Data 100 & 200A Principles and Techniques of Data Science

Spring 2019

Final Solutions

INSTRUCTIONS

- You have 2 hours and 50 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except for the provided midterm 1 reference sheet and up to three 8.5" × 11" sheets of notes of your own creation.
- Mark your answers **on the exam itself**. We will *not* grade answers written on scratch paper.

Last name	
First name	
Student ID number	
CalCentral email (`_@berkeley.edu`)	
Exam room	
Name of the person to your left	
Name of the person to your right	
All the work on this exam is my own. **(please sign)**	

Terminology and Notation Reference:

$\exp(x)$	e^x
$\log(x)$	$\log_e x$
Linear regression model	$E[Y\|X] = X^T\beta$
Logistic (or sigmoid) function	$\sigma(t) = \frac{1}{1+\exp(-t)}$
Logistic regression model	$P(Y=1\|X) = \sigma(X^T\beta)$
Squared error loss function	$L(y,\theta) = (y-\theta)^2$
Absolute error loss function	$L(y,\theta) = \|y-\theta\|$
Cross-entropy loss function	$L(y,\theta) = -y\log\theta - (1-y)\log(1-\theta)$
Bias	$\text{Bias}[\hat{\theta},\theta] = E[\hat{\theta}] - \theta$
Variance	$\text{Var}[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$
Mean squared error	$\text{MSE}[\hat{\theta},\theta] = E[(\hat{\theta}-\theta)^2]$

1. (6 points) Visualization

You have data on 500 rental housing units, including rent price, number of bedrooms, and square footage.

(a) (2 pt) Fill the box for **all** suitable visualizations for comparing the distribution of rent price for 2, 3, and 4 bedroom units.

□ Scatterplot ■ Violin plots □ Barplot ■ Boxplots □ None of these

A scatterplot using a discrete 3-valued variable for one axis would almost certainly lead to severe overplotting. A barplot of means or medians would not capture important aspects of the distributions, e.g., spread, skewness.

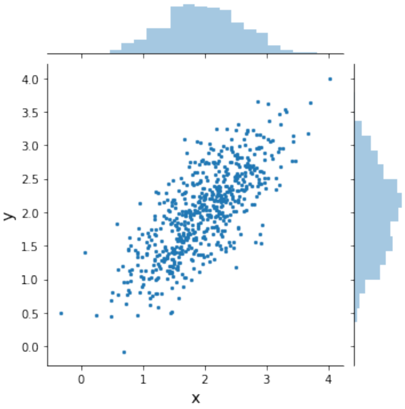
(b) (2 pt) Fill the box for **all** suitable visualizations for examining how rent price and square footage relate.

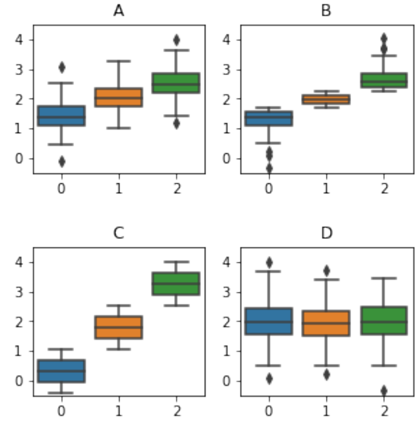
□ Density plots ■ Scatterplot □ Histograms □ Boxplots □ None of these

A two-dimensional density plot is also an acceptable response, since it conveys information similar to a scatterplot. Histograms and boxplots only reflect marginal distributions (i.e., one variable at a time), not joint distributions (i.e., relationships between variables).

The scatterplot generated by `sns.jointplot` below to the left visualizes a dataset containing 600 (x, y) pairs. You partition the pairs by their x value into three groups:

- The 200 with the lowest x values are in group 0.
- The 200 with intermediate x values are in group 1.
- The 200 with the highest x values are in group 2.





(c) (1 pt) Bubble the letter of the boxplot that shows the distribution of x values in each of the three groups.

○ A ⊗ B ○ C ○ D

(d) (1 pt) Bubble the letter of the boxplot that shows the distribution of y values in each of the three groups.

