# Data 8 Summer 2019 — Foundations of Data Science

## Midterm Exam Solutions

### INSTRUCTIONS

- The exam is worth 80 points. You have 110 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`) | |
| Name of Lab GSI | |
| Your seat number (e.g. A1) & 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)** | |

**This page is intentionally left blank. You can use it for scratch work but it will not be graded.**

**1. (22 points) Get Off Your Phone**

A digital media company commissions a Berkeley Data Science club to conduct an analysis about how much time people spend on their phones. The club surveyed 500 people and collected the resulting data in the table `phones` which has 5 columns:

- `Name`: The person's name
- `Model`: The type of phone the person has, an 'iPhone', 'Samsung Galaxy', 'Google Pixel', or 'Non-smartphone'
- `Screentime`: The average daily minutes that person spends on their phone
- `Data (MB/day)`: The average daily cell data usage of that person, in megabytes
- `Provider`: The person's cell phone carrier: 'AT&T', 'Verizon', 'Sprint', or 'T-Mobile'

Here are the first 5 rows of the table, which has 500 rows total:

| Name | Model | Screentime | Data (MB/day) | Provider |
|---|---|---|---|---|
| Tam | iPhone | 180.3 | 50 | Verizon |
| Adeel | iPhone | 300 | 22 | AT&T |
| Maddy | Non-smartphone | 65 | 10.4 | T-Mobile |
| Tanay | Samsung Galaxy | 108.9 | 45 | Sprint |
| Jiayi | Google Pixel | 199 | 33.6 | Verizon |

**Fill in the blanks with Python expressions to compute the desired output.** ONLY use the blank 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 tables or arrays. Do not write any code outside the blanks provided. The expression in the last line should evaluate to the value described in the question.

**(a) (2 pt)** Write code to assign `least_screen` to the smallest value of daily screentime in the `phones` table.

```
least_screen = min(phones.column("Screentime"))
```

**(b) (3 pt)** Write code that assigns `num_iphones` to the number of people with iPhones in the `phones` table.

```
iphones = phones.where("Model", are.equal_to("iPhone"))
num_iphones = iphones.num_rows
```

**(c) (4 pt)** Write code that assigns `high_provider` to the cell phone provider of the person with the highest data usage. You can assume that data usages are unique in the `phones` table.

```
sorted = phones.sort("Data MB/day", descending=True)
high_provider = sorted.column("Provider").item(0)
```

**(d) (3 pt)** Write code that assigns `average` to a table with two columns: one containing all of the phone models, and another containing the average screentime for each model.

```
average = phones.group("Model", np.mean).select("Model", "Screentime mean")
```

**(e) (6 pt)** Later on in the project, a club member creates a table with information about cell phone providers and data costs. The `providers` table has two columns:

- `Company`: The name of the cell phone provider (AT&T, Verizon, Sprint, or T-Mobile)
- `Price ($/MB)`: The price the company charges per megabyte of data used.

| Company | Price ($/MB) |
|---|---|
| Verizon | 0.05 |
| AT&T | 0.1 |

... (3 rows omitted)

Write code to generate a table called `costs` which contains two columns:

- `Name`: The name of the person in the study
- `Monthly Cost`: The price the person has to pay for 30 days of data usage. Assume that every day, the person used the usage amount listed in the `phones` table.

```
all_info = phones.join("Provider",providers,"Company")

daily_usage = all_info.column("Data MB/day")

prices = all_info.column("Price $/MB")

daily_cost = daily_usage * prices

monthly_cost = daily_cost * 30

costs = Table().with_columns(

            'Name', all_info.column("Name"),

            'Monthly Cost', monthly_cost)
```

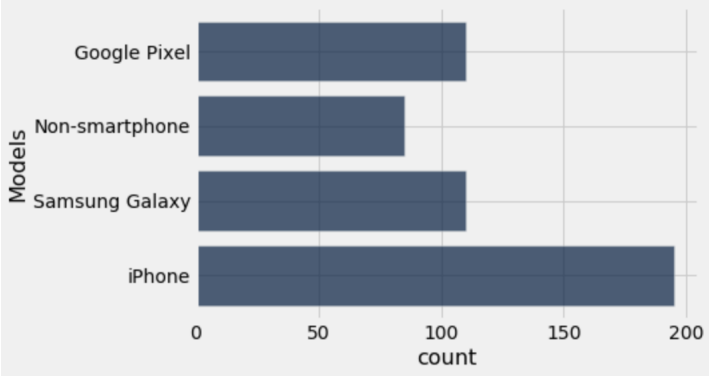
**(f) (2 pt)** Write code so that `summary` results in the following table:

| Provider | Google Pixel | Non-smartphone | Samsung Galaxy | iPhone |
|---|---|---|---|---|
| AT&T | 21 | 18 | 18 | 35 |
| Sprint | 12 | 20 | 26 | 39 |
| T-Mobile | 18 | 20 | 26 | 42 |
| Verizon | 47 | 35 | 38 | 85 |

The value in each cell represents the number of people in the survey that had the corresponding phone/provider combination.

```
summary = phones.pivot("Model", "Provider")
```

**(g) (2 pt)** The following is a plot displaying the count of each phone model in the `phones` table:




Which line of code generated the above plot?

- ◯ `phones.barh('Model')`
- ● `phones.group('Model').barh('Model')`
- ◯ `phones.hist('Model')`
- ◯ `phones.group('Model').hist('Model')`

**2. (8 points) Street Smarts**

In 2015, economists Melissa Kearney and Phillip Levine published a study investigating the effects of watching *Sesame Street* on the future academic performance of young viewers in the early 1970s. *Sesame Street* is a popular childrens show that was first introduced in 1969 and provides free educational content on public television to children who are too young to attend school.

However, there was variation in exposure to *Sesame Street* across the country, due to technological constraints. Approximately 1/3 of US viewers lived in counties where they were unlikely to be able to view *Sesame Street*. Whether or not a county had high rates of access to *Sesame Street* was close to random and was not associated with any particular characteristics of the county, such as income.

Since the researchers did not have data on whether individual children watched *Sesame Street*, they instead investigated county-level elementary-school success metrics of groups of children who started school after the show debuted and who lived in locations where broadcast reception for the show was high, and compared these metrics to those among older groups of children (who would not have watched *Sesame Street* before starting school) and those who lived in locations with limited broadcast reception (poor likelihood of access to *Sesame Street*).

The authors found that "children who lived in places with better access to the show did better [on average] in elementary school, as compared to those with limited access and those who were older at the time the show was introduced. They were more likely to start school on time and progress at the appropriate grade for age."

**(a) (2 pt)** Who were the individuals in this study?

◯ Individual children who started school after *Sesame Street* debuted in the United States

◯ Individual children who started school just before or after *Sesame Street* debuted in the United States

● Groups of children in different counties in the United States

◯ Groups of children in different counties with high likelihood of access to *Sesame Street* in the United States

**(b) (2 pt)** The authors made two comparisons:

- They compared the educational outcomes of groups of children in counties with high likelihood of access to *Sesame Street* to the educational outcomes of groups of children with low likelihood of access to *Sesame Street*.
- They compared the educational outcomes of groups of children in counties with high access to *Sesame Street*, and who started school after *Sesame Street* debuted, to groups of children **in those same high-access counties** who started school before *Sesame Street* debuted on television.

Why did the authors include both comparisons in their study?

◯ The authors thought children may have moved from one county to another, so they wanted to study the children who moved by looking at the same group over time.

◯ By studying the same children before and after they watched *Sesame Street*, the authors could determine whether *Sesame Street* was the reason that some individual students performed better than others.

◯ The authors wanted to include more children in their study (a larger sample size) so that the children would be more representative of the population of all children in the United States, instead of just looking at one age group.

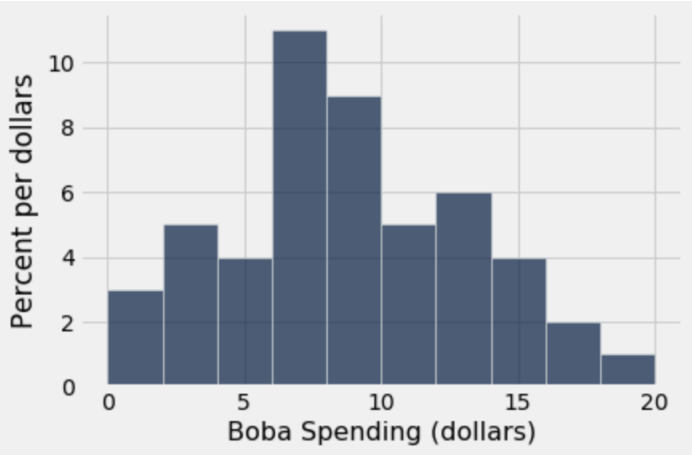
● Making both comparisons helps eliminate the possibility of confounding factors due to broadcast variability (which counties had *Sesame Street* access) or variation in educational quality across years (if education in general improved or got worse over the study period).

