# Spring 2023 Midterm Exam

Foundations of Data Science

Name: **Sample Solutions**

Total Score: _____ of 100 Points

## Instructions

- Write your name in the space provided on this page.
- Remove the last page of this exam that contains the list of Tables to reference while you complete the exam.
- Make sure you have a copy of the provided Midterm Reference Guide.
- Try to provide your responses in the spaces provided. If you find that you need additional space, write your extended response(s) on one of the provided blank sheets of paper and number them, so we can connect your response to the question.
- For Sections A and C, select the correct response(s) or provide a written response depending on the question type.
- For Section B, provide Python code that would provide the requested response if run in one of our notebooks. You should either provide your response in the box provided or fill in the template provided.
- You can assume the following code has been run, when you are writing your responses for Section B:

```
from datascience import *
import numpy as np
import matplotlib
%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
```

- Ask for clarification if you need it.
- Once you are finished, turn in your exam and you are welcome to leave. Have a nice Spring Break!

# Section A

1. (2 points) You have learned that an association between two factors does not necessarily mean that one causes the other ("correlation does not imply causation"). Which of the following would be the best example to demonstrate that? Choose one.

   ◯ Data shows an association between smoking and lung cancer. This is because smoking causes lung cancer.

   ◯ John Snow discovered that S&V customers died from cholera at a higher rate than Lambeth customers, but this was due to Lambeth's water being cleaner than S&V's; it had nothing to do with dirty water causing cholera.

   ★ **An observational study found an association between poor reading comprehension and having experienced a heart attack in the past 10 years, but this is not because strong reading skills prevent heart attacks; it is because poverty tends to cause weaker reading comprehension (because schools in poor areas tend to be weaker) and poverty tends to increase the risk of heart attacks (because poor people have worse access to medical care).**

   ◯ A study found an association between poor reading comprehension and having experienced a stroke in the past 10 years, but this is not because strong reading skills prevent strokes; it is because strokes often cause a cognitive decline and worse reading comprehension.

2. (4 points) Which of the following must be true, for an experiment to count as a randomized controlled experiment? Select all that apply.

   ★ **There is a control group.**

   ☐ The experimenters control who is selected to participate in the experiment.

   ★ **Randomness is used to determine whether each participant will be part of the control group(s) or treatment group(s).**

   ☐ Each participant is informed whether they are in the treatment group or not.

   ☐ The distribution of ages of the participants in the experiment are representative of the distribution of ages in the population at large.

3. (3 points) Suppose you perform a randomized controlled experiment where people in the treatment group take vitamin C regularly and people in the control group do not take vitamin C. In each group you measure the proportion that get sick over the next year. Which of the following statements are correct? Select all that apply.

   ◯ If the control group has a higher proportion of sickness, then it is reasonable to conclude that vitamin C causes a reduction in the chance of getting sick.

   ★ **If the control group has a higher proportion of sickness and a hypothesis test finds that this difference is statistically significant, then it is reasonable to conclude that there is an association between taking vitamin C and not getting sick.**

   ★ **If the control group has a higher proportion of sickness and a hypothesis test finds that this difference is statistically significant, then it is reasonable to conclude that vitamin C causes a reduction in the chance of getting sick.**

4. A real estate company has a dataset of all their buildings, with three attributes for each building: its size (in square feet), its type (residential or commercial), and its estimated value (sale price) if sold (in dollars).

   (a) (3 points) Select all that are correct:

   ★ **The size attribute is a numerical variable.**

   ☐ The type attribute is a numerical variable.

   ★ **The value attribute is a numerical variable.**

   (b) (2 points) The best visualization to understand the distribution of building types is: (choose one)

   ★ **A bar chart**

   ◯ A line plot

   ◯ A scatter plot

   ◯ A histogram

   (c) (2 points) The best visualization to understand the distribution of building sizes is: (choose one)

   ◯ A bar chart

   ◯ A line plot

   ◯ A scatter plot

   ★ **histogram**

   (d) (2 points) The best visualization to check for an association between building size and building value is: (choose one)

   ◯ A bar chart

   ◯ A line plot

   ★ **A scatter plot**

   ◯ A histogram

   ◯ Two histograms, overlaid

   (e) (2 points) The best visualization to check for an association between building size and building type is: (choose one)