⊗ A ○ B ○ C ○ D

2. (4 points) Classification and Regression Trees (CART) and Random Forests

(a) (1 pt) When predicting a qualitative outcome, the predicted class at each terminal node is:

⊗ the most common class in the node ○ the average class for the node ○ neither

(b) (1 pt) When predicting a quantitative outcome, the predicted value at each terminal node is:

○ the most common outcome in the node ⊗ the average outcome for the node ○ neither

(c) (2 pt) The main idea behind ensemble prediction methods such as Random Forests is that aggregating predictions from predictors applied to multiple bootstrap samples of the learning set:

○ reduces bias ⊗ reduces variability ○ reduces both bias and variability ○ none of these

3. (8 points) Feature Engineering

These instructions are <u>not</u> identical to the similar question on midterm 2, so please read carefully. For each dataset depicted below in a scatterplot, fill in the squares next to **all** of the letters for the functions f that would make it possible to choose scalars β_0, β_1, and β_2 such that $y_i = \beta_0 + x_i\beta_1 + f(x_i)\beta_2$ for all (x_i, y_i) pairs in the dataset. The input to each f is a scalar x shown on the horizontal axis, and the corresponding y value is shown on the vertical axis.

(A) $f(x) = -x$

(B) $f(x) = x$

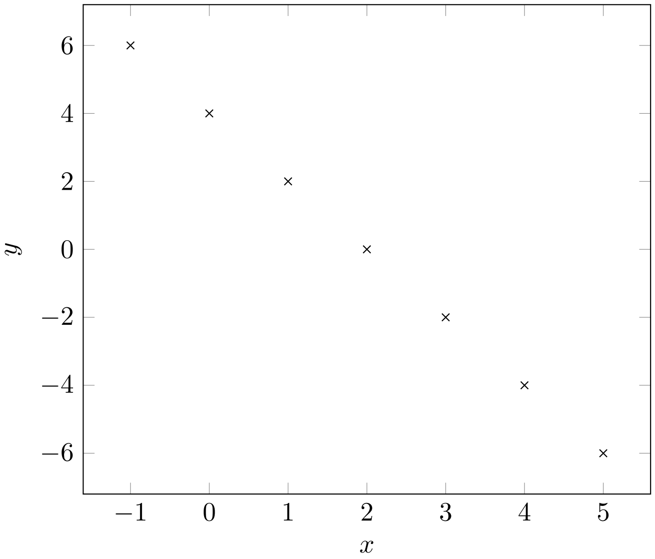
(C) $f(x) = x^2$

(D) $f(x) = |x|$

(E) None of the above

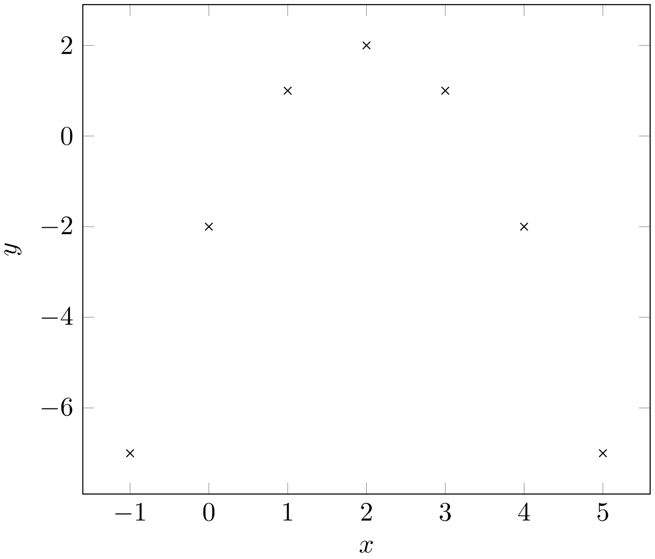
(i) (2 pt) ■ A ■ B ■ C ■ D □ E

