DS-100 Midterm Exam

Fall 2017

Name: ______________________________

Email: ______________________________@berkeley.edu

Student ID: ______________________________

Instructions:

- This exam must be completed in the **1.5 hour time** period ending at **8:30PM**.
- Note that some questions have bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**.
- When selecting your choices, you must **shade** in the box/circle. Checkmarks will likely be mis-graded.
- You may use a single page (two-sided) study guide.
- Work quickly through each question. There are a total of 116 points on this exam.

Honor Code:

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: ______________________________

Syntax Reference

Regular Expressions

`"^"` matches the position at the beginning of string (unless used for negation `"[^]"`)

`"$"` matches the position at the end of string character.

`"?"` match preceding literal or sub-expression 0 or 1 times. When following `"+"` or `"*"` results in non-greedy matching.

`"+"` match preceding literal or sub-expression *one* or more times.

`"*"` match preceding literal or sub-expression *zero* or more times

`"."` match any character except new line.

`"[ ]"` match any one of the characters inside, accepts a range, e.g., `"[a-c]"`.

`"( )"` used to create a sub-expression

`"\d"` match any *digit* character. `"\D"` is the complement.

`"\w"` match any *word* character (letters, digits, underscore). `"\W"` is the complement.

`"\s"` match any *whitespace* character including tabs and newlines. `\S` is the complement.

`"\b"` match boundary between words

Some useful `re` package functions.

`re.split(pattern, string)` split the `string` at substrings that match the `pattern`. Returns a list.

`re.sub(pattern, replace, string)` apply the `pattern` to `string` replacing matching substrings with `replace`. Returns a string.

Useful Pandas Syntax

```
df.loc[row_selection, col_list]  # row selection can be boolean
df.iloc[row_selection, col_list] # row selection can be boolean
df.groupby(group_columns)[['colA', 'colB']].sum()
pd.merge(df1, df2, on='hi') # Merge df1 and df2 on the 'hi' column
```

```
pd.pivot_table(df,                   # The input dataframe
               index=out_rows,       # values to use as rows
               columns=out_cols,     # values to use as cols
               values=out_values,    # values to use in table
               aggfunc="mean",       # aggregation function
               fill_value=0.0)       # value used for missing comb.
```

Data Generation and Probability Samples

For each of the following questions select the **single best answer**.

1. [2 Pts] A political scientist is interested in answering a question about a country composed of three states with exactly 10000, 20000, and 30000 voting adults. To answer this question, a political survey is administered by randomly sampling 25, 50, and 75 voting adults from each town, respectively. Which sampling plan was used in the survey?

   ○ cluster sampling

   √ **stratified sampling**

   ○ quota sampling

   ○ snowball sampling

2. [2 Pts] A deck with 26 cards labeled `A` through `Z` is thoroughly shuffled, and the value of the **third** card in the deck is recorded. What is the probability that we observe the letter `C` on the third card?

   √ $\frac{1}{26}$ ○ $\frac{3}{26}$ ○ $\frac{25}{26} \cdot \frac{24}{26} \cdot \frac{1}{26}$ ○ $\frac{1}{26} \cdot \frac{1}{26} \cdot \frac{24}{26}$ ○ None of the above.

3. [3 Pts] Suppose Sam visits your store to buy some items. He buys toothpaste for $2.00 with probability 0.5. He buys a toothbrush for $1.00 with probability 0.1. Let the random variable X be the total amount Sam spends. What is $\mathbf{E}[X]$? Show your work in the space provided.

   √ **$1.10**

   ○ $1.5

   ○ $3.00

   ○ The toothpaste purchase may not be independent of the toothbrush purchase so we can't compute this expectation.

   You may show your work in the following box for partial credit:

Solution: Let $X_{\text{toothpaste}}$ be the amount Sam spends on toothpaste, and $X_{\text{toothbrush}}$ be the amount Sam spends on a toothbrush.

