DATA C8
Spring 2026

# FOUNDATIONS OF DATA SCIENCE

FINAL
Solutions

PRINT Your Name: _____

PRINT Your Student ID: _____

PRINT Your Exam Room: _____

PRINT the Name of Person to your Left: _____

PRINT the Name of Person to your Right: _____

PRINT Your TA's Name (Write N/A if in Self-Service): _____

## INSTRUCTIONS

You have **2 hours and 50 minutes** to complete the exam. There are **8 questions** and **18 pages** on this exam, including this cover page.

| Question | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
|---|---|---|---|---|---|---|---|---|---|
| Points | 18 | 25 | 14 | 21 | 32 | 6 | 4 | 0 | 120 |

- This exam is closed book, closed computer and closed calculator, except the Reference Sheet provided for you.
- You may only have with you: a pencil, an eraser, and your student ID (unless you have pre-approved accommodations).
- If you need to use the restroom, bring your phone, exam, reference sheet and student ID to the front of the room.
- For written questions:
  - answers written outside the boxes provided will not be graded;
  - if your answer is ambiguous or you provide multiple answers, the worst interpretation will be graded.
- For coding questions:
  - blank spaces may include multiple arguments or functions per blank, but your solution must use every blank available;
  - you may assume the `datascience` and `numpy` libraries are imported, as seen in class;
- For multiple choice questions, see question types and instructions below.

Questions with **circular bubbles**: you may select only **1 choice**.

○ Unselected option (completely unfilled)

● Single option selected (completely filled)

Questions with **square boxes**: you may select **1 or more choices**.

■ You may select multiple squares

■ as long as they are completely filled

You must fill in the bubbles **completely**. Ticks, crosses, or other check marks will **not** receive credit.

## HONOR CODE

*"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."*

SIGN Your Name: _____

This page intentionally left blank

The exam begins on the next page.

## 1. [18.0 points] Blind Box Bonanza

Dagny and Tiffany are opening animal blind boxes! In a set of 10 blind boxes, they are guaranteed 3 pandas, 2 elephants, 2 owls, 2 raccoons, and 1 otter. Once they open a blind box, they cannot put it back. You can leave answers as unsimplified equations.

