DATA 8 Foundations of Data Science

Summer 2022 — Midterm Exam

INSTRUCTIONS

You have 1 hours and 50 minutes to complete the exam. This exam is worth 90 points.

- The exam is closed book, closed notes, closed computer/calculator, except the provided midterm reference sheet.
- Mark your answers on the exam itself in the spaces provided. We will not grade answers written on scratch paper or outside the designated answer spaces.
- If you need to use the restroom, bring your phone, student ID and exam to the front of the room.

For questions with **circular bubbles**, you should fill in exactly *one* choice.

◯ You must choose either this option

◯ Or this one, but not both!

For questions with **square checkboxes**, you may fill in *multiple* choices.

☐ You could select this choice.

☐ You could select this one too!

****Important****: Please completely **fill in** circles and squares to indicate answers and cross out or erase mistakes.

- **Valid** : ■ or ●
- **Invalid** : ⓥ, ☑, ☒ or ⊗

Preamble

You can complete these questions before the exam starts.

(a) What is your full name?

(b) What is your student ID number?

(c) Who is sitting to your left? (Write *no one* if no one is next to you.)

(d) Who is sitting to your right? (Write *no one* if no one is next to you.)

1. (15.0 points) Tire-lessly

Gabe got a flat tire. He wants to use data to inform his purchase of a new one. He has access to the `tires` table which contains the following data:

- *name*: (string) The name of the tire
- *width*: (int) The width of the tire in inches
- *efficient*: (string) Indicates whether a tire is efficient (`'Yes'`) or not (`'No'`)
- *price*: (int) Pre-tax price in dollars

The first three rows are shown below.

name	width	efficient	price
Michelin Winter	62	Yes	220
Continental All-Year	70	No	150
Goodyear Performance	63	No Data	180

`... (147 rows omitted)`

(a) (3.0 pt) When collecting the data, we could not find the efficiency data for some of the tires. These tires contain `'No Data'` in the *efficient* column. Write a line of code that outputs the total number of tires on which we do not have the efficiency data.

```
tires.where("efficient", "N/A").num_rows   or   tires.group("efficient",
np.average).column("price average").item(1)
```

(b) (3.0 pt) Gabe wants to buy a spare tire for his car, but he is on a budget. Write a line of code that outputs the name of the cheapest tire.

```
tires.sort("price").column("name").item(0)
```

(c) (4.0 points)

The tax rate on tires is 9%. Efficient tires are better for the environment, as they increase mileage. The government wants to incentivize people to buy these tires through a tax exemption - that is, efficient tires have no tax.

```
def tax_calc(price, efficiency):
    '''Input: price: (int) pre-tax price in dollars
              efficiency: (string) efficiency label
    Output: final_price: (float) final price in dollars
    '''
    if _____(a)______:
        return 1.09 * price
    else:
        return price
post_tax = tires.____(b)_____(____(c)____)
tires = tires.with_column("price", post_tax)
```

The code above adds the post-tax price to the `tires` table and assigns it to `tires`. Assume that all rows with `'No Data'` in the *efficient* column have been removed.

i. **(1.0 pt)** Fill in blank (a) so that the function works as described.

```
efficiency == 'No' OR efficiency != 'Yes'
```

ii. **(1.0 pt)** Fill in blank (b).

```
apply
```

iii. **(2.0 pt)** Fill in blank (c).

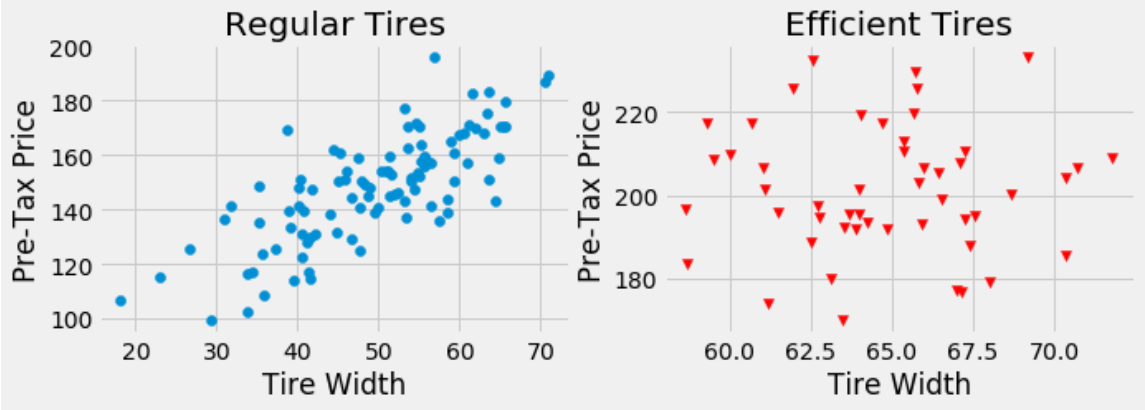
```
tax_calc, 'price', 'efficient'
```

(d) **(2.0 pt)** Gabe does not want too wide of a tire, so he is interested in visualizing the mean `width` for both the efficient and inefficient tires. Which of the following visualizations could you use to do this without any additional calculations? *Select all that apply*

