Lecture 13 – Midterm Review

DSC 10, Fall 2023

Announcements

- The Midterm Exam is **this Monday during lecture**. See **this post on Ed** for lots of details, including what is covered, what to bring, and how to study.
- Homework 3 is due **tomorrow at 11:59PM**.
  - Finish Homework 3 before the exam, since the material on it is all in scope for the exam.
- The Midterm Project is due on **Saturday 11/4 at 11:59PM**. Only one partner needs to submit.
- Quiz 2 and Homework 2 scores have been released. Along with them, we've released a Grade Report on Gradescope, which summarizes your scores in the class so far. See **this post on Ed** for details.

Agenda

- We'll work through selected problems from the Spring 2023 Midterm.
- We won't write any code, since you can't run code during the exam. Instead, we'll try to think like the computer ourselves.
- These annotated slides will be posted after lecture is over.
- **Try the problems with us!**

Spring 2023 Midterm

Access the exam **here**. Make sure to read the data info sheet at the top before starting.

Problem 4.1

Consider the following block of code.

number of students

could have separate rows

2 4

2 5

```
A = survey.shape[0]
B = survey.groupby(["Unread Emails", "IG Followers"]).count().shape[0]
```

Suppose the expression `A == B` evaluates to `True`. Given this fact, what can we conclude?

- There are no two students in the class with the same number of unread emails.
- There are no two students in the class with the same number of Instagram followers.
- There are no two students in the class with the same number of Instagram followers, and there are no two students in the class with the same number of unread emails.
- There are no two students in the class with both the same number of unread emails and the same number of Instagram followers.

number of unique combinations of unread emails and IG followers

Problem 4.2

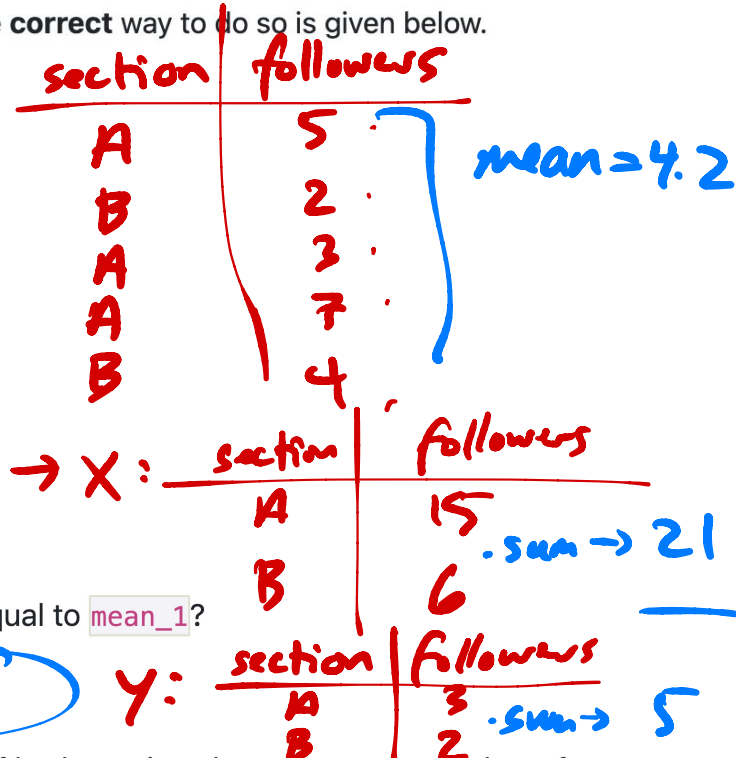
We'd like to find the mean number of Instagram followers of all students in DSC 10. One **correct** way to do so is given below.

```
mean_1 = survey.get("IG Followers").mean()
```

Another two **possible** ways to do this are given below.

```
# Possible method 1.
mean_2 = survey.groupby("Section").mean().get("IG Followers").mean()

# Possible method 2.
X = survey.groupby("Section").sum().get("IG Followers").sum()
Y = survey.groupby("Section").count().get("IG Followers").sum()
mean_3 = X / Y
```



Is `mean_2` equal to `mean_1`?

- Yes
- Yes, if both sections have the same number of students, otherwise maybe.
- Yes, if both sections have the same number of students, otherwise no.
- No.

what if everyone has the same # of followers?

Is `mean_3` equal to `mean_1`?

- Yes.
- Yes, if both sections have the same number of students, otherwise maybe.
- Yes, if both sections have the same number of students, otherwise no.
- No.

Teresa flips the coin 21 times and sees 13 heads and 8 tails. She stores this information in a DataFrame named `teresa` that has 21 rows and 2 columns, such that:

- The `"flips"` column contains `"Heads"` 13 times and `"Tails"` 8 times.
- The `"Wolftown"` column contains `"Teresa"` 21 times.

