DATA 8 Foundations of Data Science

Spring 2024 — Midterm Exam

INSTRUCTIONS

You have 1 hour and 50 minutes to complete the exam.

- The exam is closed book, closed notes, closed computer/calculator, except for the provided 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, exam, and student ID 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 **fill in** circles and squares to indicate answers and clearly cross out or erase mistakes.

Preliminaries

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. (22.0 points) True or False

(a) (2.0 pt) All of the work on this exam is your own.

◯ True

◯ False

(b) (2.0 pt) When a histogram is right skewed, the median is usually larger than the average.

◯ True

◯ False

(c) (2.0 pt) By default, the area of each bar in a histogram generated by `.hist()` in the datascience package always represents the percent of data within the corresponding bin.

◯ True

◯ False

(d) (2.0 pt) If event A is the complement of event B, then the chance of either A or B happening is always 100%.

◯ True

◯ False

(e) (2.0 pt) Standing outside Chipotle on a Friday evening, flipping a coin for each customer that walks out, and surveying a customer every time you flip "Heads" would constitute a random sample of evening Chipotle customers.

◯ True

◯ False

(f) (2.0 pt) When writing a `for` loop in Python, indenting a block with multiple lines of code after the `for` statement is optional.

◯ True

◯ False

(g) (2.0 pt) When using difference of means for a hypothesis test, the p-value will be the same regardless of whether you define your test statistic as group A minus group B or group B minus group A.

◯ True

◯ False

(h) (2.0 pt) The empirical distribution of a categorical variable contains the unique values of the variable along with how often each value is observed.

◯ True

◯ False

(i) **(2.0 pt)** In lecture, we could not conclude that smoking causes a decrease in birth weights of babies because we did not randomly assign mothers into smoking and non-smoking groups.

◯ True

◯ False

(j) **(2.0 pt)** If you sample all rows from a table without replacement, the average value of any numerical column will be the same as taking the average of that numerical column from the original table.

◯ True

◯ False

(k) **(2.0 pt)** Suppose a table contains two categorical columns that each have 3 unique categories. If you're trying to look at the counts of each combination of these 2 columns, using a pivot table will result in the same or smaller number of rows than using a multi-column group.

◯ True

◯ False

2. (30.0 points) Game Show

Joel and Ellie have been selected to participate in a quiz show!

There are four topics on the show: art, history, science, and pop culture. Every time a question is asked, it is randomly chosen from one of these four topics (each topic is equally likely to be selected).

Joel and Ellie have different strengths. The probability of them getting particular types of questions right is listed below:

Category	Joel	Ellie
Art	0.7	0.15
History	0.75	0.4
Science	0.9	0.25
Pop Culture	0.1	0.9

Before a question is revealed, Joel and Ellie will flip a fair coin to determine who will answer the question (if it's "Heads", Joel will answer the question).

(a) (13.0 points)

They are trying to determine the likelihood of them winning the show.

i. (2.0 pt) What is the probability that the first question is about art and Ellie answers?

You can write your answer as a mathematical expression without simplifying it.

ii. (3.0 pt) What is the probability that the first question is about science, Joel answers, and Joel answers correctly?

You can write your answer as a mathematical expression without simplifying it.

iii. (4.0 pt) What is the probability that the first question is about pop culture and the team (either Joel or Ellie) answers correctly?

- ◯ `0.25 * (0.1 + 0.9)`
- ◯ `(0.5 * 0.1 + 0.5 * 0.9)`
- ◯ `0.25 + (0.5 * 0.1 + 0.5 * 0.9)`
- ◯ `0.25 * (0.5 * 0.1 + 0.5 * 0.9)`
- ◯ None of the Above

iv. (4.0 pt) What is the probability that Ellie answers the first question and she answers correctly?

- ◯ `0.5 + (0.15 + 0.4 + 0.25 + 0.9)`
- ◯ `0.5 * (0.15 + 0.4 + 0.25 + 0.9)`
- ◯ `0.5 + 0.25 + (0.15 + 0.4 + 0.25 + 0.9)`
- ◯ `0.5 * 0.25 * (0.15 + 0.4 + 0.25 + 0.9)`
- ◯ None of the Above

(b) (14.0 points)

Suppose that Joel gets sick, so Ellie has to compete alone!

Joel writes a Python function for Ellie to simulate whether she will beat the current top score of correct answers out of 20 questions. For example, if the top score is `18`, Ellie could call `did_beat_top_score(18)` to simulate whether she would beat the score.