   ◯ A bar chart

   ◯ A line plot

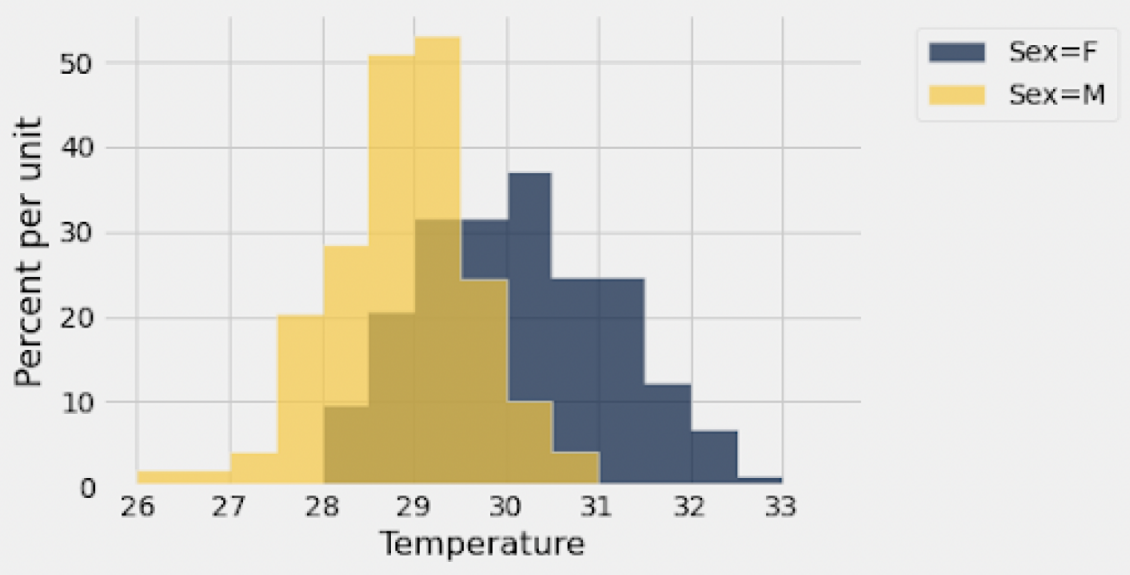
   ◯ A scatter plot

   ◯ A histogram

   ★ **Two histograms, overlaid**

5. When hatching a baby turtle from an egg, we incubate the egg at some temperature. A researcher read that the temperature an egg is incubated at influences whether or not the turtle that hatches will be male or female. They randomly sample turtle eggs, and record the incubation temperature (in Celsius) and the sex of the turtle that hatches. The following histogram shows the distribution of temperatures based on the sex of the turtle.

*You can assume that 100% of the data is captured in this visualization.*




(a) (3 points) In the sample, more than 50% of the male turtles were incubated at a temperature between 29.5 and 30.0 degrees.

- ◯ True.
- ★ **False.**
- ◯ This is not possible to determine based on the provided information.

(b) (3 points) In this sample, the number of male turtles with incubation temperatures between 29.5 and 30 degrees is the same as the number of female turtles incubated between 30.5 and 31 degrees.

- ◯ True.
- ◯ False.
- ★ **This is not possible to determine based on the provided information.**

(c) (3 points) If the bins used to form the histogram for male turtles were replaced with a single bin from 23 to 33, how tall would the resulting bar be? Make sure to include the units.

**Sample Solution:** The width of the bin would be 33 - 23 = 10 degrees. The bin would create 100% of the male turtle data. Together, this means that the height of the resulting bar would be 100% / 10 degrees = 10 percent per degree.

# Section B

**For this section, your goal is to provide Python code that could be run in our notebooks that will produce the answer to the questions asked. In most cases, we have provided a template to get you thinking. You can alternatively ignore the template and write your own code from scratch.**

6. In San Francisco, the Existing Buildings Energy Performance Ordinance (Environment Code Chapter 20) requires that each non-residential building with at least 10,000 square feet of conditioned (heated or cooled) space and each residential building with at least 50,000 square feet of conditioned space must be benchmarked using Energy Star Portfolio Manager annually. Each non-residential building specified above is also required to undergo an energy audit or retrocommissioning at least once every 5 years.

   The table `building_data` contains relevant San Francisco building information and 2021 energy use (measured in thousands of BTUs (British thermal units)). On the table reference page, you can see a preview of this table.

   (a) (4 points) How many 'Commercial' buildings are there in `building_data`.

   ```
   commercial_buildings = __________.__________(__________, __________)
   commercial_buildings.___________
   ```

   **Sample Solution:**

   ```
   commercial_buildings = building_data.where('property_type', 'Commercial')
   commercial_buildings.num_rows
   ```

   (b) (4 points) What is the address for the building with the largest floor area?

