Data 8 Spring 2020

Foundations of Data Science

MIDTERM SOLUTIONS

INSTRUCTIONS

- The exam is worth 87 points. You have 100 minutes to complete it.
- The exam is closed book, closed notes, closed computer/phone/tablet, closed calculator, except the official midterm exam reference guide provided with the exam.
- Write/mark your answers on the exam in the blanks/bubbles provided. Answers written anywhere else will not be graded. Unless the question specifically asks you to explain your answer, you do not need to do so, and if you write an explanation it will not be graded.
- If you need scratch paper, you are welcome to use the reference guide and the back of this cover page. Scratch work will not be graded.
- For all Python code, you may assume that the statements `from datascience import *` and `import numpy as np` have been executed. Do not use features of the Python language that have not been described in this course.
- In any part, you are free to use any tables, arrays, or functions that have been defined in previous parts of the same question, and you may assume they have been defined correctly.

Last name	
First name	
Student ID number	
Calcentral email (`_@berkeley.edu`)	
Lab GSI	
← 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)**	
Your room & seat number (for example, 155 Dwinelle A1)	

1. (12 points) Making Smoothies

Professor Sahai makes a smoothie every morning before coming to Data 8 lecture. When he makes a smoothie, he picks two fruits completely at random from the fridge. Suppose that his fridge has 3 pears, 5 guavas, 4 watermelons, and 3 kiwis when fully stocked. Each fruit is equally likely to be chosen.

In each part below write a mathematical expression (not Python) that evaluates to the probability described. **You do not need to simplify any arithmetic. Please do not multiply by 100 to get percents.**

(a) (3 pt) The probability that his next smoothie only contains kiwis. (Assume the fridge is fully stocked before making the smoothie.)

```
(3/15) * (2/14)
```

(b) (3 pt) The probability that his next smoothie has at least one watermelon. (Assume the fridge is fully stocked before making the smoothie.)

```
1 - (11 / 15) * (10 / 14)
```

d

(c) (3 pt) The probability that his next smoothie has a kiwi and a pear. (Assume the fridge is fully stocked before making the smoothie.)

```
2 * (3 / 15) * (3 / 14)
```

(d) (3 pt) The probability that his next smoothie has two different fruits. (Assume the fridge is fully stocked before making the smoothie.)

```
1 - (3/15 * 2/14) - (3/15 * 2/14) - (5/15 * 4/14) - (4/15 * 3/14)
```

2. (18 points) Graduates

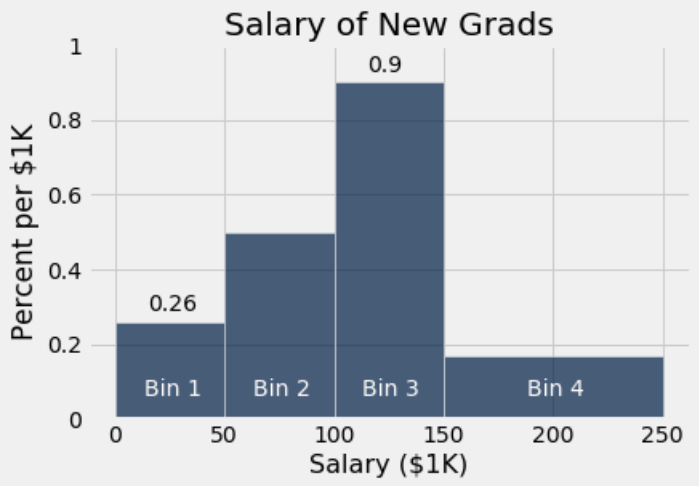
Cindy is curious about the financial situations of recent Berkeley graduates so she surveys 200 graduates from Spring 2019 and collects her results in the table `graduates`, which contains a row for each student in the survey. Here are the first few rows:

major	internships	salary	debt
Data Science	2	120,000	10,000
Psychology	1	79,500	20,000
Business Administration	0	95,000	14,500

The table contains four columns:

- **major**: a string, the student's primary major declaration
- **internships**: an int, the number of summer internships the student did before graduating
- **salary**: an int, the student's starting annual salary (in USD)
- **debt**: an int, the student's unpaid debt (in USD)

Suppose she makes the histogram of `'salary'` below using 4 bins. The heights of Bin 1 and Bin 3 are shown on the histogram.



For parts (a)-(c), write a mathematical expression (not Python) that evaluates to the quantity described. **You do not need to simplify any arithmetic.** Assume all students surveyed make less than $250K.