- ☐ Histogram
- ■ Bar Chart
- ☐ Line Plot
- ☐ Scatter Plot
- ☐ None of the above

Histograms (even overlaid) is not a valid answer since the mean cannot be directly extracted in every scenario.

(e) **(3.0 pt)** He is also interested in analyzing the relationship between tire widths and the pre-tax prices for both types of tires. He creates the scatter plots shown below. What conclusion(s) can you draw from just this visualization? *Select all that apply*



- ☐ The production of wider tires generally causes the pre-tax prices to increase.
- ■ For regular tires, higher prices are associated with wider tires.
- ☐ For efficient tires, higher prices are associated with wider tires.
- ☐ The price range for regular tires is the same as the range for efficient tires.
- ☐ The price range for regular tires is smaller than the range for efficient tires.
- ■ The price range for regular tires is larger than the range for efficient tires.

2. (14.0 points) It's Called Football!

The table `players` contains the following data of 970 soccer players:

- *Name*: (string) The player's name
- *Position*: (string) The player's position
- *Club*: (string) The player's club team
- *Nation*: (string) The player's national team allegiance
- *Net Worth*: (int) The player's net worth in millions of US dollars

Each row corresponds to one player. Here are the first few rows of the table:

Name	Position	Club	Nation	Net Worth
Cristiano Ronaldo	ST	Manchester United	Portugal	490
Lionel Messi	RW	PSG	Argentina	420
Fabinho	CDM	Liverpool	Brazil	9
Karim Benzema	CF	Real Madrid	France	46

`... (966 rows omitted)`

(a) (3.0 pt) We are interested in how much players make on average. Write a line of code that outputs the average net worth in millions of US dollars of all players in the table.

```
np.mean(players.column("Net Worth"))
```

(b) (4.0 pt) Now we are given the `countries` table that has the following columns:

- *Country*: (string) Name of the country
- *Continent*: (string) Continent of country

Assume that all the countries the players play for are also contained in the `countries` table. Write one line of code that outputs a table that has the same columns as the table `players`, as well as a new `Continent` column. Note that these columns in the output do not have to be in the same order.

```
players.join("Nation", countries, "Country")
```

(c) (7.5 points)

For each club, we are interested in whether 'ST' players have a strictly higher average net worth than 'CF' players. The last line of the code below should evaluate to an array of booleans, where `True` indicates that the 'ST' position on a particular club has a higher average net worth than the 'CF' players on the same club.

```
nw_grid = players._____(a)____(_____________(b)_____________)
_____________________________(c)_______________________________
```

i. (2.5 pt) Which of the following should fill in blank (a)?

- ○ `apply`
- ○ `join`
- ○ `group`
- ● `pivot`

ii. **(3.0 pt)** Fill in blank (b).

```
'Position', 'Club', 'Net Worth', np.average
```