   ```
   sorted_data = __________.__________(__________, __________)
   __________.__________(__________).__________
   ```

**Sample Solution:**

```
sorted_data = building_data.sort('floor_area', True)
sorted_data.column('building_address').item(0)
```

(c) (2 points) You've received a CSV file called `zip_code.csv`. Write code that will create a table called `zip_codes` from that CSV file that contains all the information in the `zip_code.csv` file.. On the table reference page, you can see a preview of what `zip_codes` looks like. Zip codes and postal codes are equivalent in this context.

`zip_codes = _____`

**Sample Solution:**

```
zip_codes = Table.read_table('zip_codes.csv')
```

(d) (3 points) Use the `join` method to create a table called `building_data_geo` that adds the latitude, longitude, and population estimate information from `zip_codes` to the data in `building_data`. You do not need to do any additional sorting or re-ordering beyond using the join method. On the table reference page, you can see a preview of what `building_data_geo` should look like.

`building_data_geo = _____`

**Sample Solution:**

```
building_data_geo = building_data.join('postal_code', zip_codes, 'zip')
```

(e) (3 points) When reading the data, it seems that Python assumed the postal code (zip code) values were numerical. Write code that will check if the data type of the values in the `postal_code` column of `building_data_geo` is `float`. Your code should output the `bool` value `True` or `False`. As a hint, `type(2.0)` would evaluate to be `float`.

**Sample Solution:**

```
type(building_data_geo.column('postal_code').item(0)) == float
```

(f) (4 points) The postal codes in `building_data_geo` are actually float values, but they need to be strings. Create a function called `float_to_str` that takes a float and returns a string version of the float ignoring any decimal part.

For example, `float_to_str(94118.0)` should return `'94118'`.

Hints: `str(94118.0)` would create the string `'94118.0'`, not `'94118'`. Also, `int(94118.0)` would produce the integer `94118`.

**Sample Solution:**

```
def float_to_str(a_float):
    return str(int(a_float))
```

(g) (3 points) Use the `float_to_str` function to create an array called `postal_codes` of the postal codes formatted as strings.

```
postal_codes = ________________________________________
```

**Sample Solution:**

```
postal_codes = building_data_geo.apply(float_to_str, 'postal_code')
```

(h) (3 points) Update the `building_data_geo` table such that postal code values in the `'postal_code'` column are strings, not floats.

Remember that `postal_codes` is an array of the postal codes as strings.

**Sample Solution:**

```
building_data_geo = building_data_geo.with_column('postal_code', postal_codes)
```

(i) (3 points) Create a bar chart showing the distribution of the postal codes in the `building_data_geo` table. Make sure the bars are in order such that the longest bars are at the top of the visualization.

```
building_data_geo_by_zip = ___________________________________
___________________.__________________________.________________________
```

**Sample Solution:**

```
building_data_geo_by_zip=building_data_geo.group('postal_code')
building_data_geo_by_zip.sort('count', True).barh('postal_code')
```

(j) (4 points) Create a table with two columns showing the median energy use for 2021 for each postal code based on the data in `building_data_geo`. Your table should have a row for each postal code showing the median energy use for the buildings with that postal code.

```
reduced_data = _________________.____________(_____________, _________________)
_________________.____________(_________________, ___________________)
```

**Sample Solution:**

```
reduced_data = building_data_geo.select('postal_code', 'energy_use_2021')
reduced_data.group('postal_code', np.median)
```

(k) (3 points) Using the data in `building_data_geo`, create a visualization to show the relationship between the floor area of a building and it energy usage.

```
building_data_geo.______________.(____________________, ___________________)
```

**Sample Solution:** `building_data_geo.scatter('floor_area', 'energy_use_2021')`

7. (3 points) One way to support concluding whether or not simulated data supports a model is to compare a related p-value to a specified cutoff percentage. Write a function called `consistent` with arguments `p_value` (`float`) and `cutoff` (`float`) that returns `False` if the `p_value` is less than the `cutoff` value (signifying support against the null model). It will return `True` otherwise.

```
def consistent(p_value, cutoff):
    ________________________:
        return False
    ________________________:
        return True
```

**Sample Solution:**

```
def consistent(p_value, cutoff):
    if p_value < cutoff:
        return False
    else:
        return True
```

8. (4 points) Create a function called `roll` with arguments `k`, `n`, and `trials` that simulates trials (the number of trials) rolls of n fair 6-sided dice, and each time counts how many of those dice show k or higher, and then displays an empirical histogram of those counts.