From the linearity of expectation, we have:

$$\mathbf{E}[X] = \mathbf{E}[X_{\text{toothpaste}} + X_{\text{toothbrush}}] = \mathbf{E}[X_{\text{toothpaste}}] + \mathbf{E}[X_{\text{toothbrush}}]$$

We know that $\mathbf{E}[X_{\text{toothpaste}}] = (0.5)(0) + (0.5)(2) = 1$, and $\mathbf{E}[X_{\text{toothbrush}}] = (0.9)(0) + (0.1)(1) = 0.1$. Thus, $\mathbf{E}[X] = 1.1$.

4. [3 Pts] Suppose we have a coin that lands heads 80% of the time. Let the random variable X be the *proportion* of times the coin lands tails out of 100 flips. What is **Var**$[X]$? You must show your work in the space provided.

○ 0.8 ○ 0.16 ○ 0.04 ✓ 0.0016 ○ 0.008

Solution: Let X_i be the outcome of the i^{th} spin. If the i^{th} spin lands heads than we say $X_i = 1$ and otherwise $X_i = 0$. Then the *proportion of times* X_i lands heads is given by:

$$Y = \frac{1}{100}\sum_{i=1}^{n} X_i$$

We can compute the variance of Y using the following identities:

$$\begin{aligned}
\textbf{Var}\left[Y\right] &= \textbf{Var}\left[\frac{1}{100}\sum_{i=1}^{n} X_i\right] && (1)\\
&= \frac{1}{100^2}\textbf{Var}\left[\sum_{i=1}^{n} X_i\right] && \text{(Squared variance of constant multiple.)}\\
&= \frac{1}{100^2}\sum_{i=1}^{n}\textbf{Var}\left[X_i\right] && \text{(Ind. Variables implies linearity of var.)}\\
&= \frac{1}{100^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{100}\\
&= \frac{.8(1-.8)}{100} = \frac{.16}{100} = .0016
\end{aligned}$$

5. A small town has 5 houses with the following people living in each house:

Abe, Ben
Cat, Dan, Emma
Frank, George
Hank, Ira, Jen
Kim, Lars

Suppose we take a **cluster sample** of 2 houses (without replacement), what is the chance that:

(1) [2 Pts] Kim and Lars are in the sample

○ 0 ○ 1/20 ○ 1/10 ○ 1/6 ○ 1/5 ✓ **2/5** ○ 1

You may show your work in the following box for partial credit:

Solution: The chance that Kim and Lars are in the same sample is given by the chance of choosing their house. The chance of choosing the their house on the first draw is $\frac{1}{5}$. Because we are drawing without replacement. The chance of choosing their house on the second draw is given by the chance of not choosing their house on the first draw ($\frac{4}{5}$) times the chance of choosing their house on the second draw ($\frac{1}{4}$). Thus the total chance of choosing them in the first two draws is:

$$\frac{1}{5} + \frac{4}{5} \times \frac{1}{4} = \frac{2}{5}$$

(2) [2 Pts] Kim, Abe, and Ben are in the sample

○ 0 ○ 1/20 ✓ **1/10** ○ 1/6 ○ 1/5 ○ 2/5 ○ 1

You may show your work in the following box for partial credit:

Solution: To draw Kim, Abe, and Ben we would need to draw both of their houses. This can be done two ways (draw Abe and Ben's house first and then Kim's or vice versa). Each way has probability:

$$\frac{1}{5} \times \frac{1}{4}$$

Thus the total probability is:

$$2 \times \frac{1}{5} \times \frac{1}{4} = \frac{2}{20} = \frac{1}{10}$$

(3) [1 Pt] Kim and Dan are in the sample - **Select all that apply**

□ The same as the chance Kim and Lars are in the sample

√ **The same as the chance Kim, Abe, and Ben are in the sample**

□ Neither of the above

Data Cleaning and EDA

6. **True or False.** For each of the following statements select true or false.

(1) [1 Pt] Exploratory data analysis is the process of testing key hypotheses.

○ True √ **False**

Solution: False. Exploratory data analysis is the process of gaining understanding about data to inform future analysis.

(2) [1 Pt] The structure of the data describes how it is formatted and organized.

√ **True** ○ False

Solution: True. The structure of data includes its formatting (e.g., JSON, CSV, XML, raw text) as well as the fields and organization of records.

(3) [1 Pt] Throughout the process of exploratory data analysis it is often necessary to transform and clean data.