All options include a linear term and intercept



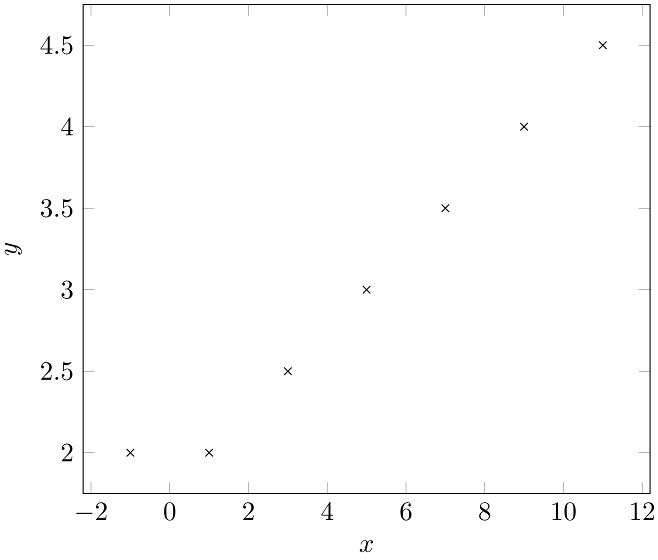
(ii) (2 pt) □ A □ B ■ C □ D □ E

A quadratic term is needed



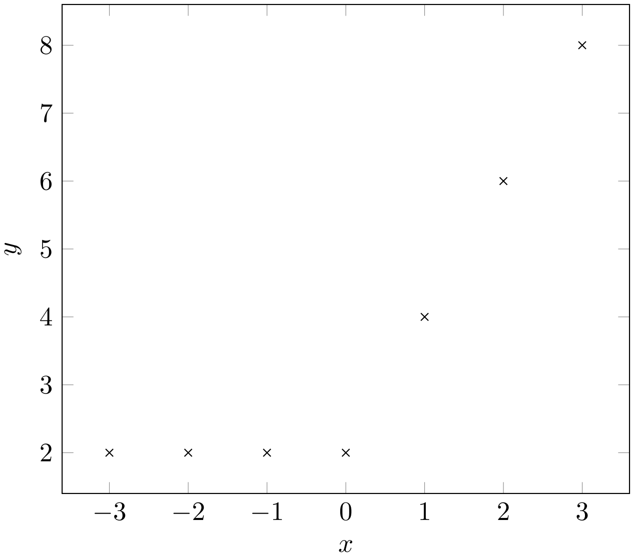
(iii) (2 pt) □ A □ B □ C ■ D □ E

$y_i = 1.75 + 0x_i + 0.25 * |x_i|$



(iv) (2 pt) □ A □ B □ C ■ D □ E

$y_i = 2 + 1x_i + 1|x_i|$



4. (11 points) Regression

i	x_i	y_i
1	$\begin{bmatrix}0\\3\end{bmatrix}$	-3
2	$\begin{bmatrix}3\\0\end{bmatrix}$	0
3	$\begin{bmatrix}6\\0\end{bmatrix}$	3

Suppose you are interested in predicting a one-dimensional quantitative random variable Y (outcome) in terms of a two-dimensional quantitative random variable X (covariates). You have a learning set $(x_1, y_1), \ldots, (x_3, y_3)$ of three observations, given in the table to the right.

(a) (2 pt) You estimate the regression function $E[Y|X]$ using ridge regression with no intercept term and a shrinkage parameter of λ. Select the correct expression for the regression function, where $\beta = [\beta_1 \quad \beta_2]^T$.

⊗ $E[Y|X] = X^T\beta$

○ $E[Y|X] = X^T\beta + \xi$

○ $E[Y|X] = X^T\beta + \lambda \cdot (\beta_1^2 + \beta_2^2)$

○ $E[Y|X] = X^T\beta + \xi + \lambda \cdot (\beta_1^2 + \beta_2^2)$

○ None of these

(b) (2 pt) Why are the letters X and Y capitalized in the models in (a) for the regression function?

⊗ They represent random variables.

○ They represent matrix-valued elements of the learning set.

○ They represent unknown parameters that need to be estimated from the learning set.