(a) (2 pt) The percent of new grads who have a salary of $100K-$150K (not including $50K).

```
0.9 % per $1K * $50K
```

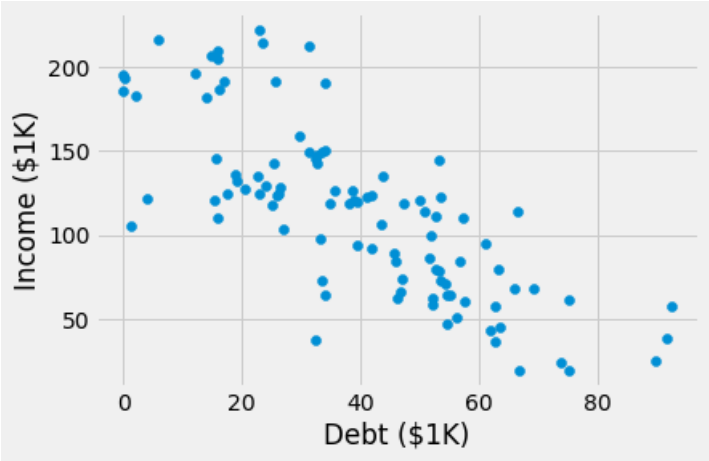
(b) (3 pt) The height of Bin 2 in the histogram above, assuming 50 students have a salary of $50K-$100K (not including $100K).

```
50/200 * 100% / $50K
```

(c) (3 pt) The number of survey respondents who make less than $50K.

```
(0.26% per $1K) / 100 % * 200 * $50K
```

Suppose Cindy wants to understand how a graduate's debt might be associated with their salary, so she makes the following scatterplot:



(d) (0 pt) Which of the following are valid conclusions that can be drawn from this graph above? **Choose all that apply.**

- ○ There is a positive association between student debt and salary.
- ● There is a negative association between student debt and salary.
- ○ There is no association between student debt and salary.
- ○ There are no Berkeley graduates with a debt greater than $100K.
- ○ There are no Berkeley graduates with a salary greater than $200K who also have a debt greater than $40K.
- ○ There are more Berkeley graduates with high debt (greater than $60K) than with small debt (less than $60K).
- ● Among the graduates surveyed, 3 of them have debt greater than $80K.
- ○ Among the graduates surveyed, higher debt caused them to have lower starting salaries.

(e) (4 pt) Suppose Cindy wants to add a new column called `'interned'` to her table `graduates` that contains whether the students interned during their undergraduate studies (e.g., the first three rows would be `True, True, False`). Fill in the blanks below to create and add the column to the table.

```
internships = graduates.column('internships')
graduates = graduates.with_column('interned', internships > 0)
```

(f) (6 pt) Suppose Cindy wants to now create a table showing the average salary for each combination of `'interned'` and `'major'`. You may assume that the `'interned'` column was defined correctly and added to the `graduates` table. Write two separate lines of code that are two ways to create such a table. The two separate lines of code must use two different table methods. Using `np.mean` and `np.average` does not count as two different ways.

First way

```
graduates.pivot('interned', 'major', 'salary', np.average)
```

Second way

```
graduates.select('interned', 'major', 'salary').group(
['interned', 'major'], np.mean)
```

3. (36 points) Crossroads

Shelly, a Nutrition Science major, is studying the nutrition in the campus dining halls. She randomly samples 500 items served at Crossroads in February 2020 and puts the sampled items into a table `menu` that contains one row for each item. Here are the first few rows:

item	category	priority	calories	plates
Herbed Pasta Salad	Salad Bar	Low	150	10,421
Raspberry Sammie	Desserts	High	200	20,497
Lobster Bisque	Soups	Medium	240	4,249

The `menu` table contains five columns:

- **item**: a string, the name of the item on the menu
- **category**: a string, the food category of the item
- **priority**: a string, the priority that Cal Dining puts on making the item (`'Low'`, `'Medium'` or `'High'`)
- **calories**: an int, the number of calories per plate
- **plates**: an int, the number of plates served to students in the month

She also creates a table `allocations` that contains one row for each food category in `menu`. Here are the first few rows:

category	availability	budget
Salad Bar	Both	220,000
Desserts	Dinner	250,000
Soups	Both	300,000

The `allocations` table contains three columns:

- **category**: a string, the name of the food category
- **availability**: a string, when the category is offered on the menu (`'Brunch'`, `'Dinner'` or `'Both'`)
- **budget**: an int, the Cal Dining budget for the food category (in dollars)

Shelly wants to answer the following questions. For parts (a) - (d), choose which kind of visualization would be the best choice to answer it. **Choose only one answer for each question.**

(a) (2 pt) Are there more soup items than salad bar items?