√ **True** ○ False

Solution: True. A key step in exploratory data analysis is identify and in some cases correcting anomalies and issues with data.

(4) [1 Pt] During the data cleaning process it is generally a good idea to drop records that contain missing values.

○ True √ **False**

Solution: False. Nooooo. It is very important that the cleaning process is done with care to avoid introducing transformations that might bias subsequent analysis. Dropping records with missing values, for example, missing addresses, could substantially bias the data (e.g., removing homeless people).

7. In homework 3, we analyzed ride sharing data comparing the weekday and weekend patterns for both casual and registered riders.

(1) [1 Pt] On **weekdays**, the number of casual riders was most frequently _____ the number of registered riders.

○ higher than √ **lower than** ○ similar to

(2) [1 Pt] Which group of riders demonstrated a pronounced bi-modal daily usage pattern:

○ Casual Riders √ **Registered Riders** ○ Both casual and registered riders.

8. Using the following snippet of data to answer each of the questions below.

Business.data

```
"business_id","name","address","phone"
10,"TIRAMISU KITCHEN","033 BELDEN PL","+14154217044"
19,"LIFESTYLE CAFE","1200 VAN NESS AVE","+14157763262"
24,"OMNI S.F. HOTEL","    ","9999999999999999"
42,"The "Best", Food!","500 CALIFORNIA ST","+14156211114"
43,"The "Best", Food!","3716 Cesar Chavez","+14156211114"
```

(1) [1 Pt] Which of the following **best** describes the format of this file.

○ Raw Text

○ Tab Separated Values

✓ **Comma Separated Values**

○ JSON

(2) [1 Pt] Which of the following **best** describes the granularity of each record?

○ Restaurant Chains

✓ **Individual Restaurant Locations**

○ Strings

○ Daily

(3) [4 Pts] Select **all** the true statements.

✓ **From the available data the `business_id` appears to be a primary key.**

□ There appear to be no missing values

✓ **While the data appears to be quoted there may be issues with the quote character.**

□ There are nested records.

□ None of the above statements is true.

Transformations and Smoothing

9. [3 Pts] Which of the following are reasonable motivations for applying a power transformation? **Select all that apply**:

   √ **To help visualize highly skewed distributions**

   □ Bring data distribution closer to random sampling

   √ **To help straighten relationships between pairs of variables.**

   □ Reduce the dimension of data

   □ Remove missing values

   □ None of the above

10. [3 Pts] Which of the following transformations could help make linear the relationship shown in the plot below? **Select all that apply**:

    √ $\log(y)$ √ x^2 √ $\sqrt{y}$ □ $\log(x)$ □ y^2 □ None of the above

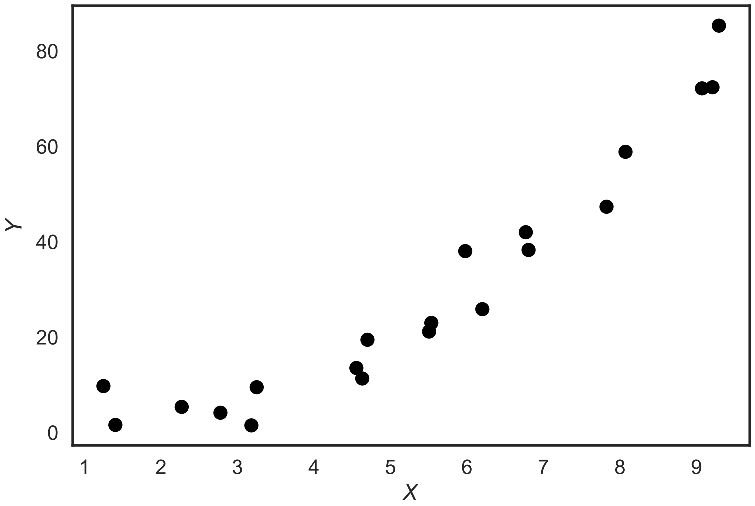


Figure 1

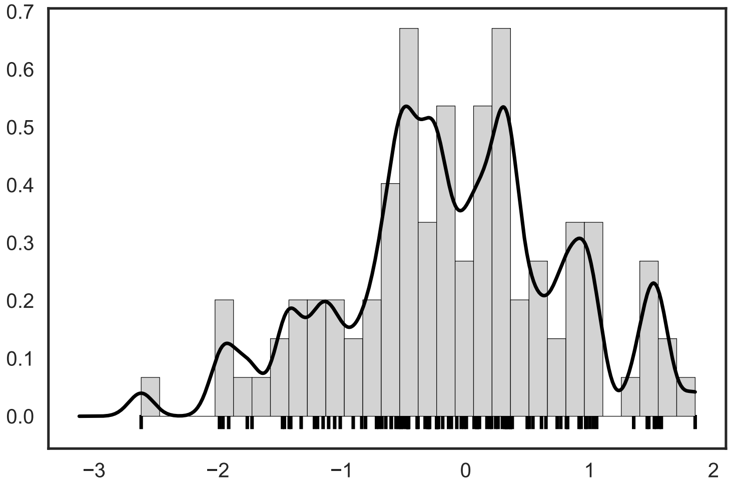


Figure 2

11. [2 Pts] The above plot contains a histogram, rug plot, and Gaussian kernel density estimator. The Gaussian kernel is defined by:

$$K_\alpha(x, z) = \frac{1}{\sqrt{2\pi\alpha^2}} \exp\left(-\frac{(x-z)^2}{2\alpha^2}\right)$$

Judging from the shape of separate standing peaks, which of the following is the most likely value for the kernel parameter α.

○ $\alpha = 0$ ✓ $\alpha = 0.1$ ○ $\alpha = 10$ ○ $\alpha = 100$

Regular Expressions

12. [2 Pts] Select **all** the strings that **fully match** the regular expression: `[^dp]an`

    √ **Dan** □ pan √ **fan** √ **man** □ None of the above.

13. [2 Pts] Select **all** the strings that **fully match** the regular expression: `<[a-z]*@\w+.edu>`

    □ `<xin.wang@berkeley.edu>`

    √ **`<@berkeley$edu>`**

    √ **`<xinwang@berkeley#edu>`**

    □ `<xinwang@.edu>`

    □ None of the above strings match.

14. [2 Pts] Select **all** the strings that **fully match** the regular expression: `^Go.*`

    □ `Way to ^Go!`

    √ **`Go Bears!`**

    □ `go trees?`

    □ None of the above strings match

15. [2 Pts] What is the result of evaluating the following python command?

```
len(re.split(r"\d+", "You get a 99.9 on the exam."))
```

○ 2 √ **3** ○ 4 ○ 5

16. For the following tasks, write the corresponding Python code or regular expression.

    (1) [2 Pts] Write a regular expression that only matches sub strings consisting of an `a` immediately followed by zero or one `b` characters.

    regx = r'____'

    Solution:

    regx = r'ab?'

    (2) [3 Pts] Suppose we've run the code below:

```
text = 'Data\t \t Science  100'
```

Use a method in the `re` module to replace all the continuous segments of spaces with a single comma. The resulting string should look like "`Data,Science,100`".

`re.`____

Solution:

```
re.sub(r'\s+', ',', text)
```

DataFrames, Joins, and Aggregation

17. The `ti` and `fare` DataFrames contain data of the people aboard the Titanic when it crashed:

```
>>> ti.head()                          |  >>> fare.head()
   survived  class      sex     id     |          fare  alone    id
0         0  Third     male   1410     |  0  73.5000   True  1457
1         1  First   female   1522     |  1   9.2250   True  1645
2         1  Third   female   1864     |  2   8.6625   True  1716
3         1  First   female   1687     |  3  59.4000  False  1367
4         0  Third     male   1173     |  4  18.0000  False  1639
```

Both tables contain one row for each passenger, uniquely identified by the `id` column. Here's a description of the columns in each DataFrame:

DataFrame `ti`	**DataFrame `fare`**
`survived`: 1 if the person survived, else 0	`fare`: Price of ticket in USD
`class`: ticket class (First, Second, or Third)	`alone`: True if the person was alone at purchase.
`sex`: Sex of person (male or female)	

Fill in the blanks to compute the following statements. You may assume that the pandas module is imported as `pd`. **You may not use more lines than the ones provided.**

(1) [2 Pts] The total number of survivors.