Fill in the blanks for the function below, which goes through the process of simulating her answering 20 randomly selected questions and returns `True` if she did better than the top score or `False` if she did not.

```
def did_beat_top_score(top_score):

    types = make_array('art', 'history', 'science', 'pop')

    correct_probs = make_array(0.15, 0.4, 0.25, 0.9)

    questions = np.random.choice(_________, _________)
                                    (a)        (b)

    total_num_correct = 0

    for i in np.arange(len(types)):

        question_type = types.item(i)

        num_questions_for_type = np.count_nonzero(_________)
                                                     (c)

        prob = correct_probs.item(i)

        correct_dist = make_array(prob, 1 - prob)

        simulation = sample_proportions(num_questions_for_type, correct_dist)

        prop_correct_for_type = simulation._________
                                              (d)

        total_num_correct = total_num_correct + _________
                                                   (e)

    return _________
             (f)
```

i. (1.0 pt) Fill in blank (a).

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

iii. **(3.0 pt)** Fill in blank (c).

iv. **(3.0 pt)** Fill in blank (d).

v. **(3.0 pt)** Fill in blank (e).

vi. **(3.0 pt)** Fill in blank (f).

vii. **(3.0 pt)** Suppose that the top score is now 19 and Ellie plans to run the following code:

```
outcomes = make_array()

for i in np.arange(10000):
    outcomes = np.append(outcomes, did_beat_top_score(19))

observed_proportion = np.sum(outcomes) / len(outcomes)
```

Which of the following should she expect `observed_proportion` to be equal to?

◯ `1 - (0.15 + 0.4 + 0.25 + 0.9)^20`

◯ `(0.15 + 0.4 + 0.25 + 0.9)^20`

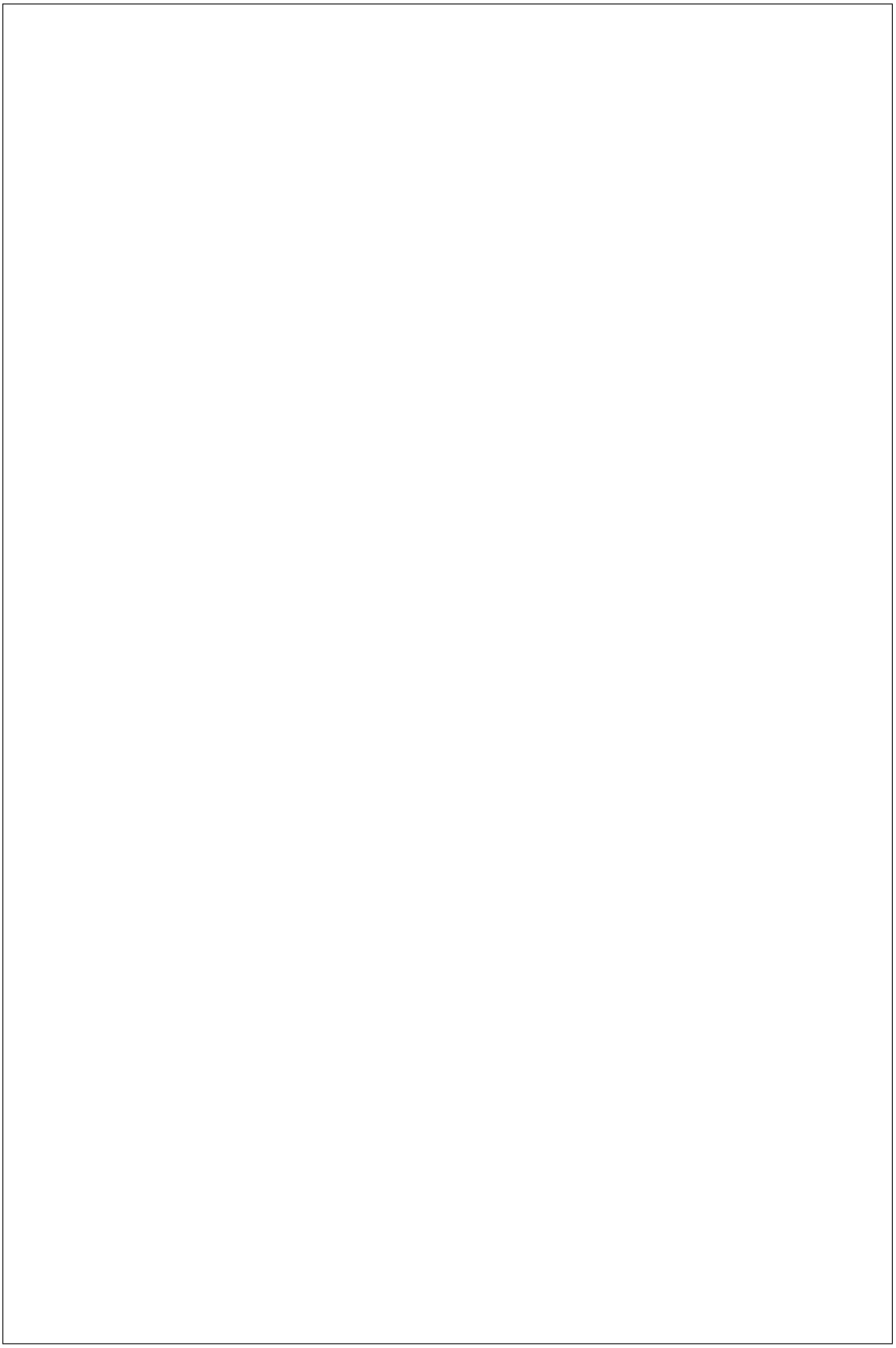
◯ `(0.25 * (0.15 + 0.4 + 0.25 + 0.9))^(20)`