Then, Sophia flips the coin 11 times and sees 4 heads and 7 tails. She stores this information in a DataFrame named `sophia` that has 11 rows and 2 columns, such that:

- The `"flips"` column contains `"Heads"` 4 times and `"Tails"` 7 times.
- The `"Makai"` column contains `"Sophia"` 11 times.

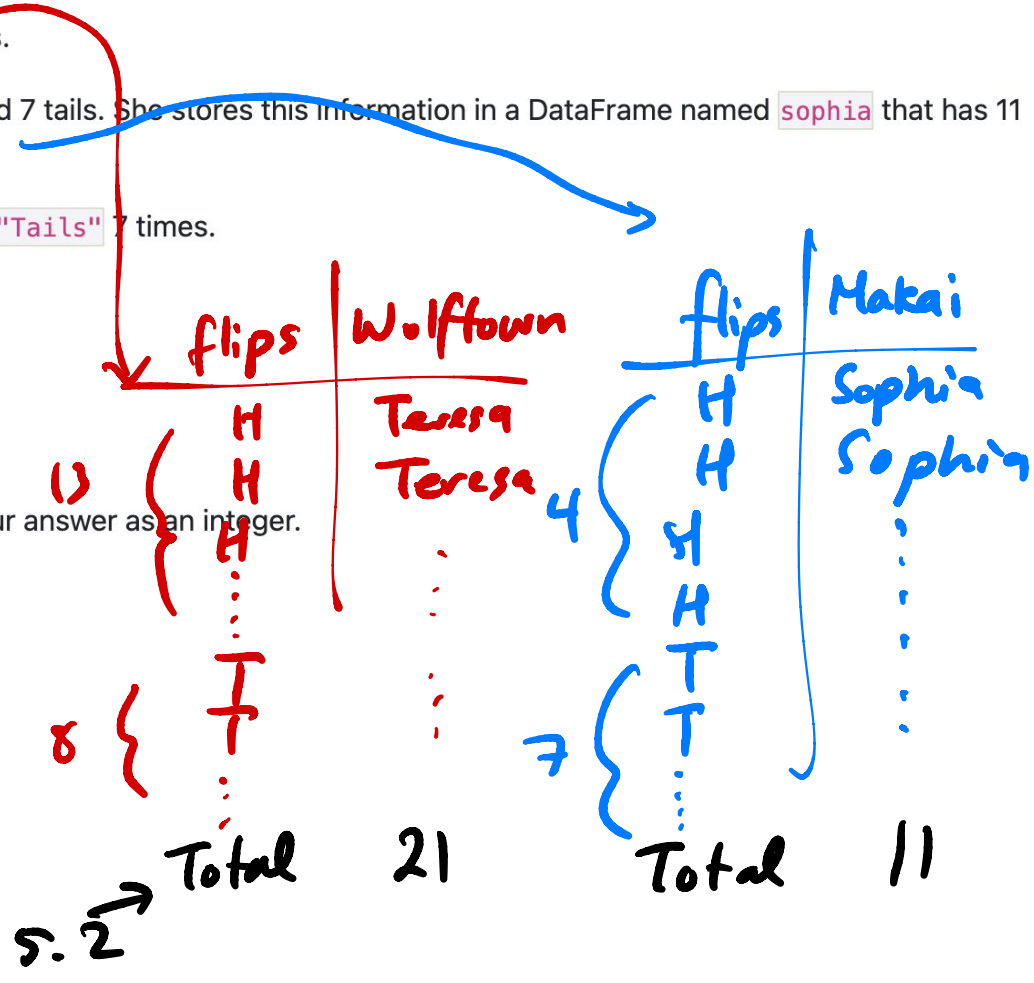
Problem 5.1

How many rows are in the following DataFrame? Give your answer as an integer.

```
teresa.merge(sophia, on="flips")
```

Hint: The answer is less than 200.

$13 \cdot 4 + 8 \cdot 7$
$= 52 + 56 = 108$



Problem 5.2

Let A be your answer to the previous part. Now, suppose that:

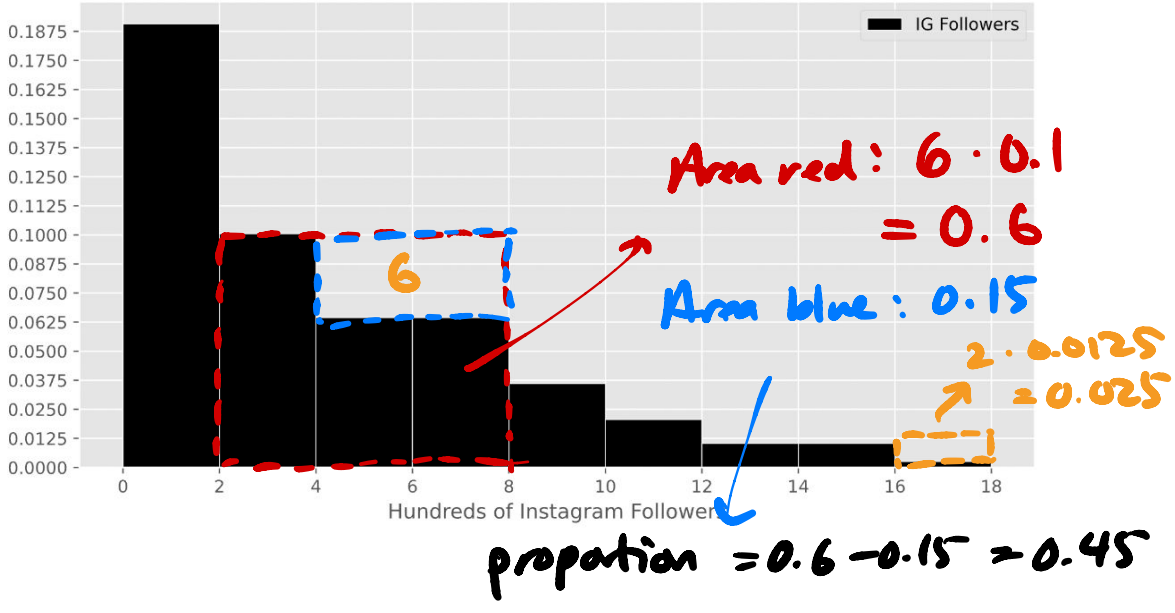
- `teresa` contains an additional row, whose `"flips"` value is `"Total"` and whose `"Wolftown"` value is 21.
- `sophia` contains an additional row, whose `"flips"` value is `"Total"` and whose `"Makai"` value is 11.

Suppose we again merge `teresa` and `sophia` on the `"flips"` column. In terms of A, how many rows are in the new merged DataFrame?

- A
- $A + 1$
- $A + 2$
- $A + 4$
- $A + 231$

Problem 6

The histogram below displays the distribution of the number of Instagram followers each student has, measured in 100s. That is, if a student is represented in the first bin, they have between 0 and 200 Instagram followers.



For this question only, assume that there are exactly 200 students in DSC 10.

students = 0.45 · 200 = 90

Problem 6.1

How many students in DSC 10 have between 200 and 800 Instagram followers? Give your answer as an integer.

Problem 6.2

Suppose the height of a bar in the above histogram is h. How many students are represented in the corresponding bin, in terms of h?

Hint: Just as in the first subpart, you'll need to use the assumption from the start of the problem.

- $20 \cdot h$
- $100 \cdot h$
- $200 \cdot h$
- $400 \cdot h$
- $800 \cdot h$

proportion in range = area of range = width · height

num students in range = proportion in range · 200

= width · height · 200

= 2 · h · 200

= 400h

can verify using answer to 6.1!

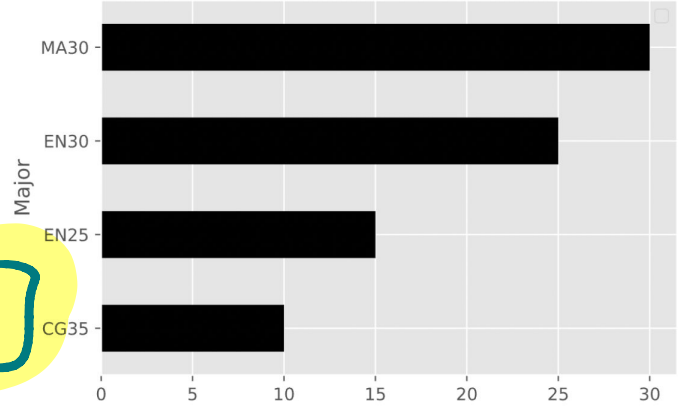
Problem 7

The four most common majors among students in DSC 10 this quarter are `"MA30"` (Mathematics - Computer Science), `"EN30"` (Business Economics), `"EN25"` (Economics), and `"CG35"` (Cognitive Science with a Specialization in Machine Learning and Neural Computation). We'll call these four majors "popular majors".

There are 80 students in DSC 10 with a popular major. The distribution of popular majors is given in the bar chart below.

CG35 10
EN25 15
EN30 25
MA30 30

[-4, -3, -2, -1]



Problem 7.1

Fill in the blank below so that the expression outputs the bar chart above.

```
(survey
 .groupby("Major").count()
 .sort_values("College")
 .take(____)
 .get("Section")
 .plot(kind="barh")
)
```

What goes in the blank?