○ Line graph ○ Histogram ● Bar chart ○ Scatter plot

(b) (2 pt) Do dessert items tend to be of high priority or low priority?

○ Line graph ○ Histogram ● Bar chart ○ Scatter plot

(c) (2 pt) How do the distributions of calories per plate vary between the high priority and low priority desserts?

○ Line graph ● Histogram ○ Bar chart ○ Scatter plot

(d) (2 pt) Is there an association between calories per plate and the number of plates served?

○ Line graph ○ Histogram ○ Bar chart ● Scatter plot

For parts (e)-(j) below, fill in the blanks of the Python expressions. **You must use ONLY the lines provided.** Some of the chained operations we might normally do in one line have been broken up into two or more lines, storing intermediate results in temporary tables. You may find the names of the temporary tables to be useful hints. Do not write any code outside the blanks provided. The expression in the last line should evaluate to the value described.

(e) (2 pt) The average number of calories of all items.

```
np.mean(menu.column('calories'))
```

(f) (4 pt) The total number of plates served of all items with over 250 calories per plate.

```
num_plates = np.sum(

    menu.where('calories', are.greater_than(250))

    .column('plates')

)

num_plates
```

(g) (4 pt) A function that takes in a menu category (as a string) and returns an array containing the proportion of items in that category that are High, Low, and Medium priority, in that order.

Hint: High, Low, Medium is in alphabetical order.

```
def priority_proportions(menu_category):

    priority_dist = tbl.where('category', menu_category).group('priority')

    priority_counts = priority_dist.sort('priority').column('count')

    return priority_counts / np.sum(priority_counts)
```

(h) (4 pt) The total variation distance between the priority distributions of dessert and salad items. You may use the function defined in part (g) and can assume that it was defined correctly.

```
dessert_props = priority_proportions('Dessert')

salad_props = priority_proportions('Salad Bar')

0.5 * np.sum(abs(dessert_props - salad_props))
```

(i) (5 pt) The most popular category at Crossroads (i.e., the one whose items, when combined, have been served the highest number of plates).

```
two_cols = menu.select('category', 'plates')

dist = two_cols.group('category', sum)

dist = dist.sort('plates sum', descending=True)

dist.column('category').item(0)
```

(j) (6 pt) The number of calories per plate of the highest-calorie item served only at brunch.

```
croads = menu.join('category', allocations)

num_calories = np.max(

    croads.where('availability', 'Brunch')

    .column('calories')

)
 num_calories
```

(k) (3 pt) In 2018, the Partnership for Healthier America (PHA) started the Healthier Campus Initiative to foster better nutrition in dining halls. One of the PHA recommendations through this initiative is that the distribution of food categories offered on the menu should be 20% salad, 20% soups, 10% desserts, and 50% entrees. Shelly wants to use her random sample of Crossroads menu items to test whether Crossroads has been following this PHA recommendation. Provide a null and alternative hypothesis Shelly could use to answer her question.

Null hypothesis

The crossroads menu items have the same distribution of categories as what the PHA recommends.

Alternative hypothesis

The crossroads menu items does not have the same distribution of categories as what the PHA recommends.

4. (21 points) Lemons

Isabelle, a lemon farmer, noticed that her crop yield this season was slightly lower than she expected. The USDA recently announced that it discovered a new lemon tree disease; they claim that an early symptom of the disease is discoloration of the leaves, and that the disease kills 2% of lemon trees. Because her crop yield was still roughly the same, Isabelle believes that the actual fatality rate of the tree disease is less than 2%.

In other to test the USDA's claim, she randomly selects 1,000 of her lemon trees for data collection. She prepares a table `lemons`, containing 1,000 rows, one for each tree. Here are the first few rows:

treeage	disease	discolored	dead
5	True	False	False
12	False	False	True
7	True	True	True

The table contains four columns:

- **treeage**: an int, the tree's age (in years)
- **disease**: a boolean, whether the tree has the disease or not
- **discolored**: a boolean, whether the tree leaves are discolored
- **dead**: a boolean, whether the tree is dead or not

Suppose we also know the following:

- among the sampled trees, 501 tested positive for the disease
- among the sampled trees that tested positive for the disease, 8 died

This implies that the observed fatality rate is 1.60%.