◯ `1 - (0.25 * (0.15 + 0.4 + 0.25 + 0.9))^20`

◯ `0.25 * (0.15 + 0.4 + 0.25 + 0.9)`

3. (0.0 points) Quick Break: Just For Fun

(a) (Optional) Take a break and draw a picture of your experience in *Data 8* so far! Make sure to finish the rest of the exam, though!



4. (30.0 points) Oscars

Troy and Abed are analyzing movies leading up to the Oscars.

They randomly sample ratings of Oscar-nominated movies from IMDB and stores this in a table called `ratings`.

The first few rows of the table are provided below:

Movie	User	Rating
Barbie	cerave24	8
Oppenheimer	francescascorsese	1
Killers of the Flower Moon	nolanforlyfe	4
Spider-Man: Across the Spiderverse	haileebaby96	10

... (949 rows omitted)

The table has the following columns:

- *Movie*: (string) the name of the movie
- *User*: (string) the IMDB username of the person who rated the movie
- *Rating*: (int) the user's rating for the movie, between 1 and 10

(a) (2.0 pt) Abed wants to see the median rating for each movie. Which of the following visualizations could be used for this?

Select all that apply.

☐ Histogram

☐ Scatterplot

☐ Line Plot

☐ Bar Chart

☐ None of the Above

(b) (8.0 points)

Abed wants to see the names of the movies that have at least 20 users in the sample who gave a rating of 2 or less for that movie.

He writes the following partially completed code, which returns these names of movies in an array:

```
movies = ratings._________.group(_________).where(1, _________)
                    (a)             (b)                 (c)

movies._________
          (d)
```

i. (3.0 pt) Fill in blank (a)

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

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

iv. **(2.0 pt)** Fill in blank (d)

(c) (10.0 points)

Troy is curious how the ratings vary by other movie information like genre & studio, so he puts together another table called `movies`, which contains all of the movies released in 2023.

The table has the following columns:

- *Title*: (string) the name of the movie
- *Genre*: (string) the genre of the movie (either "Comedy" or "Drama")
- *Studio*: (string) the studio that made the movie (e.g., "Sony", "Fox")

He writes the following partially completed code to determine the average rating of the Oscar-nominated Comedy movies:

```
info_and_ratings = ratings.join(_________)
                                    (a)

avg_ratings_by_genre = info_and_ratings.select('Genre', 'Rating')._________
                                                                    (b)

avg_comedy_rating = avg_ratings_by_genre.sort(0).column(_________)._________
                                                  (c)        (d)
```

Recall: The `ratings` table has columns *Movie*, *User* and *Rating*.

i. (3.0 pt) Fill in blank (a).

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

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

iv. (2.0 pt) Fill in blank (d).

(d) (4.0 pt) Assume that the `info_and_ratings` table in the previous question was correctly constructed.

Troy next wants to create a table where each unique studio gets its own row, each unique genre its own column, and the values inside the table correspond to the average IMDB rating for each combination of studio and genre.

He writes the following partially completed code:

```
info_and_ratings._________
```

Recall: The `ratings` table has columns *Movie*, *User* and *Rating* and the `movies` table has columns *Title*, *Genre* and *Studio*.