(c) (2 pt) There are two derivations below: left and right. Select the one that expresses the regularized empirical risk (mean squared error) for ridge regression for this dataset in terms of β_1 and β_2 **when** $\lambda = 1$.

○ $\frac{1}{3}\left[(-3+3\beta_2)^2 + (3\beta_1)^2 + (3+6\beta_1)^2\right] + (\beta_1^2+\beta_2^2)$
$= 16\beta_1^2 + 12\beta_1 + 4\beta_2^2 - 6\beta_2 + 6$

⊗ $\frac{1}{3}\left[(-3-3\beta_2)^2 + (-3\beta_1)^2 + (3-6\beta_1)^2\right] + (\beta_1^2+\beta_2^2)$
$= 16\beta_1^2 - 12\beta_1 + 4\beta_2^2 + 6\beta_2 + 6$

Let R denote the regularized empirical risk. Then

$$\begin{aligned} R &= \frac{1}{3}\sum_{i=1}^{3}(Y_i - X_i^T\ \beta)^2 + \lambda(\beta_1^2 + \beta_2^2) \\ &= \frac{1}{3}\left((-3 - 3\beta_2)^2 + (-3\beta_1)^2 + (3 - 6\beta_1)^2\right) + (\beta_1^2 + \beta_2^2) \\ &= 16\beta_1^2 - 12\beta_1 + 4\beta_2^2 + 6\beta_2 + 6. \end{aligned}$$

(d) (3 pt) Find the β_1 and β_2 that minimize the regularized empirical risk objective from part (c) above. **Show your work for full credit.**

$\beta_1 = \frac{3}{8} \quad \beta_2 = -\frac{3}{4}$

The empirical risk minimizers can be obtained by taking the first partial derivatives of R and setting these equal to zero.

$$\begin{aligned} \frac{\partial R}{\partial \beta_1} &= 32\beta_1 - 12 \\ \frac{\partial R}{\partial \beta_2} &= 8\beta_2 + 6. \end{aligned}$$

(e) (2 pt) Choose the value of λ for which the values of β_1 and β_2 that would minimize the regularized empirical risk for ridge regression would also minimize the mean squared error on the learning set.

○ -1 ⊗ 0 ○ 1 ○ None of these ○ Impossible to tell

5. (12 points) Logistic Regression

(a) **(4 pt)** Select *always*, *sometimes*, or *never* to describe when each statement below is true about a logistic regression model $P(Y = 1|X) = \sigma(X^T\beta)$, where Y is binary and X is vector-valued. X and β are finite.

○ Always ○ Sometimes ○ Never: $P(Y = 1|X) > X^T\beta$ Sometimes; if $X^T\beta < 0$

○ Always ○ Sometimes ○ Never: $P(Y = 1|X) = P(Y = 0| - X)$ Always.

○ Always ○ Sometimes ○ Never: $P(Y = 1|X) < 1$ Always.

○ Always ○ Sometimes ○ Never: $\sigma(X^T\beta) \leq \sigma(X^T(2 \cdot \beta))$ Sometimes; when $X^T\beta \geq 0$. Let $Z = X^T\beta$. Then, the inequality holds if and only if

$$\begin{aligned}
\frac{1}{1+\exp(-Z)} &\leq \frac{1}{1+\exp(-2Z)} \\
1+\exp(-2Z) &\leq 1+\exp(-Z) \\
\exp(-Z) &\leq 1 \\
Z &\geq 0.
\end{aligned}$$