**(a)** [2.0 pts] What is the chance that they first get either a panda or a raccoon?

```
3/10 + 2/10
```

**(b)** [2.0 pts] What is the chance that they first get a panda and second get a raccoon?

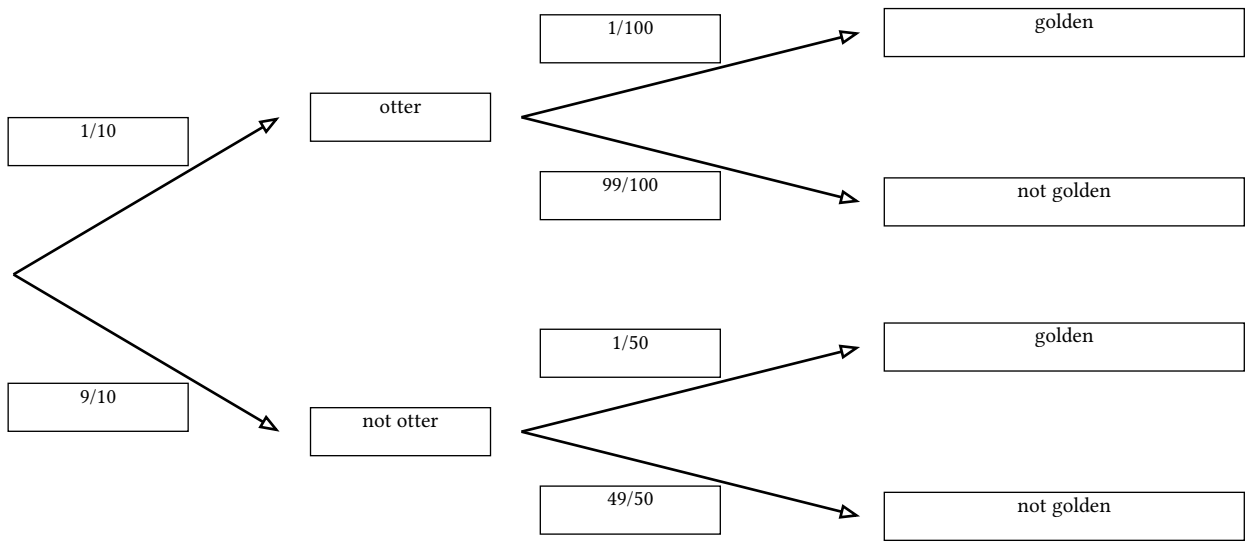
```
3/10 * 2/9
```

**(c)** [2.0 pts] If they open 3 out of the 10 blind boxes, what is the probability that at least one of them is an owl?

- ◯ `(2/10)**3`
- ◯ `1 - (2/10)**3`
- ◯ `(8/10)**3`
- ◯ `1 - (8/10)**3`
- ◯ `8/10 * 7/9 * 6/8`
- ● `1 - (8/10 * 7/9 * 6/8)`

They learn that there is a golden version of each animal! There is a 1/100 chance that each otter is golden and a 1/50 chance that each other animal is golden.

**(d)** [3.0 pts] Fill in the tree diagram that best describes this situation. You should not have to perform any calculations to do so.



**(e)** In the context of this tree diagram:

**i.** [1.0 pt] What is the **prior**?

- (●) The probability that the animal is an otter before knowing whether it is golden or not
- ( ) The probability that the animal is golden before knowing whether it is an otter or not
- ( ) The probability that the animal is an otter given whether it is golden or not
- ( ) The probability that the animal is golden given whether it is an otter or not

**ii.** [1.0 pt] What is the **likelihood**?

- ( ) The probability that the animal is an otter before knowing whether it is golden or not
- ( ) The probability that the animal is golden before knowing whether it is an otter or not
- ( ) The probability that the animal is an otter given whether it is golden or not
- (●) The probability that the animal is golden given whether it is an otter or not

**iii.** [1.0 pt] What is the **posterior**?

- ( ) The probability that the animal is an otter before knowing whether it is golden or not
- ( ) The probability that the animal is golden before knowing whether it is an otter or not
- (●) The probability that the animal is an otter given whether it is golden or not
- ( ) The probability that the animal is golden given whether it is an otter or not

**(f)** [4.0 pts] Dagny and Tiffany took a peek into the box and saw that the figure was golden (but not what animal it was). Given the figure is golden, what is the probability that it is an otter? **You must show your work and write a single fraction as the answer (not an equation).**

$$P(\text{otter} \mid \text{golden}) = \frac{P(\text{golden} \mid \text{otter}) \cdot P(\text{otter})}{P(\text{golden})} = \frac{\frac{1}{100} \cdot \frac{1}{10}}{\frac{1}{100} \cdot \frac{1}{10} + \frac{1}{50} \cdot \frac{9}{10}} = \frac{\frac{1}{1000}}{\frac{1}{1000} + \frac{9}{500}} = \frac{\frac{1}{1000}}{\frac{19}{1000}} = \frac{1}{19}$$

**(g)** [2.0 pts] If the manufacturer changed the chance that an otter is golden to be 1/50, just like the other animals, how would the probability that a golden figure is an otter change?

(●) Increase ( ) Decrease ( ) Stay the same ( ) Impossible to tell

## 2. [25.0 points] Top Dog

Marissa samples 200 Berkeley students at random and asks them which they prefer, Top Dog or Artichoke Pizza. 115 prefer Top Dog and 85 prefer Artichoke Pizza. Help her use 1,000 bootstrap resamples to create an approximate confidence interval for the proportion of Berkeley students who prefer Top Dog over Artichoke Pizza.