(e) (6.0 points)

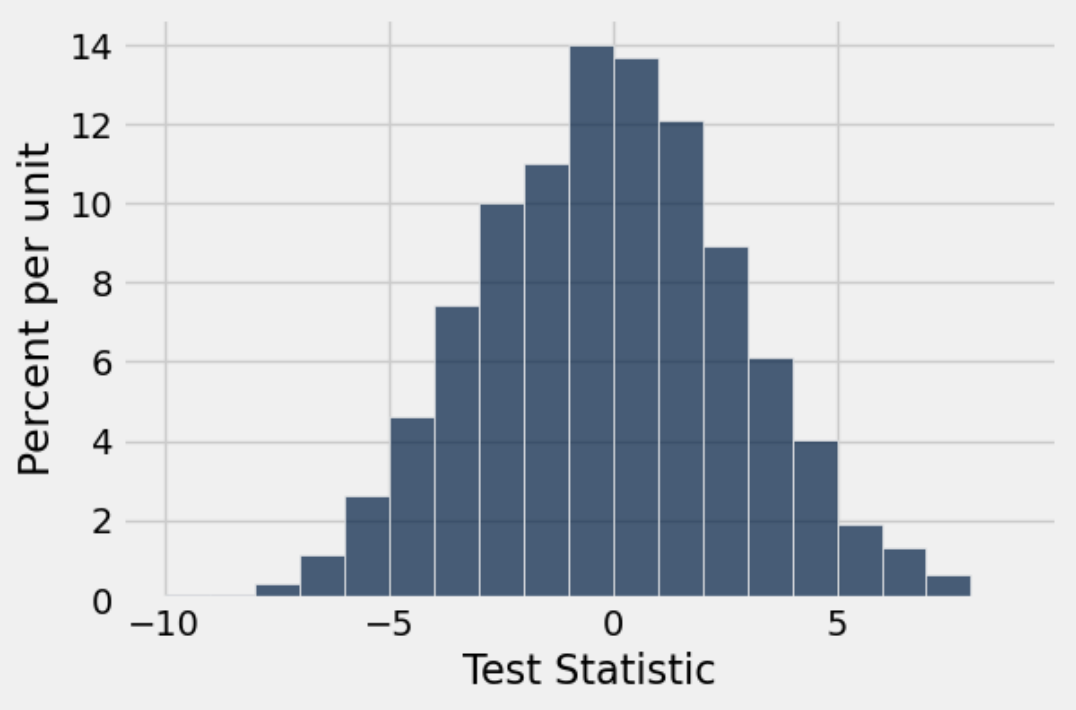
Suppose Abed wants to test the hypothesis that among Oscar-nominated films, Comedy movies get lower IMDB ratings than Drama movies.

i. (2.0 pt) Which of the following test statistics could he use?

Select all that apply.

- ☐ Mean *Rating* for Comedy movies
- ☐ Mean *Rating* for Comedy movies minus mean *Rating* for Drama movies
- ☐ Absolute difference between mean *Rating* for Comedy movies and mean *Rating* for Drama movies
- ☐ Mean *Rating* for Drama movies minus mean *Rating* for Comedy moviees
- ☐ None of the above because the sample size of Comedy ratings is different from the sample size of Drama movies

ii. (4.0 pt) Suppose Abed carries out his hypothesis test and generates the following histogram of simulated test statistics using equally spaced bins.



Suppose the observed test statistic is 5 and larger values are in favor of the alternative.

Using a **5%** cutoff, which of the following conclusions can he make?

Select all that apply.

☐ The alternative hypothesis is true.

☐ Comedy movies get lower ratings on average compared to Drama movies.

☐ Oscar-nominated Drama movies get higher ratings on average than Oscar-nominated Comedy movies do.

☐ In the original sample, Drama movies received higher ratings than Comedy movies did.

☐ The distribution of ratings is the same between Oscar-nominated Drama movies and Oscar-nominated Comedy movies.

☐ None of the Above.

5. (27.0 points) Chipotle

Edwin and Gamy want to compare the burritos from two Chipotle locations, "Shattuck" and "Telegraph." They want to determine if there are any differences in the burritos served between these two locations.

They put together a table called `burritos`, which contains 250 randomly sampled burrito orders from the past year. Here are the first few rows:

Location	Month	Weight	Guac
Telegraph	Oct	22.2	Regular
Shattuck	Mar	23.7	Extra
Telegraph	Apr	24.9	None
Telegraph	Jan	19.7	None
Shattuck	Oct	25.4	Extra

... (245 rows omitted)

The table contains the following columns:

- *Name*: (string) the name of the Chipotle location
- *Month*: (string) the month of the year the burrito was made
- *Weight*: (float) the weight of the burrito (in ounces)
- *Guac*: (string) how much guacamole was added (`'None'`, `'Regular'` or `'Extra'`)

Gamy notices that burritos served in the `'Telegraph'` location seem to weigh more than those in `'Shattuck'`. Edwin thinks the differences observed in the sample are only due to chance.