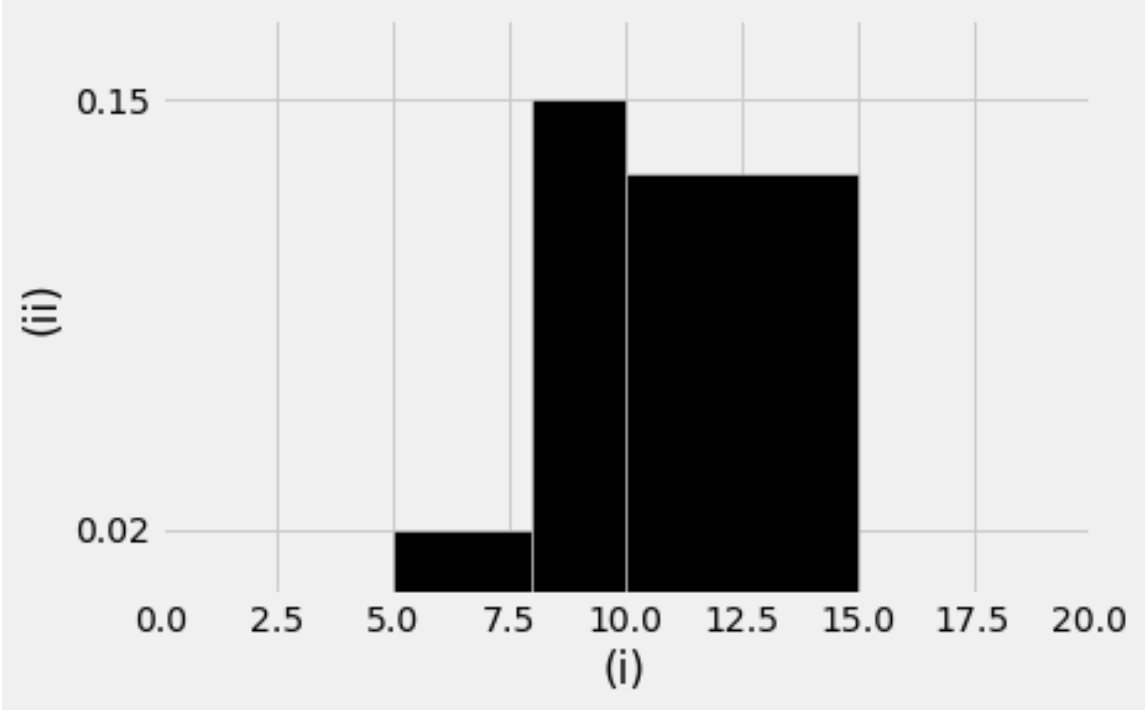
iii. **(2.0 pt)** Fill in blank (c).

```
nw_grid.column('ST') > nw_grid.column('CF')
```

3. (12.5 points) When I Was a Kid

Oscar's parents often told him the story of how when they were kids, they walked 12 miles to school through snow uphill in both ways. Oscar did not believe this story, and therefore asked some of his closest friends whether their parents had to as well. It turns out many did!

He collected information about the distances his friends' parents walked from home to school. Here is a visualization he created of some of the data he collected:



The table below includes the bins and the corresponding heights for the first two bins.

Bin	Height
[5,8)	0.02
[8,10)	0.15
[10,15)	???

(a) (2.5 pt) Oscar forgot to label the axes of his visualization. Which of the following are valid labels for (i) and (ii)? *Select All That Apply.*

- ■ (i) = *Distance*, (ii) = *Density*
- ■ (i) = *Distance*, (ii) = *Proportion per mile*
- ☐ (i) = *Distance*, (ii) = *Proportion per parent*
- ☐ (i) = *Proportion per mile*, (ii) = *Distance*
- ☐ (i) = *Proportion per parent*, (ii) = *Parent*
- ☐ None of the above

(b) **(2.0 pt)** For this and the next question, assume that 50% of parents said they walked between 10 to 15 miles to school uphill. What should the corresponding height be for this bar on the visualization?

- ● 0.1
- ○ 0.0
- ○ 10
- ○ 0.5
- ○ 50

(c) **(3.0 pt)** What percentage of parents took less than 15 miles to get to school, assuming none took fewer than 5?

```
(8 - 5) * 0.02 + (10 - 8) * 0.15 + (15 - 10) * 0.1 = 86%
```

(d) **(2.0 pt)** Is the visualization above a histogram?

- ● No, because areas do not sum to 1
- ○ No, because the bar widths are unequal
- ○ No, because distance is not a numerical variable
- ● Yes, because the visualization fulfills all the requirements of a histogram.

The intended answer was option A under the 50% assumption from part (b). Since this assumption wasn't clearly stated for this question, we also accepted option D for full credit.

(e) **(3.0 pt)** The final visualization only depicts the distribution of distances walked from home to school. Now Oscar asks parents for the total distance to and from school, instead of just the distance to school.

As a result, all distances were doubled. The table below show the new bins:

Bin
[10,16)
[16,20)
[20,30)

Which of the following statements correctly describe the resulting histogram? Assume the route taken in each direction is the same.