**(a)** [3.0 pts] Fill in the blanks so that the function `one_bootstrapped_prop` computes a single bootstrapped proportion of Berkeley students who prefer Top Dog.

```
def one_bootstrapped_prop():
    "Returns the proportion of students who prefer Top Dog in a random sample."

    props = make_array(_____, _____)

    return sample_proportions(_____, props).item(0)
```

```
115/200, 85/200, 200
```

**(b)** [4.0 pts] Fill in the blanks so that `ci` returns an approximate `level`-% confidence interval (as a two-element array of the lower and upper bound of the interval) for the proportion of Berkeley students who prefer Top Dog over Artichoke Pizza. You **must** call `one_bootstrapped_prop` on one of the blanks. The input `level` is a number.

```
def ci(level):
    props = make_array()
    for i in np.arange(1000):

        props = _____

    p = _____
    low = percentile(p, props)
    high = percentile(100 - p, props)
    return make_array(low, high)
```

```
np.append(props, one_bootstrapped_prop())
(100 - level) / 2
```

**(c)** [2.0 pts] What would you expect to be the relationship between `ci(95)` and `ci(99)`?

- ◯ The `low` and `high` values would be about the same.
- ◯ The `low` and `high` values would both be lower for `ci(95)`
- ◯ The `low` and `high` values would both be lower for `ci(99)`
- ◯ The `low` value would be lower and `high` value would be higher for `ci(95)`
- ● The `low` value would be lower and `high` value would be higher for `ci(99)`

**(d)** [2.0 pts] How would you expect the width of `ci(95)` to change if the number of bootstrapped samples were increased from 1000 to 4000?

- ● The interval returned would have about the same width.
- ◯ The interval returned would be about one quarter as wide.
- ◯ The interval returned would be about half as wide.

- ◯ The interval returned would be about twice as wide.
- ◯ The interval returned would be about four times as wide.

**(e)** [3.0 pts] If `ci(95)` returns `array([0.525, 0.62])`, what can we say about the proportion of Berkeley students who prefer Top Dog over Artichoke Pizza? Select **all** the ways of completing, "The data are consistent with the hypothesis that the proportion of Berkeley students who prefer Top Dog over Artichoke Pizza is ..."

- ☐ "Equal to 1/2 when using a p-value cut-off of 1%"
- ☐ "Equal to 1/2 when using a p-value cut-off of 5%"
- ☐ "Equal to 1/2 when using a p-value cut-off of 10%"
- ☐ "Not equal to 1/2 when using a p-value cut-off of 1%"
- ■ "Not equal to 1/2 when using a p-value cut-off of 5%"
- ■ "Not equal to 1/2 when using a p-value cut-off of 10%"
- ☐ None of these

**(f)** [3.0 pts] Select all statements that are true about calling `ci(95)` multiple times.

- ☐ Every call to `ci(95)` will return the same interval.
- ☐ About 95% of calls to `ci(95)` will contain the population proportion of Berkeley students who prefer Top Dog.
- ☐ About 95% of calls to `ci(95)` will contain the sample proportion of Berkeley students who prefer Top Dog.
- ■ None of these

**(g)** [2.0 pts] Complete this sentence: For about 95% of _____, a 95% confidence interval created using the bootstrap will contain the population parameter.

- ◯ individuals in the population
- ◯ individuals in the original sample
- ◯ random bootstrap samples from the original sample
- ● random samples from the population

**(h)** [4.0 pts] Marissa thinks she needs a larger sample. She plans to construct a 99.7% confidence interval for the proportion of Berkeley students who prefer Top Dog, and she would like to ensure that the width of this interval is no wider than about 0.1. What is the minimum sample size needed to satisfy this requirement? **Show your work and write a number.**

Width of a 99.7% CI $= 6 \cdot \frac{\text{SD}}{\sqrt{n}}$. Using the max SD of 0.5:

$0.1 = 6 \cdot \frac{0.5}{\sqrt{n}}$, so $\sqrt{n} = 30$ and $n = 900$.

**(i)** [2.0 pts] Marissa asks every Berkeley student, they all respond, and 52% of them prefer Artichoke Pizza! What should she do next to make a reliable inference about the proportion of Berkeley students who prefer Top Dog?

- ◯ Sample from these responses many times to create a confidence interval.
- ◯ Permute these responses many times to conduct a hypothesis test about whether the true proportion is 50%.

- ◯ Compute the maximum standard deviation of sample proportions based on the total number of Berkeley students to find the width of a confidence interval.
- ● Subtract 52% from 100%.