(a) (3.0 pt) Which of these are **null** hypotheses that Gamy could use to assess his claims?

Select all that apply.

☐ Burritos with high weight have the same *Location* distribution as burritos with low weight.

☐ `'Shattuck'` burritos have the same weight distribution as `'Telegraph'` burritos.

☐ `'Shattuck'` burritos have lower weight than `'Telegraph'` burritos on average.

☐ `'Telegraph'` burritos have the same weight distribution as `'Shattuck'` burritos.

☐ None of the above.

(b) (3.0 pt) Which of the following function calls would be helpful when simulating data under the null hypothesis?

Assume `props` is an array containing the categorical distribution of *Location* in `burritos`.

Select all that apply.

☐ `burritos.sample()`

☐ `burritos.sample(with_replacement=False)`

☐ `burritos.sample(250, with_replacement=False)`

☐ `sample_proportions(250, props)`

☐ None of the above.

(c) (5.0 points)

Suppose Gamy decides to use a test statistic such that **lower values** are in favor of the alternative hypothesis.

He simulates 1,000 values of the test statistic and assigns them to the array `test_stats`. He then writes the following code, which assigns the p-value to `p_val`:

```
num_extreme_test_stats = ___________
                              (a)
p_val = num_extreme_test_stats / ___________
                                      (b)
```

Assume the observed test statistic has been assigned to `observed_stat`.

i. (3.0 pt) Fill in blank (a).

ii. (2.0 pt) Fill in blank (b).

(d) (3.0 pt) Suppose that `num_extreme_test_stats` from the previous question is equal to `39`.

Assuming he uses a **5%** cutoff, which of the following can Gamy conclude?

Select all that apply.

- ☐ In the population, `'Shattuck'` burritos and `'Telegraph'` burritos have the same *Weight* distribution.
- ☐ In the population, burritos with high weight have a different *Location* distribution than burritos with low weight.
- ☐ In the population, `'Telegraph'` burritos have higher *Weight* on average compared to `'Shattuck'` burritos.
- ☐ In the sample, `'Shattuck'` burritos have a lower *Weight* on average compared to `'Telegraph'` burritos.
- ☐ In the population, `'Telegraph'` burritos always have higher *Weight* than `'Shattuck'` burritos.
- ☐ None of the Above.

(e) (10.0 points)

Edwin notices that `'Telegraph'` burritos have a different *Guac* distribution than `'Shattuck'` burritos.

Gamy thinks the differences observed in the sample are only due to chance.

Edwin wants to make a function to calculate the total variation distance of the *Guac* distributions between `'Telegraph'` and `'Shattuck'` burritos.

He writes the following partially completed code:

```
def props(values):
    # This function converts an array of counts into an array of proportions.

    return _________
              (a)

def tvd(data, category_a, category_b):

    dist_a = data.where('Location', category_a)._________
                                                  (b)

    counts_a = dist_a.sort(0)._________
                                 (c)

    dist_b = data.where('Location', category_b)._________
                                                  (b)

    counts_b = dist_b.sort(0)._________
                                 (c)

    abs_diff = abs(props(counts_a) - props(counts_b))

    return _________
              (d)
```

Note: The argument for ***`props`*** *is* ***`values`*** *(array). The arguments for* ***`test_stat`*** *are* ***`data`*** *(Table),* ***`category_a`*** *(string) and* ***`category_b`*** *(string).*

i. **(3.0 pt)** Fill in blank (a).

ii. **(2.0 pt)** Fill in blank (b).

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

iv. **(3.0 pt)** Fill in blank (d).

(f) **(3.0 pt)** Edwin creates a histogram of `test_stats` and uses the area principle to calculate that at least 6% of the values are greater than `0.47`.

If his p-value cutoff is **5%** and the observed test statistic is `0.49`, which of the following can he conclude? *Select all that apply.*

- ☐ The data are consistent with the null hypothesis.
- ☐ The data are consistent with the alternative hypothesis.
- ☐ The null hypothesis is true.
- ☐ The null hypothesis is false.
- ☐ There is not enough information to make any of these conclusions.

6. (0.0 points) Optional

(a) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question (including the title of the section, e.g., Experiments) and state your assumptions. Be warned: We only plan to consider this information if we agree that the question was erroneous or ambiguous and we consider your assumption reasonable.