Select all that apply

- ☐ Bin heights would stay the same
- ☐ Bin heights would be doubled
- ■ Bin heights would be halved
- ■ Area of each bin would stay the same
- ☐ Area of each bin would be doubled
- ☐ Area of each bin would be halved

4. (18.0 points) Sampling and Simulations

(a) (3.0 pt) When the code below is run, `final` evaluates to a number that is approximately equal to one of the values below. Which one?

```
count = 0
for i in np.arange(10000):
    one_sim = np.random.choice(make_array(0, 0, 1),10)
    if sum(one_sim) > 0:
        count = count + 1
final = count / 10000
```

- ○ `1/3`
- ○ `2/3`
- ○ `10 * 1/3`
- ○ `10 * 2/3`
- ○ `(1/3)**10`
- ○ `(2/3)**10`
- ○ `1 - 10* 1/3`
- ○ `1 - 10* 2/3`
- ○ `1 - (1/3)**10`
- ● `1 - (2/3)**10`

(b) (3.0 pt) Suppose we have a table `staff` of all 45 Data 8 Staff members, which contains several columns, one of which is the staff member's `Name` (a string).

```
def mystery():
    s = staff.sample(10)
    if s.where('Name', 'Ellen Persson').num_rows > 0:
        return True
    else:
        return False
```

The `mystery` function shown above uses the `staff` table and returns either True or False. What is the approximate probability that it will return True?

You can assume that there is exactly one staff member with the name `Ellen Persson`.

`1 - (44/45) ** 10`

(c) (2.0 pt) Select the appropriate word to fill in each of the blanks.

One simulation entails flipping a fair coin 1000 times and calculating the proportion of times seeing more than 550 Heads. If we run this simulation many times, the proportion of times we get more than 550 Heads should be _____ the proportion of times we get more than 550 Tails.

- ○ Larger than
- ○ Smaller than
- ● Approximately equal to

(d) **(2.0 pt)** According to the Law of Large Numbers, if a simulation is repeated a large number of times, then the proportion of times that an event occurred in the simulation is very likely to be close to the _____ probability of the event.

- ○ Empirical
- ● Theoretical
- ○ Average
- ○ Observed

(e) **(2.0 pt)** If you don't know the theoretical distribution of a random variable but you have access to a large random sample, it is _____ to simulate the approximate probabilities of each value.

- ○ Possible
- ● Impossible

(f) **(3.0 pt)** Which of the following functions can be used to simulate a sample from a categorical distribution, given the appropriate tables/arrays? *Select All That Apply*

- ■ np.random.choice
- ■ tbl.sample(...)
- ■ sample_proportions
- ☐ None of the above

(g) **(3.0 pt)** The below code simulates the proportion of times that a fair die rolls a number greater than or equal to four, but there's an error in the code. What does `prop_higher_than_four` evaluate to?

```
rolls = make_array()
for i in np.arange(1000):
    roll = np.random.choice(np.arange(1, 7))
    if roll >= 4:
        np.append(rolls, roll)
prop_higher_than_four = len(rolls) / 1000
```

- ● 0
- ○ 0.333
- ○ 0.571 (Approximately 4/7)
- ○ 33
- ○ 50
- ○ None of the above

Since the result np.append is not reassigned to the `rolls` array, the final length evaluates to 0.

5. (12.0 points) Jolly Ranchers!

It's Meghan's birthday! She buys a bag of 100 Jolly Ranchers to celebrate. There are 4 flavors in total (orange, apple, cherry and peach). The bag Meghan got has the following flavor breakdown:

- 38 orange
- 22 apple
- 25 cherry
- 15 peach

She draws at random from the bag. After each draw, she returns the chosen Jolly Rancher to the bag. Please select the probability of the described events.

(a) (3.0 pt) The probability that Meghan picks a cherry first, and then an orange Jolly Rancher.

● $\frac{38}{100} * \frac{25}{100}$

○ $\frac{38}{100} + \frac{25}{100}$

○ $1 - \frac{38}{100} * \frac{25}{100}$

○ $\left(\frac{38}{100}\right)^2 + \left(\frac{25}{100}\right)^2$

○ None of the above

(b) (3.0 pt) The probability that Meghan picks one, two or three peach Jolly Ranchers in three draws.

○ $\left(\frac{15}{100}\right) + \left(\frac{15}{100}\right)^2 + \left(\frac{15}{100}\right)^3$

○ $3 * \left(\frac{15}{100}\right) + 2 * \left(\frac{15}{100}\right)^2 + \left(\frac{15}{100}\right)^3$

● $1 - \left(\frac{85}{100}\right)^3$

○ $3 * \left(\frac{15}{100}\right) * \left(\frac{85}{100}\right)^2 + 2 * \left(\frac{85}{100}\right) * \left(\frac{15}{100}\right)^2 + \left(\frac{15}{100}\right)^3$

○ None of the above

(c) (3.0 pt) The probability that Meghan ends up with exactly two orange Jolly Rancher in three draws.