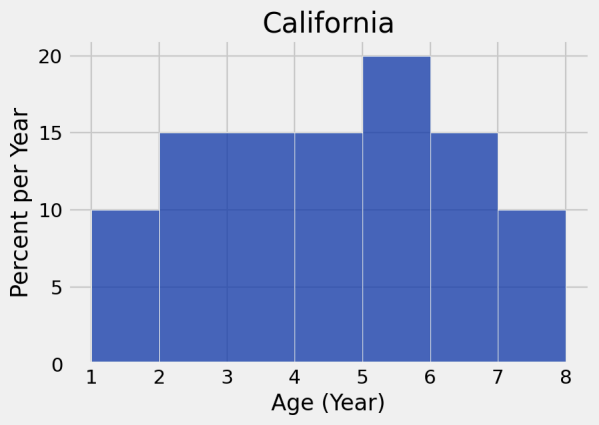
## 3. [14.0 points] Banana Slugs

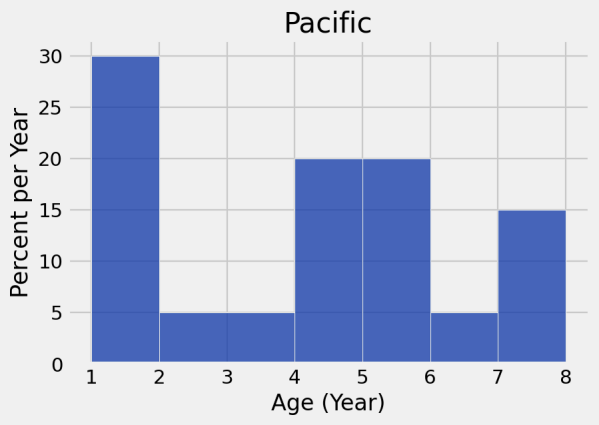
Mariel recorded observations of 100 banana slugs at UC Santa Cruz: 80 California slugs and 20 Pacific slugs. Each slug is recorded in a row of `slugs_table` with columns `Age` (int), `Length` (float), `Width` (float), `Building` (string), and `Species` (string).

| Age | Length | Width | Building | Species |
|---|---|---|---|---|
| 5 | 21.9 | 4.0 | Stevenson College | California |
| 4 | 22.1 | 3.8 | Oakes College | California |
| 6 | 13.1 | 2.5 | Porter College | California |

... (97 rows omitted)

A histogram of `Age` for the California slugs appears to the left, and a histogram of `Age` for the Pacific slugs appears to the right.




**(a)** [2.0 pts] What **percentage** of **Pacific** slugs are either 2 or 3 years old?

○ 5% ● 10% ○ 15% ○ 20% ○ 25% ○ 30% ○ Cannot tell

**(b)** [3.0 pts] How many slugs in total (both species) are less than 2 years old? **Write a number and show your work.**

10% x 80 + 30% x 20 = 8 + 6 = 14

**(c)** [3.0 pts] If we drew a histogram of **all 100 slugs**, what would be the height of a bar from age 2 to 4? **Write a number.**

(30% x 80 + 5% x 20) / 2 = 26 / 2 = 13

**(d)** [4.0 pts] Write a Python expression that evaluates to a table with one row per unique age and one column per unique building. For a given age (row) and building (column), the table value is the average length of all slugs of that age and found near that building. Assume that there is at least one slug of each age found near each building.

```
slugs_table._____
        pivot("Building", "Age", "Length", np.mean)
```

**(e)** [2.0 pts] The length distribution of slugs has mean 15 and standard deviation 5. Width and length are linearly associated, and the correlation coefficient $r$ is 0.8. What is the average length of slugs that have a width that is one SD below the mean width?

○ 5 ○ 7 ○ 10 ● 11 ○ 14 ○ 15 ○ 19 ○ 20 ○ 23 ○ 25

## 4. [21.0 points] Climb of Best Fit

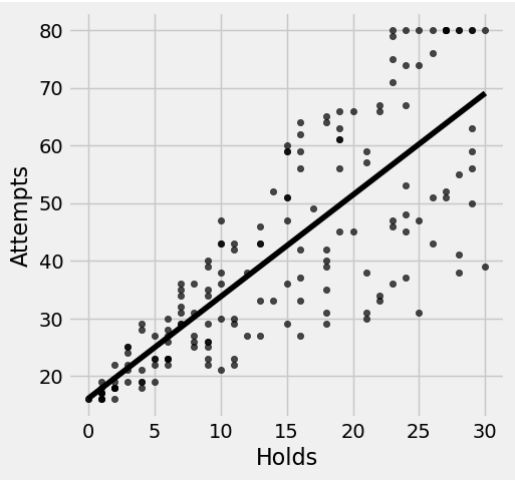
A rock climbing route has a number of `Holds` (`int`) and a `Type` (`str`). The `Type` is always either `"Overhang"` or `"Slab"`. Nao records how many `Attempts` (`int`) she required to complete 55 different climbing routes in the table `climbs`.