____

Solution:

```
ti[’survived’].sum()
```

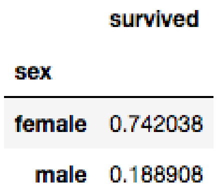
(2) [4 Pts] The proportion of females who survived (a `float`).

```
ti.loc[________________________________,____________].mean()
```

Solution:

```
ti.loc[ti[’sex’] == ’female’, ’survived’].mean()
```

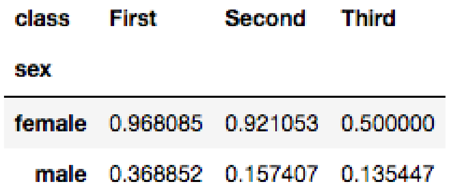
(3) [4 Pts] A DataFrame containing the proportion of survivors for each sex. It should look like:



	survived
sex	
female	0.742038
male	0.188908

Solution:
```
ti[['survived', 'sex']].groupby('sex').mean()
```

(4) [5 Pts] A DataFrame containing the proportion of survivors for each sex and class. It should look like:



class	First	Second	Third
sex			
female	0.968085	0.921053	0.500000
male	0.368852	0.157407	0.135447

Solution:
```
pd.pivot_table(ti, values='survived',
                    index='sex', columns='class')
```

(5) [8 Pts] A DataFrame containing the proportion of survivors for each sex after filtering out those that bought their ticket alone. The table should have the same structure as (3) but with different numbers.

```
merged = ________________________________________________

(merged__________________________________________________

 ________________________________________________________

 ________________________________________________________)
```

Solution:

```
merged = pd.merge(ti, fare, on='id')
(merged[merged['alone']]
 .loc[:, ['survived', 'sex']]
 .groupby('sex').mean())
```

18. [3 Pts] From the following list **select all** statements that are true for Pandas Data Frames.

√ **All data frames must have an index.**

□ All columns must be the same type.

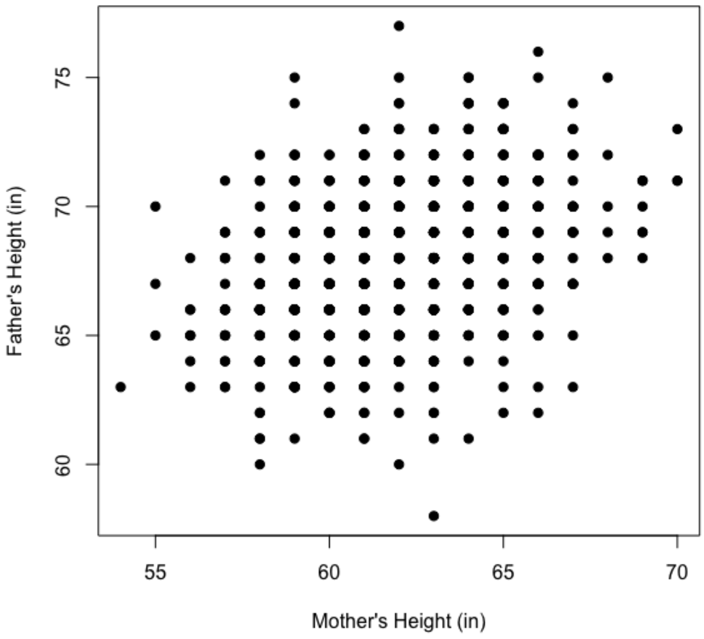
√ **You can always index a record by its row number.**

□ Missing values in **string** columns are always encoded as `NaN`.

□ None of the above

Visualizations

19. The figure below is a scatter plot of the heights of mothers (in) and fathers (in) of a sample of 1000 UC Berkeley students.



(1) [2 Pts] What is the main problem with this plot?

○ Choice of scale

○ Jiggling the baseline

○ Aspect ratio

√ **Overplotting**

○ Lack of context

○ Perception (length, angle, area)

(2) [2 Pts] What is the remedy for this problem?