(b) **(8 pt)** Complete the code below to assign `best_`λ to the regularization parameter λ among `choices` that gives the smallest 2-fold cross-validation risk for L$_1$-regularized logistic regression. The learning set `x` has n observations and m features. Correct labels are named `y`. Assume `minimize` can minimize any function.

```
n, m = x.shape
assert len(y) == n
from scipy.optimize import minimize as scipy_minimize
def minimize(f, k):
    "Return the k-length array that minimizes f."
    return scipy_minimize(f, np.zeros(k))['x']
def sigma(t):
    return 1/(1 + np.exp(-t))
def log_loss(y, z):
    return -y * np.log(z) - (1 - y) * np.log(1 - z)

def risk(λ):
    "Return 2-fold cross-validation risk for regularization hyperparameter λ."
    def objective(b):
        losses = log_loss(y_train, sigma(x_train @ b))
        return np.mean(losses) + λ * sum(np.abs(b))
    half = n // 2
    x_1, x_2 = x[:half, :], x[half:, :]
    y_1, y_2 = y[:half], y[half:]
    x_train, y_train = x_1, y_1
    b = minimize(objective, m)
    risk_1 = np.mean(log_loss(y_2, sigma(x_2 @ b)))
    x_train, y_train = x_2, y_2
    b = minimize(objective, m)
    risk_2 = np.mean(log_loss(y_1, sigma(x_1 @ b)))
    return (risk_1 + risk_2)/2
choices = [2 ** c for c in range(-10, 10)]
best_λ = min([(risk(c), c) for c in choices])[1]
```

6. (13 points) Taxi Trips

The streets of Anytown, USA, are a perfect grid with every block exactly one tenth of a mile long. Taxi drivers only pick up and drop off at intersections. You have acquired a log of all 200,000 taxi trips in Anytown during 2018, which includes measurements for the following 3 variables: number of blocks, total distance, and fare. Taxi fare in Anytown is \$2/mile + \$3/trip. **Assume no measurement error and no missing values.**

(a) **(1 pt)** What is the rank of the matrix with rows corresponding to the 200,000 taxi trips and columns to the 3 variables?

○ 1 ⊗ 2 ○ 3 ○ 4 ○ None of these ○ Impossible to tell

Distance is a constant multiple of number of blocks.

(b) **(1 pt)** What is the rank of a four-column matrix containing columns for the three variables as well as a column of all ones?

○ 1 ⊗ 2 ○ 3 ○ 4 ○ None of these ○ Impossible to tell

Fare is a linear function of distance (a linear combination of distance and a constant intercept feature).

(c) **(4 pt)** If you apply PCA to this four-column matrix, which of the following statements will be true? U, Σ, and V refer to the singular value decomposition of the centered four-column matrix X: $X = U\Sigma V^T$.

○ True ○ False: The sum of the singular values will be the total variance of the dataset. False: the sum of *squared* singular values is the total variance.

○ True ○ False: All singular values will be greater than zero. False: The number of singular values is the number of columns (4), but the number of positive singular values is the rank.

○ True ○ False: The dot product of the first and second columns of V will be equal to 1. False: it will be zero.

○ True ○ False: A scatterplot of the first two columns of $U\Sigma$ will form a straight line. True: there is only one positive singular value.

A simple random sample of 500 taxi trips in Anytown is taken on May 1, 2018. The data matrix for this sample has 4 columns: number of blocks, total distance, fare, and the duration of the trip.

(d) **(5 pt)** Fill the boxes for **all** of the following that are possible to compute reliably via bootstrapping, using this 4-column sample from May 1 as well as the original 3-column population dataset for all of 2018.

□ A confidence interval for the mean trip duration in 2018. False; May 1, 2018 may not be representative of all 2018.

□ A confidence interval for the mean trip duration in 2019. False; May 1, 2018 may not be representative of all 2019.

□ An estimate of the variance of the average duration computed for each possible sample of 500 trips taken on May 1, 2018. True; estimate the variance via bootstrapping.

□ An estimate of the variance of the average duration computed for each possible sample of 100 trips taken on May 1, 2018. True; subsample 100 from the 500 and estimate the variance via bootstrapping.

□ A confidence interval for the difference between the average trip distance on May 1, 2018, and the average trip distance for all of 2018. True; the average trip distance for all 2018 is known, and the May 1, 2018 trip distance can be estimated from the sample.

□ None of the above

(e) **(2 pt)** Using this random sample, you decide to fit an ordinary least squares linear regression model with no intercept using the normal equations, where the outcome is trip duration and the design matrix contains three covariates: number of blocks, total distance, and fare. Which problem below do you expect to encounter **first**?

○ The model will overfit because the sample is too small.