**(c) (4 pt)** Select **all** correct statements:

- [x] This was an observational study.
- [ ] This was a randomized controlled experiment.
- [ ] It is only possible to investigate a causal link with a randomized controlled experiment, and this study did not help establish a causal link between *Sesame Street* and better educational outcomes.
- [ ] Whether or not a child lived somewhere with access to *Sesame Street* was random, because the researchers split the groups of children into a treatment and a control group.
- [x] The authors were attempting to use a natural experiment, the variation in TV signals, to establish a causal link between watching *Sesame Street* and better educational outcomes.

**3. (12 points) Boba Distributions**

Anna is curious about her boba spending habits. The following is a histogram of her weekly spending on boba, over the span of **50** weeks (she doesnt drink boba on her yearly two week vacation):




The spending is divided into bins of width 2: [0,2), [2,4) ... [18,20).

Assume that all of the data is shown on the histogram (Anna never spent $20 or more in a given week).

For the following questions, write a **mathematical expression** for the answer, or write *Not possible* if it is not possible to calculate the desired quantity with the information in the given histogram.

**(a) (3 pt)** In what percent of weeks did Anna spend $16 or more on boba?

2*2 + 1*2 = 6%

**(b) (3 pt)** In what percent of weeks did Anna spend between $12 and $15 on boba?

Not possible to determine

**(c) (3 pt)** During how many weeks did Anna spend between $10 and $13 dollars?

- ○ 11 weeks
- ● Between 5 and 11 weeks
- ○ 22 weeks
- ○ Between 10 and 22 weeks
- ○ Not possible to calculate using the histogram

**(d) (3 pt)** After looking over her notes, Anna finds that she entered the wrong data for two of the weeks when constructing the histogram. She is not sure whether her spending in those weeks should go in the [12, 14) or [14, 16) bin, so she decides to combine all of the data in those two bins into one bin of width 4, from [12,16). What is the height of this new bin?

(0.6*2 + 0.4*2)/4 = 5 %/$

**4. (10 points) Steph Curry Fan Club**

A group of current Data 8 students who are big fans of Steph Curry are inspired by the current course material to conduct an analysis of Curry's basketball shot record. It turns out Steph Curry has a 58% chance of successfully shooting the ball in the basket any time he tries.

**(a) (2 pt)** Curry shoots the ball twice. What is the probability that both of the shots make it in the basket?

- ○ $\frac{58}{100} \times \frac{58}{99}$
- ● $0.58^2$
- ○ $2 \times 0.58$
- ○ $0.42^2$

**(b) (2 pt)** Curry shoots the ball twice. What is the probability that one shot makes it into the basket, and one shot does not?

- ○ $(0.58 \times 0.42)^2$
- ○ $0.58 \times 0.42$
- ● $(0.58 \times 0.42) + (0.42 \times 0.58)$
- ○ $1 - (0.58)^2$

Digging deeper into Steph Curry's basketball record, the students find that 80% of Currys shots are contested - that is, someone is actively trying to block his shot or make him miss. When a shot is contested, Curry has a 55% chance of successfully shooting the ball in the basket any time he tries. When his shots are not contested, Curry has a 70% chance of shooting the ball in the basket.

**(c) (2 pt)** Given that Steph Curry's shot was uncontested, what is the probability the shot did not make it in the basket?

- ○ $\frac{0.2 \times 0.3}{(0.2 \times 0.3 + 0.8 \times 0.45)}$
- ○ $0.42$
- ● $0.30$
- ○ $(0.2 \times 0.3) + (0.8 \times 0.45)$

**(d) (2 pt)** Given that Steph Curry took three contested shots in a row, what is the probability that at least one of them went in?

