S&DS 123 YData: An Introduction to Data Science
Spring 2021 Midterm Solutions

INSTRUCTIONS

- You have 60 minutes to complete the exam, once you start.
- The exam is open book, open notes, open computer.
- **It is strictly forbidden to communicate with other people about the exam during the 24 hour period when the exam is available.**

First name	
Last name (Surname)	
NetID	

Problem	Score	Out of
1		9
2		9
3		18
4		6
Total		42

1. **Tables (9 points)**

In class, we illustrated various concepts using the New York Times Covid-19 Database, a publically-available database of confirmed cases and deaths, compiled from state and local governments and health departments across the United States.

Two tables `ct_data` and `ma_data` are constructed from these data by taking the data on a single day, showing the cumulative cases and deaths from Covid-19 in Connecticut and Massachusetts. This is the table `ct_table`:

date	county	state	cases	deaths
2021-03-23	Fairfield	Connecticut	86392	2098
2021-03-23	Hartford	Connecticut	73835	2339
2021-03-23	Litchfield	Connecticut	12195	282
2021-03-23	Middlesex	Connecticut	11216	349
2021-03-23	New Haven	Connecticut	77638	1986
2021-03-23	New London	Connecticut	20189	419
2021-03-23	Tolland	Connecticut	8356	177
2021-03-23	Windham	Connecticut	9692	185

And the table `ma_table` looks like this:

date	county	state	cases	deaths
2021-03-23	Barnstable	Massachusetts	11395	430
2021-03-23	Berkshire	Massachusetts	5211	269
2021-03-23	Bristol	Massachusetts	58392	1602
2021-03-23	Dukes	Massachusetts	894	0
2021-03-23	Essex	Massachusetts	87487	2255
2021-03-23	Franklin	Massachusetts	2128	104
2021-03-23	Hampden	Massachusetts	44466	1422
2021-03-23	Hampshire	Massachusetts	8087	279
2021-03-23	Middlesex	Massachusetts	118787	3587
2021-03-23	Nantucket	Massachusetts	1243	0

... (4 rows omitted)

Provide Python expressions to compute the values described below. You can assume the statements `from datascience import *` and `import numpy as np` have been executed.

(a) The name (as a Python string) of the Massachusetts county having the largest number of cases.
ma_data.sort('cases', descending=True).column(1).item(0)

(b) The assignment to the name `ct_ma_data` of a table that combines the data for both Connecticut and Massachusetts, looking like this:

date	county	state	cases	deaths
2021-03-23	Fairfield	Connecticut	86392	2098
2021-03-23	Hartford	Connecticut	73835	2339
2021-03-23	Litchfield	Connecticut	12195	282
2021-03-23	Middlesex	Connecticut	11216	349
2021-03-23	New Haven	Connecticut	77638	1986
2021-03-23	New London	Connecticut	20189	419
2021-03-23	Tolland	Connecticut	8356	177
2021-03-23	Windham	Connecticut	9692	185
2021-03-23	Barnstable	Massachusetts	11395	430
2021-03-23	Berkshire	Massachusetts	5211	269

... (12 rows omitted)

ct_ma_data = ct_data.append(ma_data)

(c) The reassignment to `ct_ma_table` of a table with an additional column, which is the number of deaths divided by the number of cases in each county, looking like this:

date	county	state	cases	deaths	deaths per case
2021-03-23	Fairfield	Connecticut	86392	2098	0.024
2021-03-23	Hartford	Connecticut	73835	2339	0.032
2021-03-23	Litchfield	Connecticut	12195	282	0.023
2021-03-23	Middlesex	Connecticut	11216	349	0.031
2021-03-23	New Haven	Connecticut	77638	1986	0.026
2021-03-23	New London	Connecticut	20189	419	0.021
2021-03-23	Tolland	Connecticut	8356	177	0.021
2021-03-23	Windham	Connecticut	9692	185	0.019
2021-03-23	Barnstable	Massachusetts	11395	430	0.038
2021-03-23	Berkshire	Massachusetts	5211	269	0.052