○ Processing the dataset will require more than one computer.

○ Gradient descent will reach a local minimum that is not the global minimum.

⊗ It will be impossible to fit the parameters of the model. The model matrix does not have full column rank, so the matrix inversion in the normal equation is not possible.

○ None of these

7. (22 points) Tables

Fill in both the Python code and the SQL query to produce each result below, assuming that the following tables are stored both as Pandas DataFrames and Sqlite tables. **Only the first few rows are shown for each table. No tables have missing values.** The `scores` table contains a row for each submitted homework in a course and contains columns for the `id` of the student who submitted, the `hw` number, and the `score`. **Assume that** a student can submit a homework at most once, every enrolled student has submitted at least one homework, and every homework has been submitted by at least one student. The `hws` table contains a row for each homework and contains columns for the `name`, which is always the letters "hw" followed by the homework number, as well as the `max` possible score for the homework. Each homework appears exactly once in the `hws` table. The `ugrads` table contains the student ID (`sid`) for each undergraduate enrolled at the university. There are both undergrads and graduate students in the course.

`scores`

id	hw	score
90210	1	4
94720	3	3
90210	3	2

`hws`

name	max
hw1	5
hw2	5
hw3	3

`ugrads`

sid
94709
94114
94720

(a) (6 pt) Select the number of undergrads enrolled in the course.

Python:
```
sum(ugrads['sid'].isin(scores['id']))
```

SQL:
```
SELECT COUNT(*) FROM ugrads WHERE sid IN (SELECT id FROM scores);
```

(b) (6 pt) List the IDs for students who are missing at least one homework.

Python:
```
counts = scores.groupby('id').size()

list(counts[counts < hws.shape[0]].index)
```

SQL:
```
SELECT id FROM scores GROUP BY id HAVING COUNT(*) < (SELECT COUNT(*) FROM hws);
```

(c) (6 pt) Build a table with one row per homework containing the total number of points scored by undergraduates on that homework.

Python:
```
scores.merge(ugrads, left_on='id', right_on='sid')

 .groupby('hw')['score'].sum())
```

SQL:
```
SELECT hw, SUM(score) FROM scores JOIN ugrads ON id=sid GROUP BY hw;
```

(d) (4 pt) Select the average fraction of possible points that was scored on each submitted homework.

Python:
```
names = scores['hw'].apply(lambda n: f'hw{n}')

np.mean(scores['score'] / np.array(hws.set_index('name').loc[names, 'max']))
```

SQL:
```
SELECT AVG(score / max) FROM scores JOIN hws ON name='hw'||hw;
```

8. (8 points) Bias and Variance

(a) (2 pt) Select the best description of the *bias* of an estimator $\hat{\theta}$ for a parameter θ, where $\hat{\theta}$ is computed based on a simple random sample from the population of interest.

○ The difference between θ and the value of $\hat{\theta}$ for a particular sample.

○ The average difference between θ and $\hat{\theta}$ computed for each element of the sample.

⊗ The average difference between θ and $\hat{\theta}$ over all possible simple random samples from the population.

○ The difference between the expected value of θ and the value of $\hat{\theta}$ for a particular sample.

Bias is defined as the difference between the expected value of an estimator and the parameter of interest, $E[\hat{\theta}] - \theta$, where the expected value refers to the sampling distribution of the estimator.

You estimate the maximum possible outcome (number of dots) of rolling of a fair 6-sided die using the outcome of a single roll of the die (pretending you don't already know the number of dots on the faces of the die).

(b) (2 pt) What is the bias of this estimator?

○ -3.5 ⊗ -2.5 ○ -1.5 ○ 0 ○ 1.5 ○ 2.5 ○ 3.5 ○ Other: _________

The expected number of dots for one roll of a die is

$$\frac{1}{6}\sum_{i=1}^{6} i = \frac{21}{6} = \frac{7}{2}.$$

Thus bias is $3.5 - 6 = -2.5$.

(c) (1 pt) What is the variance of this estimator?