○ $$1 - \left(\frac{38}{100}\right)^3$$

● $$3 * \left(\frac{38}{100}\right)^2 * \left(\frac{62}{100}\right)$$

○ $$1 - 3 * \left(\frac{62}{100}\right)^2 * \left(\frac{38}{100}\right)$$

○ $$1 - (\frac{38}{100})^2$$

○ None of the above

(d) (3.0 pt) Meghan gets hungry. Each time she draws now, she eats the Jolly Rancher. What is the probability of her drawing exactly one apple Jolly Rancher in three draws?

○ $$\left(\frac{22}{100}\right) * \left(\frac{78}{100}\right)^2$$

○ $$3 * \left(\frac{22}{100}\right) * \left(\frac{78}{100}\right)^2$$

● $$3 * \left(\frac{22}{100}\right) * \left(\frac{78}{99}\right) * \left(\frac{77}{98}\right)$$

○ $$1 - 3 * \left(\frac{22}{100}\right) * \left(\frac{78}{99}\right) * \left(\frac{77}{98}\right)$$

○ None of the above

6. (18.0 points) I Will Walk Instead

Walking uphill to campus can be quite tedious. Fortunately, there is a ZC-transit bus stop in front of your house. ZC-transit operates 3 buses: new, regular, and old. According to the ZC transit app, a bus ride from your place to campus always takes exactly 20 minutes on a regular bus, 15 minutes on a new bus, and 30 minutes on an old bus. Max, the ZC-transit customer service representative, told you that every morning for your route, a bus is uniformly sampled from the three buses they have.

You, a true data scientist, notice that in the last 10 days that it doesn't seem uniform at all. There were many days that you arrived in 15 minutes or 30 minutes, but few days that your commute took 20 minutes.

(a) (4.0 pt) Given the information above, state a clear and complete null hypothesis.

- **The 10 buses in the sample are drawn uniformly at random with replacement from the possible bus options.**
- **Each of the buses in the sample has an equal chance of being each type of bus, regardless of the other buses**
- **The buses in the sample were drawn uniformly at random with replacement from the distribution of buses given by Max.**

These are examples of valid null hypotheses.

(b) (6.0 pt) *For the following questions, consider the scenario and hypotheses provided below.*

ZC-transit is livid that you are investigating them. Of the last 10 days, there were 7 days where your trip took 30 minutes. You now suspect that ZC-transit is consistently scheduling old buses on your route to annoy you.

Null Hypothesis: Old buses are equally likely to show up as any of the other buses, that is - with a 1/3 probability. Any deviation is due to chance.

Alternative Hypothesis Old buses are more likely to show up than the other buses.

Given the information above, which of the following test statistics are valid to test your hypothesis? *Select all that apply*

- ■ Total commute time
- ■ Number of trips with old buses
- ☐ Proportion of trips with regular buses
- ☐ Difference of number of trips with old buses and trips with new buses
- ☐ Total Variation Distance between the observed distribution and theoretical distribution of buses
- ☐ (1/3) * Total Variation Distance between the observed distribution and theoretical distribution of buses
- ☐ None of the above

(c) **(3.0 pt)** We decide to use the difference between the proportion of trips with old buses and the corresponding expected proportion.

You decide to recycle a function from lab that simulates one test statistic. Unfortunately, your notebook crashed and left these blanks in one line. Copy the line and fill in the blanks.

```
def simulate_one_stat():
    expected_props = make_array(1/3, 1/3, 1/3)
    prop_old = _____________(___________, ______).item(0)
    return prop_old - expected_props.item(0)
```

```
prop_old = sample_proportions(10, expected_props).item(0)
```

(d) **(2.5 pt)** You obtained a p-value of 0.04. Assume a p-value cutoff of 0.05. Which of the following statements are correct? *Select all that apply*

- ☐ You fail to reject the null hypothesis.
- ☐ The alternative hypothesis is true.
- ☐ There is a 4% chance that your null hypothesis is true.
- ☐ There is a 96% chance that your null hypothesis is true.
- ☐ If the null hypothesis is true, then there is a 4% chance we will reject the null.
- ■ None of the above

(e) **(2.5 pt)** Your friend brings in a new test statistic: the absolute difference of the proportion of old buses and expected proportion. Which of the following statements are correct? *Select all that apply*

- ☐ The new test statistic is valid for testing your null and alternative.
- ☐ The new test statistic is invalid for your null and alternative because it does not consider the new and regular buses.
- ■ The new test statistic is invalid for your null and alternative because it is undirectional.
- ☐ The histograms of both our old and new simulated test statistics would look identical.
- ■ The shape of the histograms of both our old and new simulated test statistics would look different.