... (12 rows omitted)

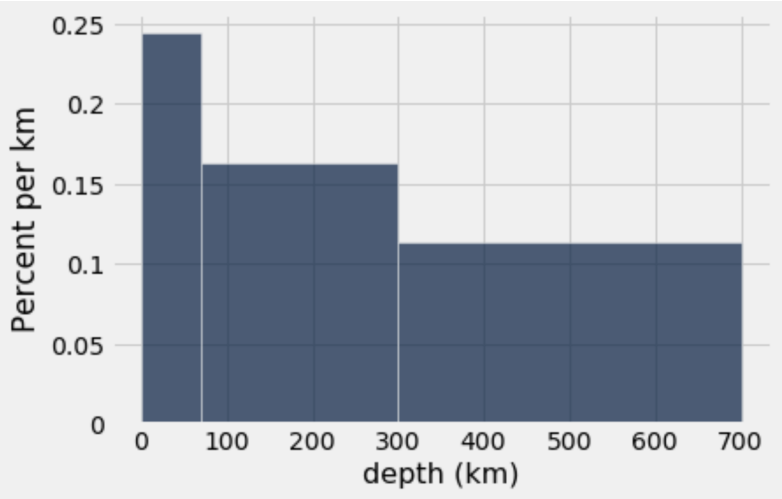
ct_ma_data = ct_ma_data.with_column('deaths per case', np.round(ct_ma_data['deaths']/ct_ma_data['cases'],3))

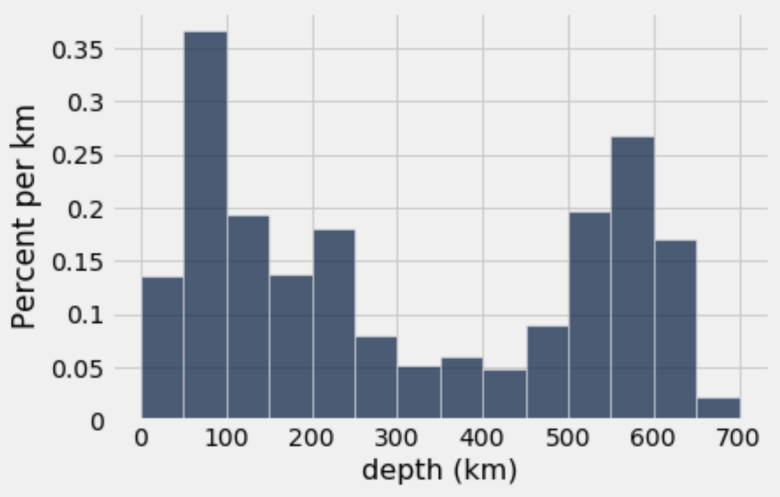
2. **Histograms (9 points)**

Earthquakes are relatively common in Fiji. Below are two histograms, each summarizing the same database of earthquakes near Fiji since 1964. Earthquakes can occur at a depth between the Earth's surface and about 700 kilometers below the surface. This earthquake depth range of 0–700 km is commonly divided into three zones: shallow, intermediate, and deep. Shallow earthquakes are less than 70 km deep, intermediate earthquakes are between 70 and 300 km deep, and deep earthquakes are between 300 and 700 km deep.

The histograms below were generated as follows:

```
quakes = Table.read_table("earthquakes.csv")
quakes.hist("depth", bins=make_array(0, 70, 300, 700), unit="km")
quakes.hist("depth", bins=np.arange(0, 750, 50), unit="km")
```





Give estimates for the following quantities by inspecting the histograms—do not write Python expressions to evaluate the quanties from the data. Show your work.