○ 0 ○ $\frac{5}{2}$ ○ $\frac{7}{2}$ ⊗ $\frac{35}{12}$ ○ $\frac{47}{12}$ ○ $\frac{55}{6}$ ○ $\frac{61}{6}$ ○ Other: _________

The variance for the number of dots in one roll of a die is

$$\frac{1}{6}\sum_{i=1}^{6} i^2 - 3.5^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

(d) (1 pt) What is the mean squared error of this estimator?

○ 0 ○ $\frac{5}{2}$ ○ $\frac{7}{2}$ ○ $\frac{35}{12}$ ○ $\frac{47}{12}$ ⊗ $\frac{55}{6}$ ○ $\frac{61}{6}$ ○ Other: _________

The mean squared error is the sum of the variance and of the square of the bias

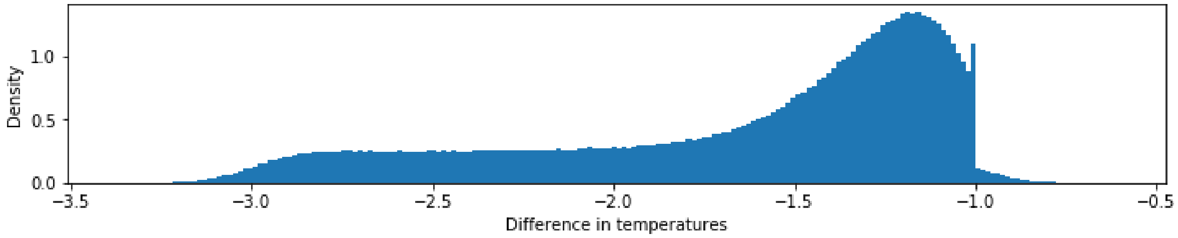
$$\frac{35}{12} + \frac{25}{4} = \frac{35+75}{12} = \frac{55}{6}.$$

(e) (2 pt) How would the magnitude (absolute value) of the estimator's bias change if your estimator was the maximum outcome that appeared in 100 rolls of the die instead of just one roll?

○ It would increase ⊗ It would decrease ○ It would stay the same

9. (4 points) Sampling Distributions

The temperature difference between the outside and inside of a room measured each minute has this distribution:



You sample 10 differences at random with replacement. For each distribution below, select the histogram that best visualizes the distribution.

○ A ○ B ○ C ○ D ○ E ○ F: The sampling distribution of the sample 90th percentile.

E: A distribution similar in shape to normal (but not that normal because of the small sample size) that is centered near the 90th percentile of the population

○ A ○ B ○ C ○ D ○ E ○ F: The sampling distribution of the sample median.

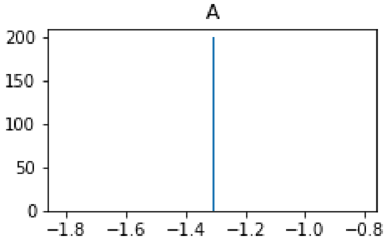
B: Centered at the median of the population

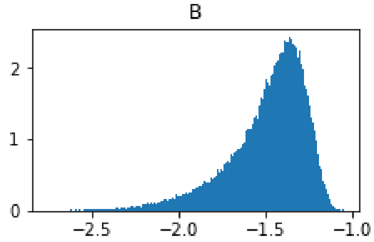
○ A ○ B ○ C ○ D ○ E ○ F: The sampling distribution of the sample standard deviation.

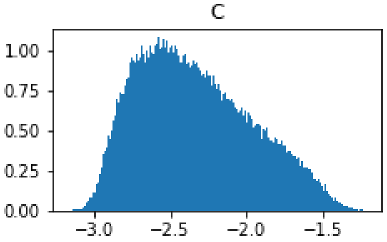
D: The only distribution over postiive numbers.

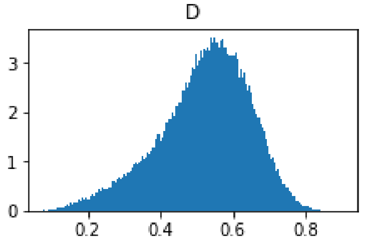
○ A ○ B ○ C ○ D ○ E ○ F: The bootstrap distribution of the sample median.