- ● $1 - (0.45)^3$
- ○ $1 - (3 \times 0.45)$
- ○ $\frac{0.80 \times 0.55}{[(0.80 \times 0.55) + (0.20 \times 0.70)]^3}$
- ○ $0.55^3$

**(e) (2 pt)** Given that Steph Curry made the shot, what is the probability the shot was contested?

- ○ $\frac{0.20 \times 0.70}{((0.80 \times 0.55) + (0.20 \times 0.70))}$
- ○ $1 - (0.45)^3$
- ○ $0.8$
- ● $\frac{0.80 \times 0.55}{((0.80 \times 0.55) + (0.20 \times 0.70))}$

**5. (28 points) Issues with Influenza**

The national news reports widespread hospitalizations from flu this year in the United States. Since you don't remember the media reporting on this in the past, you want to know if this year's incidence rate of flu (the proportion of the population who become infected with influenza) in the U.S. is higher than usual. Since you only have the resources to take a random survey of 1000 people, you decide to conduct a hypothesis test to find out.

After conducting your random survey of 1000 people, you find that the incidence rate among this sample is 0.35. In modern history (since the invention of the influenza vaccine), the expected incidence rate is 0.2.

**(a) (3 pt)** What is the appropriate null hypothesis?

- ○ Our sample is large enough to represent the US population.
- ○ The true incidence rate is 0.35, and 0.2 is an inaccurate estimate of historical incidence rates.
- ○ The true incidence rate of influenza this year is greater than 0.2, and our survey incidence rate is representative of the true incidence rate.
- ● The true incidence rate of influenza this year is 0.2, and any deviation from this rate is due to chance in the selection of the random sample.

**(b) (3 pt)** What is the appropriate alternative hypothesis?

- ○ Our sample was not representative of the US population.
- ○ The influenza vaccine is not very effective this year.
- ● The true incidence rate of influenza this year is greater than 0.2.
- ○ The true incidence rate of influenza this year is less than 0.2.

**(c) (3 pt)** What is an appropriate test statistic?

- ○ Number of influenza cases in our sample
- ○ Population proportion of influenza cases
- ● Sample proportion of influenza cases

**(d) (8 pt)** Write code below so that `test_proportions` evaluates to 10,000 simulated values of your test statistic under the null hypothesis.

```
expected_props = make_array(0.2,0.8)
sample_size = 1000

def simulated_proportion(expected_props, sample_size):

        null_props = sample_proportions(sample_size,expected_props)

        prop_sick = null_props.item(0)

        return prop_sick


test_props = make_array()

for i in np.arange(10000):

        one_statistic = simulated_proportions(expected_props, sample_size)

        test_props = np.append(test_props, one_statistic)
```

**(e) (3 pt) Select one of the options from parts i-iii to fill in the corresponding blanks in the sentence below.**

To calculate the p-value for this test, we would find the proportion of ___(i)___ under the ___(ii)___ that were ___(iii)___ our observed test statistic.

i. ○ observed test statistic ● simulated test statistics ○ years in the past century

ii. ● null hypothesis ○ alternative hypothesis

iii. ○ equal to ○ less than or equal to ● greater than or equal to

**(f) (2 pt)** You find your p-value to be 0.04. Under which of these p-value cutoff level(s) would you **fail** to reject the null? Select all that apply.

○ 0.10 ○ 0.07 ○ 0.05 ● 0.03 ● 0.01

**(g) (3 pt)** With a 5% p-value cutoff and your p-value of 4%, and considering your original hypothesis, would you conclude that the influenza vaccine was not very effective this year?

○ No, because you did not reject the null hypothesis.
○ Yes, because you found a significant difference in the influenza rates.
● No, because the null hypothesis was about the incidence rate, not vaccine efficacy.
○ Yes, because your p-value was above 0%.

**(h) (3 pt)** You are discussing your survey methodology with a friend and reveal that the population from which your survey participants were selected was the population of all Berkeley undergraduates. Assuming you carried out the steps in parts e-f, would your method successfully test the null hypothesis that the incidence rate is higher than expected in the U.S. population? Select one option.

○ Yes, because you selected students at random for your survey.
○ Yes, because your p-value was small.
○ No, because 1000 individuals is not a big enough sample to run a hypothesis test.
● No, because your sample is not representative of the US population.

**6. (0 points)** Write your name in the space provided on one side of every page of the exam. You're done!