   For example, if `k` is 5, `n` is 3, and rolling 3 dice results in a 6, a 4, and a 5, then 2 of the 3 dice are 5 or larger (the 6 and the 5). So, `roll(5, 3, 10_000)` would output a histogram created by repeating simulation 10,000 times.

```
def roll(k, n, trials):
    """Repeatedly roll n dice and check how many results are k or larger."""
```

```
outcomes = make_array()
possible_results = np.arange(1, 7)
for _________________________________
    rolls = _______________________________
    outcomes = _____________________(outcomes, np.count_nonzero(rolls >= ______))
    Table().with_column('Outcomes', ______________).________(bins=np.arange(30))
```

**Sample Solution:**

```
def roll(k, n, trials):
    """Repeatedly roll n dice and check how many results are k or larger."""
    outcomes = make_array()
    possible_results = np.arange(1, 7)
    for i in np.arange(trials):
        rolls = np.random.choice(possible_results, n)
        outcomes = np.append(outcomes, np.count_nonzero(rolls >= k))
        Table().with_column('Outcomes', outcomes).hist(bins=np.arange(30))
```

# Section C

9. In a game called September, players take turns selecting tokens and making moves based on the selected tokens. During each player's turn, they randomly select two tokens from a container, makes a play based on the two tokens, and then put all the tokens back in the container for the next player. The distribution of tokens is:

   - Earth Token: 21 Tokens
   - Wind Token: 12 Tokens
   - Fire Token: 1 Token

   (a) (3 points) What is the probability that a player will select no Wind tokens when it is their turn?

   **Sample Solution:** `(22 / 34) * (21 / 33)`

   (b) (3 points) What is the probability that a player will select 2 Fire tokens when it is their turn?

   **Sample Solution:** `0`

   (c) (3 points) What is the probability that a player will select at least one Wind token when it is their turn?

   **Sample Solution:** `1 - (22 / 34) * (21 / 33)`

10. According to a recent survey, 28% of surveyed adults in the United States use LinkedIn. For the sake of this question, assume that the chance of a randomly sampled adult in the United States being a LinkedIn user is 28% (independently of all others).

    (a) (3 points) For which sample size below is there a higher chance that a random sample of that size will contain a percent of LinkedIn users of more than 50%?

    ★ **200**

    ◯ 1000

(b) (3 points) According to the Law of Large Numbers (Law of Averages), with a smaller sample size the percentage of surveyed adults in that sample that use LinkedIn is more likely to be closer to 28% than a larger sample size.

◯ True

★ **False**

11. In the United States, 31% of adults report being online almost constantly. A team of data scientists took a random sample of 100 adults in San Francisco and found that 37 of them reported being online almost constantly.

- One member of the team says, "The percent of San Francisco adults who are online almost constantly is more than the percent of adults nationwide."
- Another member of the team says, "No, it's just chance." (This suggests that the 37% observed in this sample is just due to chance.)

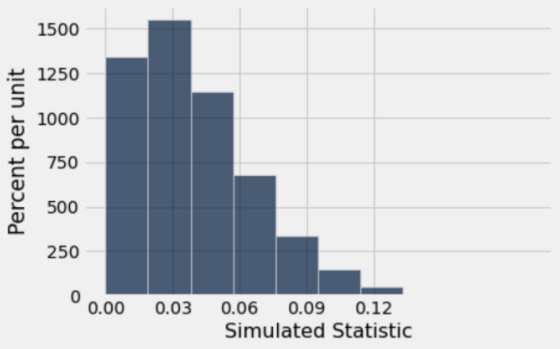
In order to decide between these two positions, the data scientists will conduct a test of hypotheses.

(a) (3 points) Provide a complete sentence for the null hypothesis.

**Sample Solution:** The sample of adults from San Francisco is drawn randomly from a population where 31% of them are labeled "online almost constantly."

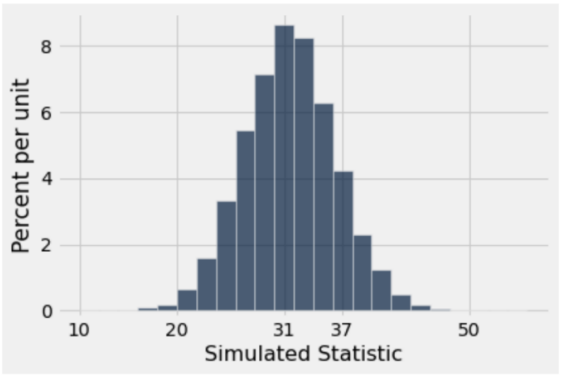
(b) (3 points) In order to decide between their two hypotheses, the data scientists have picked an appropriate test statistic and simulated it 10,000 times under appropriate conditions. Choose the most appropriate graph of the 3 options shown below that could be the histogram of their simulated values.

