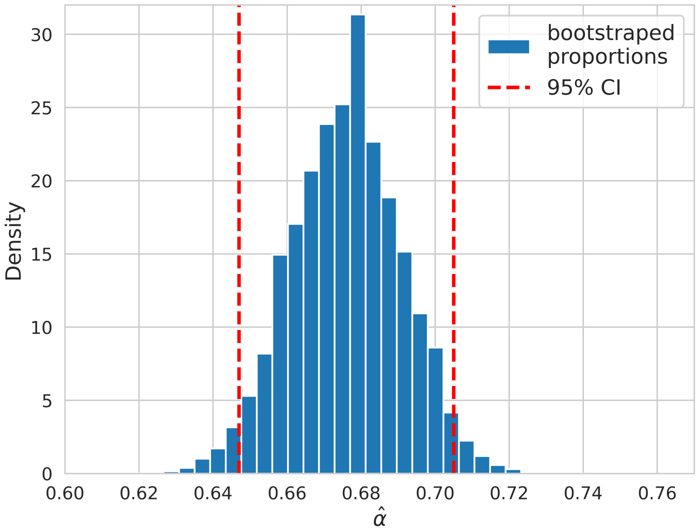
○ m^{1000}

○ $(1-m)^{1000}$

○ $\left(\frac{m}{2}\right)^{1000}$

○ 1

The team then draws 100,000 bootstrap samples and computes $\hat{\alpha}$ on each sample. The sample proportions are plotted in the histogram below. The two dashed lines represent the lower and upper bounds of the 95% confidence interval for p computed on the bootstrap samples.



(d) [1.5 Pts] Suppose $\hat{\alpha} = 0.67$ for the original sample. Using this information and the histogram, which of the following **must be true**?

○ True ○ False This histogram is an approximation of the distribution of X_1 (or any one of $X_1, \ldots, X_{1000}$ since they are i.i.d.).

○ True ○ False The two-sided null hypothesis "the proportion of adults in Berkeley that consume at least 1 soda everyday is 72%" is rejected at the 5% significance level.

○ True ○ False The two-sided null hypothesis "the proportion of adults in Berkeley that consume at least 1 soda everyday is 72%" is rejected at the 10% significance level.

After this background study, the team goes on to explore the relationship between consumption of sugary drinks and blood sugar level. When the team collected the data on `drink_soda`, it also collected the following data on the same individuals.

- `blood_sugar`: The pre-breakfast blood sugar level measured by the research team in the unit of millimoles per liter. (type: `np.float`)
- `t2d`: An indicator of whether the individual has type 2 diabetes, where a value of 1 indicates type 2 diabetes, and 0 indicates no type 2 diabetes. (type: `int`)

To study the relationship, the team fit the following linear regression model over the data

$$\widehat{\text{blood\_sugar}} = \theta_0 + \theta_1 \times \text{drink\_soda} + \theta_2 \times \text{t2d}$$

Suppose the optimal estimated parameters are $\hat{\theta}_0 = 5.1$, $\hat{\theta}_1 = 2.8$, and $\hat{\theta}_2 = 3.8$.

(f) [1 Pt] What is the interpretation of $\hat{\theta}_0$? Explain in one sentence.

(g) [1 Pt] What is the interpretation of $\hat{\theta}_1$? Explain in one sentence.

(h) [1 Pt] What is the interpretation of $\hat{\theta}_2$? Explain in one sentence.

You are done with the final- Congratulations!

Draw your favorite DATA 100/200 memory!