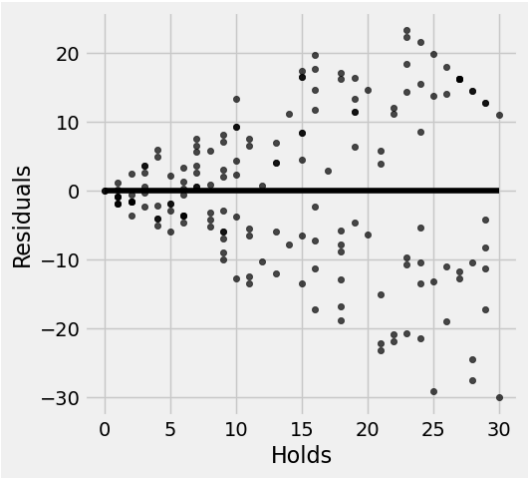
**(a)** [3.0 pts] Fill in the code to calculate the `r`, `slope`, and `intercept` of the regression line predicting `Attempts` from `Holds`. You can use `su`, which is an abbreviation for *standard units*, as well as `np.mean`, `np.std`, `h`, and `a`, but you **may not** call other functions or refer to `climbs`.

```
def su(arr):
    """Return the array arr converted to standard units."""
    return (arr - np.mean(arr)) / np.std(arr)

h = climbs.column("Holds")

a = climbs.column("Attempts")

r = _____

slope = r * _____

intercept = _____
```

```
np.mean(su(h) * su(a)); np.std(a) / np.std(h); np.mean(a) - slope * np.mean(h)
```

**(b)** [4.0 pts] Based on the scatter plot (left) and residual plot (right) below, which of the following are reasonable conclusions to make? Select all that apply.




- ☐ There is no association between the number of holds on a route and the number of attempts.
- ■ There is a positive association between the number of holds on a route and the number of attempts.
- ■ There is a positive linear association between the number of holds on a route and the number of attempts.
- ☐ This residual plot is clearly impossible under linear regression, so there must be a bug in the code.
- ☐ None of these.

**(c)** [2.0 pts] According to the plots above, is the residual positive or negative for a route that has 15 holds and 60 attempts?

- ● Positive
- ○ Negative

◯ Not enough information

**(d)** [4.0 pts] Which of the following must be true of the residual plot? Select all that apply.

■ The sum of residuals is zero.

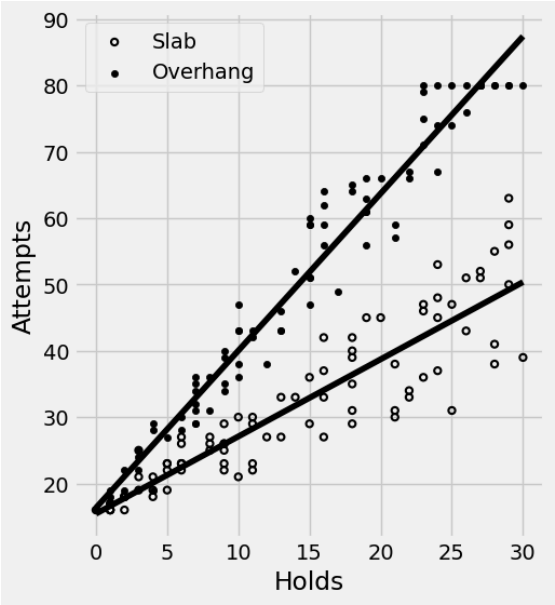
☐ The sum of squared residuals is zero.

☐ The sum of residuals is minimized.

■ The sum of squared residuals is minimized.