(a) (3 pt) Write down valid null and alternative hypotheses for the test that Isabelle wants to carry out below.

Null hypothesis

The fatality rate of the tree disease is 2%.

Alternative hypothesis

The fatality rate of the tree disease is less than 2%.

(b) (4 pt) As a first step to testing the USDA's claims, Isabelle wants to simulate 10,000 values of the test statistic under the null hypothesis. She chooses the fatality rate of diseased trees as the test statistic. Which of the following expressions will help her simulate one value of the test statistic? **Choose only one.**

○ `np.random.choice(lemons.where('disease', True).column('dead'), 501)`

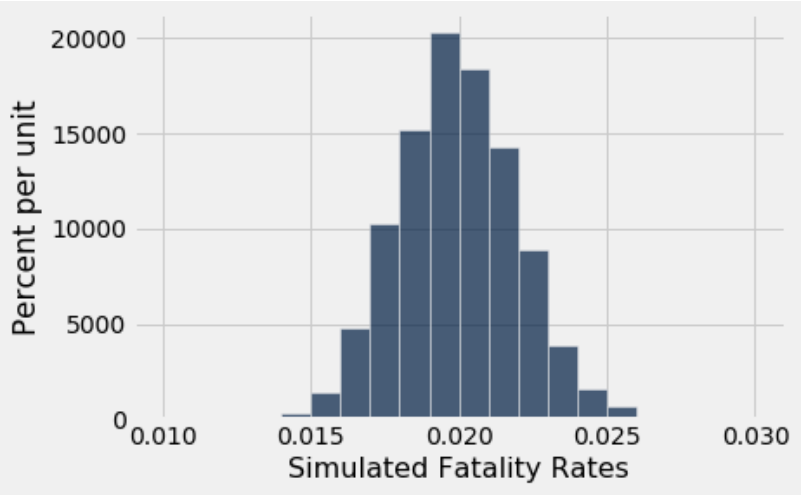
○ `lemons.where('disease', True).sample(501).column('dead')`

○ `np.count_nonzero(lemons.where('disease', True).sample(501).column('dead'))`

● `sample_proportions(501, (0.02, 0.98)).item(0)`

○ `sample_proportions(501, (0.016, 1 - 0.016)).item(0)`

(c) (4 pt) Suppose Isabelle repeats the procedure in part (b) 10,000 times and she plots a histogram of the simulated test statistics. Assume the histogram shows all of the simulated values.



Given that the observed fatality rate is 1.60%, which of the following are valid conclusions from this histogram? **Choose all that apply.**

- ○ The fatality rate in Isabelle's sample is due to chance alone
- ● There is evidence against the claim that the disease fatality rate is 2%
- ○ There is evidence for the claim that the disease fatality rate is 2%
- ○ The disease fatality rate is actually 1.60%
- ○ The disease fatality rate is actually less than 2%

Suppose that Isabelle's friend, Noah, sprayed fertilizer on random trees that were diseased, hoping that they would recover. She adds a new column `'fertilizer'` (whose rows are either `'True'` or `'False'`), and wants to do an A/B test to see whether the fatality rate differed between trees that got fertilizer and those that did not.

(d) (6 pt) The alternative hypothesis that Isabelle will use is given below; write down a valid null hypothesis and a valid test statistic that corresponds to the null and alternative hypothesis.

Isabelle's Alternative hypothesis: The fatality rate of diseased trees that were sprayed with fertilizer is lower than that of diseased trees that did not receive fertilizer.

Null hypothesis

The fatality rate of diseased trees that were sprayed with fertilizer is the same as that of diseased trees that did not receive fertilizer

Test Statistic

Fatality rate of diseased trees with fertilizer minus fatality rate of diseased trees without fertilizer.

(e) (4 pt) Why should Isabelle shuffle the `'fertilizer'` column in order to carry out her A/B test? **Choose all that apply.**

- ● under the null hypothesis, the label of being in treatment or control doesn't matter
- ● she needs to simulate two groups of trees such that their expected disease fatality rate is identical under the null hypothesis
- ○ she needs to ensure that the trees in the experiment are selected randomly
- ○ she needs to randomize treatment & control to establish causation

5. (0 points) Write your name in the space provided on one side of every page of the exam. You're done!