*Note that in each graph, some relevant values are labeled on the horizontal axis.*

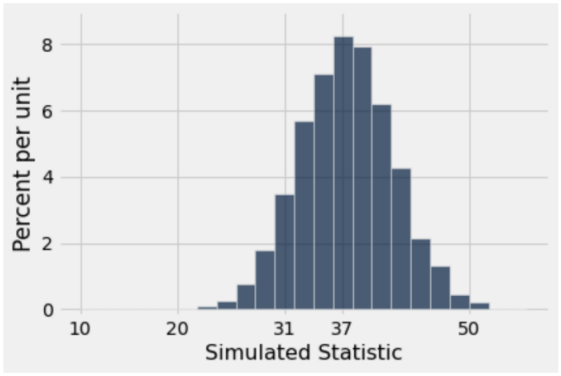



**Testing Option A**

○




★ **Testing Option B**




○ **Testing Option C**

(c) (4 points) Using one or two sentences, explain why the graph you selected is the correct graph. Make sure to address the properties of the graph you selected that made you choose it.

**Sample Solution:** The correct histogram is simulated assuming the null hypothesis is true, so the most likely value must be the value under the null. Additionally, it must contain values both above and below this peak, because the alternative is directional. Option B is the only graph that satisfies both properties. ALTERNATIVELY Option A is wrong because absolute value does not allow us to determine whether the proportion in the sample is greater or less than the proportion in the population. Option C is wrong because it is not simulated under the null - we can tell because it is centered at 37.

(d) (3 points) The 10,000 simulated values of the data scientists' test statistic are in the array `SIM_STAT_ARR`. The code `np.count_nonzero(SIM_STAT_ARR >= 37) / 10000` produces a value of `0.22`. What is the most appropriate conclusion for this hypothesis test?

★ **The test results favor the null hypothesis.**

◯ The test results favor the alternative hypothesis.

◯ It is impossible to make a decision based on the provided information.

# Table Reference

The table `building_data` contains 9 columns. The values in the columns `parcel_s`, `building_name`, `building_address`, `property_type`, `energy_audit_due_date` have a `str` data type. The values in the rest of the columns `int` or `float` data types.

| parcel_s | building_name | building_address | postal_code | floor_area | property_type | year_built | energy_audit_due_date | energy_use_2021 |
|---|---|---|---|---|---|---|---|---|
| 0010/001 | 2801 Leavenworth Street | 2801 LEAVENWORTH ST | 94109 | 133675 | Commercial | 1907 | 2024-04-01T00:00:00.000 | 6.21001e+06 |
| 0010/002 | Argonaut Hotel-SV | 495 JEFFERSON ST | 94109 | 180000 | Commercial | 1907 | 2025-04-01T00:00:00.000 | 7.34107e+06 |
| 0011/008 | Anchorage Garage | 500 BEACH ST | 94133 | 198525 | Commercial | 1974 | 2024-04-01T00:00:00.000 | 1.88699e+06 |

... (590 rows omitted)

The `zip_codes` table contains 4 columns. All the values in this table are either `float` or `int` data type.

| zip | latitude | longitude | irs_estimated_population |
|---|---|---|---|
| 94102 | 37.78 | -122.42 | 21610 |
| 94103 | 37.77 | -122.41 | 22940 |
| 94104 | 37.79 | -122.4 | 1720 |

... (48 rows omitted)

At some point, you are asked to create the table `building_data_geo`. It should look like:

| postal_code | parcel_s | building_name | building_address | floor_area | property_type | year_built | energy_audit_due_date | energy_use_2021 | latitude | longitude | irs_estimated_population |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 94102 | 0296/001 | 449 Powell Street | 449 POWELL ST | 34173 | Commercial | 1913 | 2024-04-01T00:00:00.000 | 2.08193e+06 | 37.78 | -122.42 | 21610 |
| 94102 | 0296/005 | Chancellor Hotel | 433 POWELL ST | 46800 | Commercial | 1914 | 2021-04-01T00:00:00.000 | 3.01398e+06 | 37.78 | -122.42 | 21610 |
| 94102 | 0296/006 | 400 POST ST | 400 POST ST | 61807 | Commercial | 1909 | 2020-04-01T00:00:00.000 | 9.32405e+06 | 37.78 | -122.42 | 21610 |

... (590 rows omitted)