○ Overlay plots

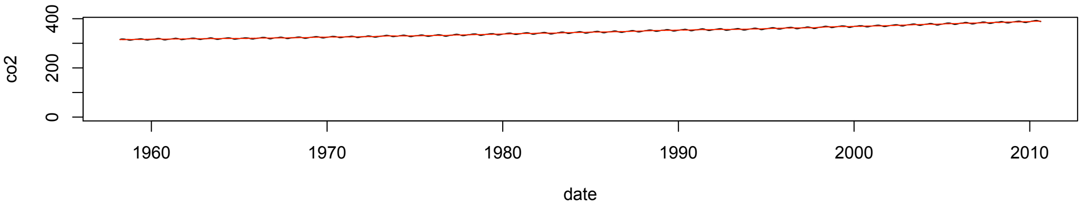
√ **Jitter values**

○ Use color to condition on student's gender

○ Transform one variable or the other or both

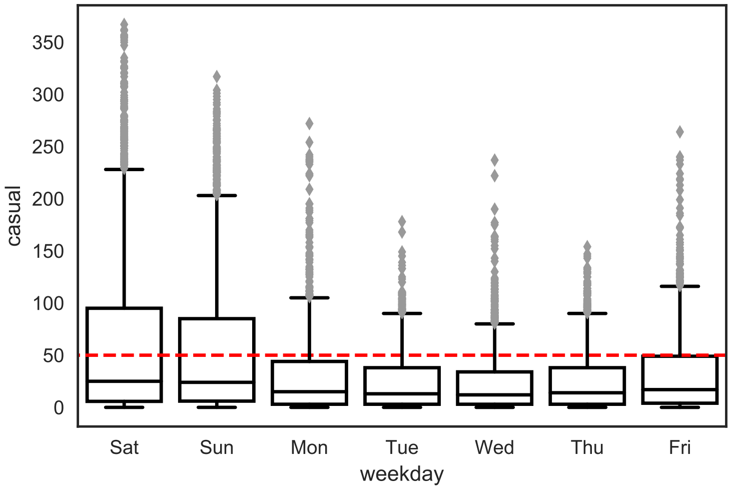
○ Improve labels and legends

20. [2 Pts] The following figure is a line plot of CO_2 emissions over time. What is the main problem with this plot?



√ **Empty data region** ○ Jiggling the baseline ○ Overplotting
○ Lack of context ○ Perception (length, angle, area)

21. Consider the following visualization of the number of casual riders per hour by day of the week, which has been constructed from the bike sharing data used in Homework 3.



(1) [2 Pts] Which days of the week frequently (at least 75% of the time) had fewer than 50 casual riders? **Select all that apply.**

□ Saturday □ Sunday √ **Monday** √ **Tuesday** □ None of the above.

(2) [3 Pts] Which of the following describe conclusions that we can draw about the distribution of rider counts on Tuesdays using the above plot? **Select all that apply.**

□ Skewed left □ Symmetric √ **Skewed right** □ Unimodal √ **Has outliers**
□ None of the above

Estimation and Loss Minimization

22. Consider the following loss function.

$$L(\theta, x) = \begin{cases} 4(\theta - x) & \theta \geq x \\ x - \theta & \theta < x \end{cases}$$

(1) [2 Pts] Select **all** statements that are true.

□ The loss function is concave.

√ **The loss function is convex.**

□ The loss function is smooth.

□ None of the above statements are true.

(2) [4 Pts] Given a sample $x_1, \ldots x_n$, which value of θ minimizes the average loss? **Show your work in the space provided.**

√ 20^{th} **percentile** ○ 25^{th} percentile ○ 75^{th} percentile ○ 80^{th} percentile

Solution:

(3) [2 Pts] The optimal value θ^* is a percentile for the

√ **sample**

○ population

23. We propose the following simple model for a dataset consisting of four points $\mathcal{D} = \{0, 2, 4, 10\}$:

$$y = \theta^*$$

Use the following plots of loss functions for this model to answer the following questions.