☐ None of these

Nao groups the data by type and fits a separate linear regression line for Overhang and Slab routes.




**(e)** [2.0 pts] Based on this plot, will considering the route type lead to more accurate predictions on average?

● Yes, because residuals are smaller when considering the two types separately.

◯ Yes, because using less data for a regression line tends to reduce prediction error.

◯ No, because using less data for a regression line tends to increase prediction error.

◯ No, because the original dataset already showed a clear linear trend.

**(f)** [4.0 pts] To predict `Holds` from `Attempts` using linear regression (instead of predicting `Attempts` from `Holds`), which of the following is true? Select all that apply.

■ The same correlation coefficient will be used.

☐ The same regression line slope will be used.

☐ The same regression line intercept will be used.

☐ The same fitted values will result.

☐ None of these

**(g)** [2.0 pts] What method from Data 8 could be used to predict the route type based on the number of holds and the number of attempts?

● k-NN Classification

◯ Linear Regression

- ◯ Regression Inference
- ◯ Total Variation Distance
- ◯ Nearest Neighbor Regression
- ◯ Least Squares

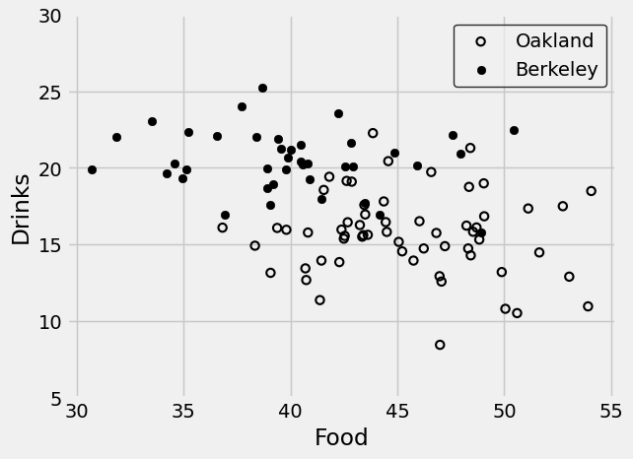
## 5. [32.0 points] A Table of Two Cities

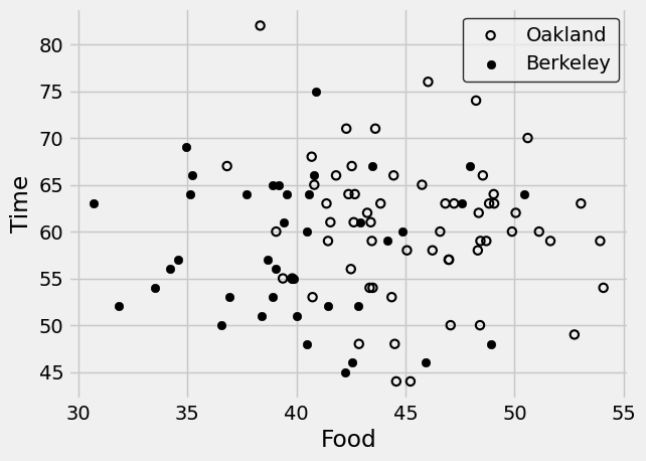
The owner of a restaurant on the border of Berkeley and Oakland creates a `patrons` table with one row per person served, how much they spent on `Food` (`float`) and `Drinks` (`float`), how much `Time` (`int`) in minutes they spent, and which `City` (`str`) they came from. Assume all patrons are either from Berkeley or Oakland, and this table contains a random sample of the patrons that come to the restaurant.

| Food | Drinks | Time | City |
|---|---|---|---|
| 46.99 | 14.56 | 66 | Oakland |
| 44.45 | 15.44 | 53 | Oakland |
| 47.59 | 12.68 | 71 | Oakland |

... (97 rows omitted)

A scatter plot of `Food` vs `Drinks` appears to the left, and a scatter plot of `Food` vs `Time` appears to the right. Filled dots represent Berkeley patrons, and hollow dots represent Oakland patrons.