AN25 1
EN25 2
EN30 10
EG 25 15
⋮ ⋮
popular 30

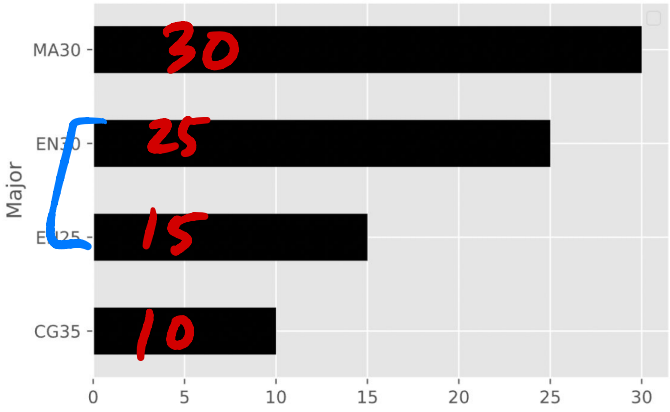
need bottom 4

Problem 7.2

Suppose we select **two** students in popular majors at random with replacement. What is the probability that both have `"EN"` in their major code? Give your answer in the form of a simplified fraction.

$P(\text{student has EN}) = \frac{\#\text{ with EN}}{\#\text{ total}} = \frac{40}{80} = \frac{1}{2}$

$P(\text{both have EN}) = \frac{1}{2} \cdot \frac{1}{2} = \boxed{\frac{1}{4}}$



Problem 7.3

Suppose we select **two** students in popular majors at random with replacement. What is the probability that we select one `"CG35"` major and one `"MA30"` major (in any order)?

- $\frac{1}{2}$
- $\frac{3}{4}$
- $\frac{3}{8}$
- $\frac{3}{16}$
- $\frac{3}{32}$
- $\frac{3}{64}$

$P(\text{CG35 then MA30}) + P(\text{MA30 then CG35})$

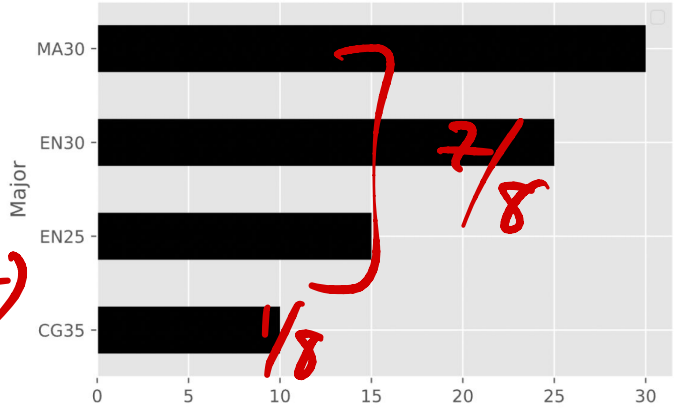
$= \frac{10}{80} \cdot \frac{30}{80} + \frac{30}{80} \cdot \frac{10}{80}$

$= \frac{3}{64} + \frac{3}{64} = \frac{6}{64} = \boxed{\frac{3}{32}}$

Problem 7.4

Suppose we select k students in popular majors at random with replacement. What is the probability that we select at least one `"CG35"` major?

- $\frac{7}{8}$
- $\frac{7^k}{8^k}$
- $\frac{1}{7^k}$
- $\frac{1}{8^k}$
- $\frac{8^k - 7^k}{7^k}$
- $\frac{8^k - 7^k}{8^k}$



$$P(\text{at least one CG35}) = 1 - P(\text{no CG35})$$

$$= 1 - \left[\frac{7}{8}\right]^k$$

$$= \frac{8^k}{8^k} - \frac{7^k}{8^k} = \frac{8^k - 7^k}{8^k}$$

Problem 8

We'd like to select three students at random from the entire class to win extra credit (not really). When doing so, we want to guarantee that the same student cannot be selected twice, since it wouldn't really be fair to give a student double extra credit.

Fill in the blanks below so that `prob_all_unique` is an estimate of the probability that all three students selected are in different majors.

Hint: The function `np.unique`, when called on an array, returns an array with just one copy of each unique element in the input. For example, if `vals` contains the values `1, 2, 2, 3, 3, 4`, `np.unique(vals)` contains the values `1, 2, 3, 4`.

```
unique_majors = np.array([])
for i in np.arange(10000):
    group = np.random.choice(survey.get("Major"), 3, __(a)__)
    __(b)__ = np.append(unique_majors, len(__(c)__))
prob_all_unique = __(d)__
```

Problem 8.1

What goes in blank (a)?

- `replace=True`
- `replace=False`

Problem 8.4

What could go in blank (d)? Select all that apply. At least one option is correct; blank answers will receive no credit.

- `(unique_majors > 2).mean()`
- `(unique_majors.sum() > 2).mean()`
- `np.count_nonzero(unique_majors > 2).sum() / len(unique_majors > 2)`
- `1 - np.count_nonzero(unique_majors != 3).mean()`
- `unique_majors.mean() - 3 == 0`