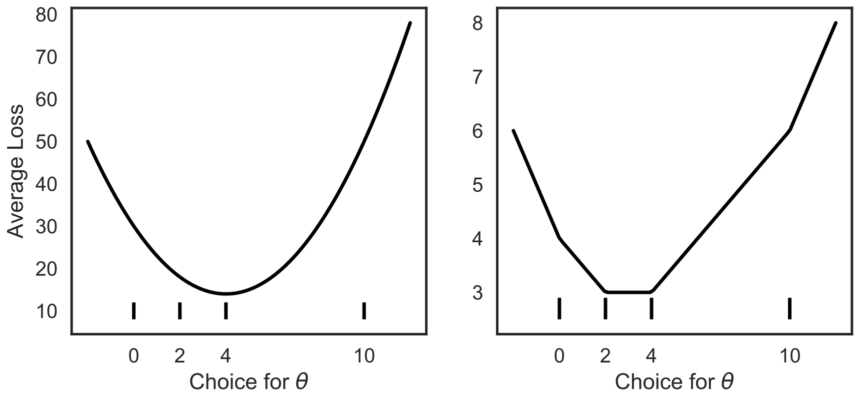


Figure 3

(1) [2 Pts] Which choice(s) for θ minimize the average squared loss? **Select all that apply.**

□ 2 □ 3 √ **4** □ 10 □ None of the above

(2) [2 Pts] Which choice(s) for θ minimize the average absolute loss? **Select all that apply.**

√ **2** √ **3** √ **4** □ 10 □ None of the above

(3) [2 Pts] Suppose we add an observation at $y_5 = 100$. Which choice(s) for θ minimize the average absolute loss? **Select all that apply.**

□ A value smaller than 3 □ 3 √ **4** □ 5 □ A value larger than 5

(4) [2 Pts] Which loss function is most sensitive to outliers?

○ absolute loss √ **squared loss** ○ Huber loss

24. [4 Pts] Which θ minimizes the following loss function for a dataset D comprised of (x_i, y_i) pairs? **Show your work in the space provided.**

$$L(\theta, D) = \sum_{i=1}^{n} (y_i - \theta x_i)^2$$

○ $\theta = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$ ✓ $\theta = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ ○ $\theta = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$ ○ $\theta = \frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i^2}$

Solution:

Sampling Distribution, Bootstrapping, and Confidence Intervals

25. **True or False.** For each of the following statements select true or false.

(1) [1 Pt] Suppose we have 100 samples drawn independently from a population. If we construct a 95% confidence interval for each sample, we expect 95 of them to include the **sample** mean.

○ True [✓ **False**]

Solution: False. All of them should include the sample mean.

(2) [1 Pt] The law of large numbers tells us that as the sample size grows, the average of a random sample with replacement from a population gets closer to the population average.

[✓ **True**] ○ False

Solution: True. The sample looks more and more like the population as the sample size grows, and we have seen that the standard error of the sample mean shrinks like $\sigma/\sqrt{n}$ as the sample mean approaches the population mean.

(3) [1 Pt] We often prefer a pseudo-random number generator because our simulations results can be exactly reproduced by controlling the seed.

[✓ **True**] ○ False

Solution: True. This is an essential aspect of reproducible data analyses and simulation studies.

(4) [1 Pt] As the sample size increases, the bootstrapped sampling distribution of a statistic will always become roughly normal.

○ True [✓ **False**]

Solution: False. The bootstrapped sampling distribution should resemble the true sampling distribution of the statistic, which may or may not be normal.

26. [2 Pts] Suppose we have a census of household incomes for the entire state of California. Which of the following histograms would most closely resemble a normal curve?

√ **A histogram of 10000 sample means from samples of size $n = 1000$.**

○ A histogram of incomes from a SRS of $n = 10000$ households.

○ A histogram of incomes from the entire census.

○ None of the above would resemble a normal distribution.

27. [5 Pts] Suppose we have a Pandas Series called **`thePop`** which contains a census of **25000 subjects**. We also have a simple random sample of **400 individuals** saved in the Series **`theSample`**. We are interested in studying the behavior of the bootstrap procedure on the simple random sample. Fill in the blanks in the code below to construct **10000 bootstrapped estimates** for the **median**.

```
boot_stats = [

        ______________________.

        sample(n = _____________, replace = _____________).

        ______________________()


        for j in range(________________)
    ]
```

Solution:

```
boot_stats = [

        theSample.

        sample(n = 400, replace = True).

        median()

        for j in range(10000)

        ]
```

End of Exam