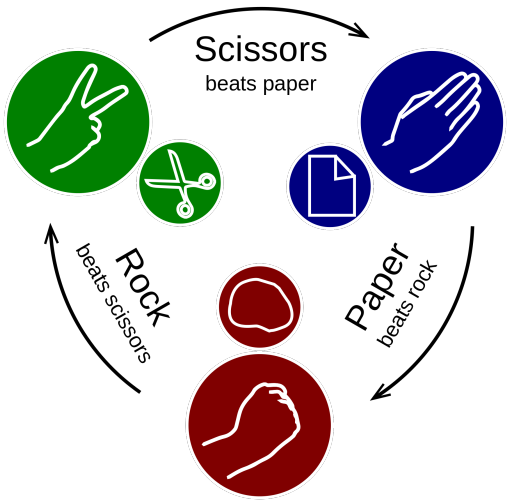
(a) The percentage of earthquakes that are shallow.
$70 \cdot 0.245 = 17.15$, so 17.15% of earthquakes are shallow.

(b) The percentage of earthquakes with depth greater than or equal to 70 km and less than 100 km.
0 to 100km : $50 \cdot 0.13 + 50 \cdot 0.36 = 24.5$
0 to 70km: $70 \cdot 0.24 = 16.8$
70 to 100km: $24.5\% - 16.8\% = 7.7\%$

(c) The chance that a new earthquake will be at least 600 km deep.
$50 \cdot 0.17 + 50 \cdot 0.02 = 9.5$, so there is 9.5% chance that a new earthquake will be at least 600 km deep.

3. **Probability and Hypothesis Testing (18 points)**

The game "Rock, Paper, Scissors" is a popular way for children to make randomized decisions. There are two players in the game. In each round, they simultaneously show one of three hand configurations: rock (closed fist), paper (flat open hand), or scissors (V-shape with two fingers). The rules of the game are that if the players show the same sign, the round is a draw (a tie). Otherwise, paper wins over rock, scissors wins over paper, and rock wins over scissors. These outcomes are shown below:



en.wikipedia.org/wiki/Rock_paper_scissors

We'll call the players A and B. Assume that each of player A and player B plays randomly, playing rock, paper, or scissors with equal probability in each round.

(a) What is the chance that a given round is a tie?

Each pair of hands has probability 1/9. There are three ways of having a draw: rock-rock, paper-paper, and scissors-scissors. So the probability is 1/3.

(b) What is the chance that player A wins a given round?

Each hand player A throws beats exactly one hand that B throws. So the probability that A wins is $3 \cdot 1/9 = 1/3$.

(c) What is the chance the game is tied after three rounds? (The game is tied if both players have won the same number of rounds.)

If • indicates a draw, `W` indicates a win for A, and `L` indicates a loss for A, then the possible outcomes that end in a tie after three rounds are • • •, • `W` `L`, • `L` `W`, `W` • `L`, `L` • `W`, `W` `L` •, `L` `W` •. Each of these has probability $1/27 = (1/3)^3$. So, the overall probability is 7/27.

Data are collected for a game of 100 rounds between two players. The data look like this:

A plays	B plays	Outcome
scissors	scissors	Draw
rock	scissors	A wins
scissors	rock	B wins
scissors	rock	B wins
scissors	scissors	Draw
paper	rock	A wins
paper	scissors	B wins
scissors	rock	B wins
paper	rock	A wins
scissors	rock	B wins

... (90 rows omitted)

The number of times the players together play each of the 9 possibilities is this:

B plays	paper	rock	scissors
paper	5	10	8
rock	13	16	17
scissors	9	14	8

For example, in 10 of the 100 rounds, player A played rock while player B played paper.

You would like to study whether the players are playing randomly, or if they are using a non-random strategy.

(d) What is the null hypothesis?

The players are playing randomly, with each of the nine possible outcomes having equal probability 1/9.

(e) What is the alternative hypothesis?

The players are not playing randomly; some other mechanism is behind the data.

You decide to run a hypothesis test by simulating random play and comparing it to the observed data. You use total variation distance as your test statistic.

Complete the code below, by entering the Python expression for each line containing a blank.