**(a)** For each pair of variables, indicate what the scatter plots above show about their association.

[1.0 pt] Time and Food:

○ Positive ○ Negative ● Very weak or none ○ Not enough info

[1.0 pt] Food and Drinks:

○ Positive ● Negative ○ Very weak or none ○ Not enough info

[1.0 pt] Time and City:

○ Clear association ● Weak/no association ○ Not enough info

[1.0 pt] Food and City:

● Clear association ○ Weak/no association ○ Not enough info

[1.0 pt] Drinks and City:

● Clear association ○ Weak/no association ○ Not enough info

**(b)** [2.0 pts] The average food cost of a rival restaurant is \$30 and the standard deviation is \$10. From this information, what is the largest possible fraction of food orders per patron that are either above \$50 or below \$10 in that restaurant?

○ 0 ○ 1/2 ○ 1/3 ● 1/4 ○ 1/8 ○ 1/9 ○ 1/16

**(c)** [3.0 pts] The `length` function is defined below. Fill in the blank so that `patrons` has a new `Length` column containing the category returned by the `length` function when called on each patron's `Time` value.

```
def length(minutes):
    if minutes <= 60:
        return 'Quick'
    elif minutes <= 75:
        return 'Normal'
    else:
        return 'Long'
```

| Food | Drinks | Time | City | Length |
|---|---|---|---|---|
| 46.99 | 14.56 | 66 | Oakland | Normal |
| 44.45 | 15.44 | 53 | Oakland | Quick |
| 47.59 | 12.68 | 71 | Oakland | Normal |

... (97 rows omitted)

```
patrons = patrons.with_column('Length', _____________________________________________________)
patrons.apply(length, 'Time')
```

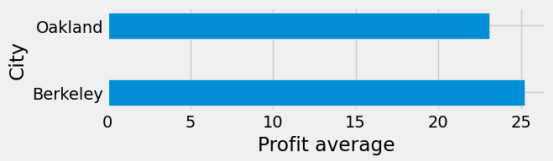
**(d)** [2.0 pts] The code below creates a table with a column called `X mean`. How many rows does it have, and what does it contain?

```
t = patrons.with_column('X', patrons.column('City') == 'Oakland')
t.select('Length', 'X').group('Length', np.mean)
```

- ◯ 2 rows containing, for Oakland, the fraction of patrons that stayed for each length
- ◯ 2 rows containing, for each length, the fraction of patrons that were from Oakland
- ◯ 3 rows containing, for Oakland, the fraction of patrons that stayed for each length
- ● 3 rows containing, for each length, the fraction of patrons that were from Oakland

Next, create this chart of the restaurant's average profit for patrons from each city. The profit is 50% of the price of food and drinks, minus the cost to serve the table, which varies depending on the length of their stay according to the `table_cost` table.

```
table_cost = Table().with_columns(
    'Stay', ['Quick', 'Normal', 'Long'],
    'Cost', [      5,        8,     10])
```




```
profit = 0.5 * (_______) - _______
                  (e)        (f)
patrons.select('City').with_column('Profit', profit)._______.barh('City')
                                                      (g)
```

**(e)** [2.0 pts] Fill in blank (e).

- ◯ `patrons.select('Food', 'Drinks')`
- ◯ `patrons.select('Food' + 'Drinks')`
- ◯ `patrons.select('Food') + patrons.select('Drinks')`
- ◯ `patrons.column('Food', 'Drinks')`
- ● `patrons.column('Food') + patrons.column('Drinks')`

- ◯ `patrons.column('Food' + 'Drinks')`

**(f)** [4.0 pts] Fill in blank (f). You **may not** use `apply`.

```
patrons.join('Length', table_cost, 'Stay').column('Cost')
```

**(g)** [2.0 pts] Fill in blank (g).

- ● `group('City', np.average)`
- ◯ `group('Profit', np.average)`
- ◯ `group('City', 'Profit', np.average)`

**(h)** [2.0 pts] The owner notices that average profit per patron is higher for Berkeley residents. If the owner is interested in whether this difference is due to chance, which null hypothesis should she test?

- ● The patron profits for each city were drawn from the same underlying distribution.
- ◯ The averages are drawn from a 50/50 distribution; any difference is due to chance.
- ◯ The costs are drawn from the fixed distribution described in `table_cost`.
- ◯ The average profit for patrons from each city in the population is the same as the average profit in the sample.