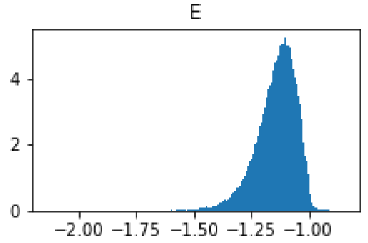
F: Medians of bootstrap samples for small sample sizes can only have a small number of possible values.

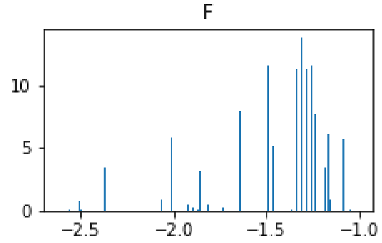












10. (8 points) Distributed File Systems

A large file is partitioned into 8 equal-sized blocks, and each block is replicated 2 times. For each block, a simple random sample of 2 machines is chosen (without replacement) out of 4 total machines, and one replica of the block is stored on each machine chosen in this way.

(a) (2 pt) What is the expected number of blocks stored on each machine?

○ 0 ○ 1 ○ 2 ⊗ 4 ○ 8 ○ 16 ○ 32 ○ Other: _________

(b) (2 pt) If one machine fails, what is the chance that the entire file can still be read?

○ 0 ○ $\frac{1}{8}$ ○ $\frac{1}{4}$ ○ $\frac{1}{2}$ ○ $\frac{3}{4}$ ○ $\frac{7}{8}$ ⊗ 1 ○ Other: _________

(c) (2 pt) If two machines fail, what is the chance that the entire file can still be read?

○ 0 ○ $\frac{1}{2}$ ○ $\left(\frac{1}{2}\right)^8$ ○ $\frac{1}{6}$ ○ $\frac{5}{6}$ ⊗ $\left(\frac{5}{6}\right)^8$ ○ $1-\left(\frac{5}{6}\right)^8$ ○ 1 ○ Other: _________

(d) (2 pt) If three machines fail, what is the chance that the entire file can still be read?

○ 0 ○ $\frac{1}{2}$ ⊗ $\left(\frac{1}{2}\right)^8$ ○ $\frac{1}{6}$ ○ $\frac{5}{6}$ ○ $\left(\frac{5}{6}\right)^8$ ○ $1-\left(\frac{5}{6}\right)^8$ ○ 1 ○ Other: _________

11. (4 points) Distributed Computing

Assume you have the long-running remote function `f` defined below and you want to call it 10 times.

```
@ray.remote
def f():
   run_time = 10 * np.random.random()  # Chosen uniformly at random from 0 to 10 seconds
   time.sleep(run_time)                # Do nothing for run_time seconds
   return run_time
```

(a) (1 pt) Assuming there is no overhead, what is the **expected** amount of time required to call `f` 10 times if all calls are executed serially?

○ 1 second ○ 5 seconds ○ 10 seconds ⊗ 50 seconds ○ 100 seconds ○ None of these

Note: Since many students were confused by the first comment in f and thought that `run_time` could range from 0 to 100, credit was also give to "none of these" if the answer for part (b) was 100.

(b) (1 pt) Assuming there is no overhead, what is the **maximum** amount of time required to call `f` 10 times if all calls to f are executed in parallel?

○ 1 second ○ 5 seconds ⊗ 10 seconds ○ 50 seconds ○ 100 seconds ○ None of these

Note: Since many students were confused by the first comment in f and thought that `run_time` could range from 0 to 100, credit was also give to 100 if the baswer for part (a) was "None of these".

(c) (2 pt) Select **all** of the expressions below that correctly compute a list of the return values from calling `f` 10 times **and** execute all calls in parallel.

□ `[f() for _ in range(10)]`

□ `[f.remote() for _ in range(10)]`

■ `ray.get([f.remote() for _ in range(10)])`

□ `[ray.get(f.remote()) for _ in range(10)]`

□ None of these

12. (0 points) Optional: Draw a Picture About Berkeley Data Science