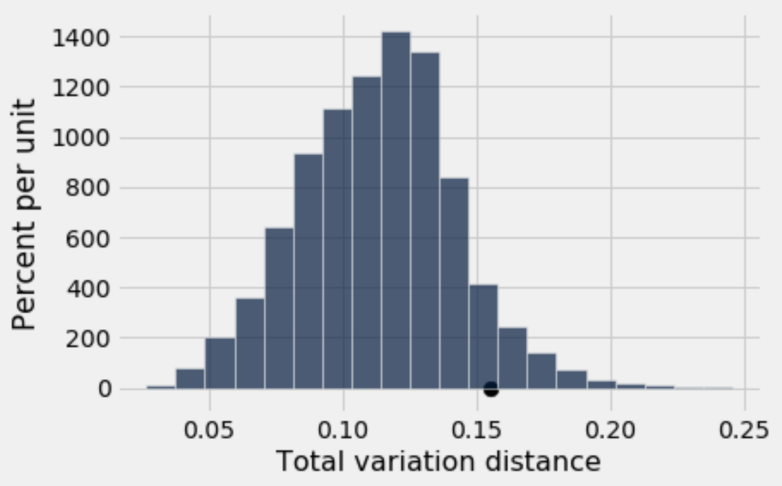
```
1   def total_variation_distance(distribution_1, distribution_2):
2       return sum(np.abs(distribution_1 - distribution_2)) / 2

3   observed_distribution = make_array(5, 10, 8, 13, 16, 17, 9, 14, 8)/100
4   null_distribution = np.ones(9)/9
5   observed_statistic = total_variation_distance(observed_distribution, null_distribution)

6   tvds = make_array()
7   for i in np.arange(10000):
8       sample_distribution = sample_proportions(100, null_distribution)
9       tvds = np.append(tvds, total_variation_distance(sample_distribution, null_distribution))

10  Table().with_column("Total variation distance", tvds).hist(bins=20)
11  plots.scatter(observed_statistic, 0, color="black", s=75)

12  p_value = sum(tvds >= observed_statistic)/10000
```



(f) Line 3: [see code above]

(g) Line 4: [see code above]

(h) Line 8: [see code above]

(i) Line 9: [see code above]

(j) Line 12: [see code above]

(k) The histogram of the sample statistics is shown above; the observed test statistic is shown as a large dot. The P-value is 0.0847. What do you conclude—do the data better support the null hypothesis or the alternative? Explain your answer.

Assuming the null hypothesis is true, the probability of observing a statistic equal to or larger than the observed statistic is 0.0847. Since this is larger than 0.05, there is insufficient evidence to reject the null hypothesis at the 5% significance level.

4. **Causality (6 points)**

Many people are trying to reduce their egg consumption, believing that consuming large quantities of eggs leads to high cholesterol levels.

To statistically test this theory, two groups of researhers conduct independent studies: study A and study B.

In study A, researchers asked 123 healthy adults, aged 21 to 67, about their dietary preferences, namely the amount of eggs each participant consumes on average per day. Other dietary preferences that might affect the cholesterol level were also recorded and accounted for. A total of 54 participants reported that they consume less than one egg a day, while the other 69 participants reported that they eat at least two eggs per day. The results showed that there is a significant difference in cholesterol levels between the two groups, with the group that ate more eggs having higher cholesterol.

In study B, researchers randomly assigned 123 healthy adults, aged 21 to 67, to one of two diet programs. Participants in the first program were told to consume at least three eggs daily while the participants from the second program were limited to only one egg per day. The results showed that there is no significant difference in cholesterol levels between the two groups.

Select all that apply:

(a) Based on the results of study A, we conclude that higher egg consumption causes higher level of cholesterol.

✓ (b) Based on the results of study A, we conclude that there is an association between higher egg consumption and higher levels of cholesterol.

✓ (c) Based on the results of study A, we cannot conclude that higher egg consumption causes higher levels of cholesterol, as there might be confounding factors.

✓ (d) Based on the results of study B, we cannot conclude that there is an association between higher egg consumption and higher levels of cholesterol.

(e) Study A is a randomized control trial.

✓ (f) Study B is a randomized control trial.