**(i)** [4.0 pts] If this null hypothesis is rejected, which of the following can the owner conclude? Select all that apply. At least one of them is correct.

- ☐ Restaurants in Berkeley have higher profit than restaurants in Oakland.
- ☐ Increasing the fraction of patrons from Berkeley by advertising there should increase profit per patron for her restaurant.
- ■ Among all patrons to her restaurant, those from Berkeley have a higher average profit than those from Oakland.
- ☐ A larger sample size is needed to determine whether this difference in average profits is due to chance.

**(j)** [2.0 pts] The owner wishes to predict whether future patrons are from Berkeley or Oakland using a k-nearest-neighbor classifier. She splits the data randomly into a training and test set. Complete this sentence describing how she should learn about what mistakes the classifier makes:

- ◯ Find `k` neighbors in the training set for each training example and look at the misclassified examples.
- ● Find `k` neighbors in the training set for each test example and look at the misclassified examples.
- ◯ Find `k` neighbors in the test set for each test example and look at the misclassified examples.

**(k)** [2.0 pts] The owner finds that using all three features (`Food`, `Drinks`, and `Time`) provides better test set accuracy than just using `Food` and `Drinks`. What can she conclude? Assume future patrons come from the same distribution as current patrons.

- ◯ Using all three features will have higher accuracy than just using `Food` and `Drinks` when applied to future patrons.
- ◯ Since the test set contains no future patrons, she cannot conclude anything about future patrons.
- ● The observed difference in accuracy might be due to chance, and so she should conduct a hypothesis test to decide if three features will have higher accuracy than just using `Food` and `Drinks` when applied to future patrons.

**(l)** [2.0 pts] Patrons leave notes in the restaurant guest book. How best can this text be processed by a neural network to provide attributes to the owner's k-NN classifier?

- ◯ The bag of words in the note can be converted to embeddings; each word becomes a numerical attribute.

- ◯ The whole note can be converted to a list of embeddings; an embedding is a numerical attribute.
- ● The whole note can be converted to one embedding; an embedding is a list of numerical attributes.

## 6. [6.0 points] Climate Project

In Project 2, we used average high temperature measurements in Phoenix, Arizona from all months of February, 1900–2021, to test a null hypothesis that February average temperatures in the present period (2019–2021) are drawn from the same distribution as February temperatures in the past (1900–1960).

**(a)** [3.0 pts] What assumptions are needed to justify using this null hypothesis? Select all that apply.

- [x] Temperature variations from day to day are like random samples from some distribution, rather than being determined by complicated natural phenomena.
- [ ] Temperatures are drawn from a bell-shaped normal distribution.
- [ ] The fraction of values that are more than $x$ standard deviations from the mean is no more than $\frac{1}{x^2}$.
- [ ] None of these.

**(b)** Complete this sentence by choosing the correct way to fill in each blank in this sentence about multiple hypothesis tests: When conducting a separate hypothesis test for each month using a confidence level of 90%, we should expect that we will __(i)__ for about __(ii)__ of months even if the __(iii)__.

**i.** [1.0 pt] Fill in blank (i).

- (●) reject the null
- ( ) not reject the null
- ( ) accept the null
- ( ) sample under the null

**ii.** [1.0 pt] Fill in blank (ii).

- (●) 10%
- ( ) 3 of 12
- ( ) 90%
- ( ) 9 of 12

**iii.** [1.0 pt] Fill in blank (iii).

- (●) null is true
- ( ) null is false
- ( ) sample size is too small
- ( ) sample size is adequate

## 7. [4.0 points] Movies Project

Implement this function from the project. Your answer must be one line following `return`, but if you run out of space, you can continue below the line.

[4.0 pts] Fill in the blank so that `most_common` returns the most common value in the `label` column of `table`. In case of a tie, it may return any one of the most common values.

```
def most_common(label, table):
    """This function takes two arguments:
      label: The label of a column, a string.
      table: A table.

    It returns the most common value in the label column of the table.
    """


    return _____
```

```
table.group(label).sort('count', descending=True).column(label).item(0)
```

## 8. Just for Fun (and Completely Optional)

**(a)** Draw a picture of your Data 8 journey.