DATA 8
Fall 2021

Foundations of Data Science

Midterm
Solutions

INSTRUCTIONS
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!

Preliminaries
You can complete these questions before the exam starts.

(i) What is your full name?

(ii) What is your Student ID number?

(iii) Who is your Lab GSI?

Preliminaries

You can complete and submit these questions before the exam starts. Note '. . . ' can mean any code after the given variable.

(a) What is your full name?

(b) What is your student ID number?

(c) Who is your Lab GSI?

1. (18 points) Working with Tables

After the Data 8 midterm, Will, Eddie, and Melissa decide to get dinner at a restaurant in Berkeley, but they're having trouble deciding on a single place. They create a table of all Berkeley restaurants, `RESTS_TBL`, with four columns:

- "REST_NAME": The name of the restaurant
- "CUISTYP": The cuisine (type of food) served at this restaurant
- "Rating": The numerical rating given to the restaurant by the Daily Cal (a `float`)
- "Distance From Sproul": The distance, in miles, the restaurant is from Sproul Hall (a `float`)

REST_NAME	CUISTYP	Rating	Distance From Sproul
Imm Thai	Thai	9.9	0.2
Berkeley Social Club	Korean	8.7	0.8
Italian Homemade	Italian	7.9	1.1

(. . . 76 more rows)

(a) (3 pt) Help Will count how many restaurants there are for each cuisine. Write a line of code that outputs a table with two columns: one column with the type of cuisine, and one column containing a count of how many restaurants there are with that cuisine.

Reminder: the columns of `RESTS_TBL` are "REST_NAME", "CUISTYP", "Rating", and "Distance From Sproul".

```
RESTS_TBL.group("CUISTYP")
```

(b) (3 pt) Will wants to eat at the highest-rated restaurant. Write a line of code that evaluates to the name of the restaurant with the highest rating. (You can assume there is only one restaurant with the highest rating; there are no ties.)

Reminder: the columns of `RESTS_TBL` are "REST_NAME", "CUISTYP", "Rating", and "Distance From Sproul".

```
RESTS_TBL.sort("Rating", descending=True).column("REST_NAME").item(0)
```

(c) (3 pt) Melissa only wants to eat at a Thai restaurant. Write a line of code that evaluates to a table containing all four columns but only the rows for restaurants whose cuisine is "Thai".

Reminder: the columns of `RESTS_TBL` are "REST_NAME", "CUISTYP", "Rating", and "Distance From Sproul".

```
RESTS_TBL.where("CUISTYP", "Thai")
```

(d) **(3 pt)** Eddie didn't want to walk to any restaurants that were further than one mile away from Sproul. Fill in the code below to assign the variable `EDDIE_CHOICE` to a table containing only restaurants that are less than one mile from Sproul.

```
EDDIE_CHOICE = ...
```

Reminder: the columns of `RESTS_TBL` are "REST_NAME", "CUISTYP", "Rating", and "Distance From Sproul".

```
EDDIE_CHOICE = RESTS_TBL.where("Distance From Sproul",
are.below_or_equal_to(1))   or   EDDIE_CHOICE = RESTS_TBL.where("Distance From
Sproul", are.below(1))
```

(e) **(3 pt)** Will decides to randomly pick a restaurant from the restaurants that are less than one mile from Sproul. Write code to randomly pick a restaurant from the `EDDIE_CHOICE` table and assigns the variable `WILL_CHOICE` to the name of that restaurant.

```
WILL_CHOICE = ...
```

Reminder: the columns of `RESTS_TBL` are "REST_NAME", "CUISTYP", "Rating", and "Distance From Sproul".

```
# Some example solutions - any are OK
WILL_CHOICE = np.random.choice(EDDIE_CHOICE.column("REST_NAME"))
WILL_CHOICE = EDDIE_CHOICE.sample().column("REST_NAME").item(0)
WILL_CHOICE = EDDIE_CHOICE.sample(1).column("REST_NAME").item(0)
```

(f) **(3 pt)** Write a line of code that evaluates to the number of different cuisines that appear in the "CUISTYP" column of the `RESTS_TBL` table.

```
RESTS_TBL.group("CUISTYP").num_rows
```

2. (10 points) Arrays and Tables

Several Data 8 staff are reserving rooms for study groups. The `rooms` table has one row per room that can potentially be reserved:

Room	Capacity	Region
110MC Kresge	10	Northside
B4 Gardner	5	Central
Warbler, 435 Moffitt	4	Central

(. . . 223 more rows)

All room names are different and every room appears only once in the `rooms` table.

The `RESEVS` table has one row per reservation they have made:

STNAME	Room	DAYCOL	Time
Meghan	Quail, 431 Moffitt	Tuesday	10
Rita	C6 Gardner	Monday	3
Margaret	110MC Kresge	Friday	12

(. . . 47 more rows)

(a) (3 pt) Write a line of code that evaluates to the total capacity if we reserved every room in the `rooms` table.

Reminder: `rooms`'s columns are "Room", "Capacity", and "Region". `RESEVS`'s columns are "STNAME", "Room", "DAYCOL", and "Time".

```
sum(rooms.column("Capacity"))
```

(b) (3 pt) Write a line of code that evaluates to the number of reservations that TARGETPERSON has made.

Reminder: `rooms`'s columns are "Room", "Capacity", and "Region". `RESEVS`'s columns are "STNAME", "Room", "DAYCOL", and "Time".

```
RESEVS.where("STNAME", "TARGETPERSON").num_rows
```

(c) (4 pt) Write code that assigns the variable `TOP_REGION` to the region of campus that has the most number of reservations. Note that the "Region" column of the `rooms` table shows the campus region for each room.

```
TOP_REGION = ...
```

Reminder: `rooms`'s columns are "Room", "Capacity", and "Region". `RESEVS`'s columns are "STNAME", "Room", "DAYCOL", and "Time".

```
grouped = RESEVS.join(rooms, "Room").group("Region")
TOP_REGION = grouped.sort("count", descending=True).column("Region").item(0)
```

3. (11 points) Chances

Each morning, Noor grabs a mug from her cabinet for coffee during the day. She has 9 mugs in total: 3 each of the colors green, black, and white.

Each morning, Noor picks one mug at random from all 9 mugs regardless of the mugs she picks on other days.

In each question below, pick the correct answer.

(a) (3 pt) The weekend (Saturday and Sunday) is coming up. What is the chance that Noor picks a green mug on both those days?

- ○ $\frac{3}{9}$
- ○ $\frac{3}{9} + \frac{3}{9}$
- ● $\frac{3}{9} \times \frac{3}{9}$

(b) (4 pt) Noor brings her mug to each Data 8 lecture. Next week, Data 8 lectures will be on Monday, Wednesday, and Friday. What is the chance that Noor brings a black mug to at least one of the three lectures?

- ○ $\frac{3}{9} + \frac{3}{9} + \frac{3}{9}$
- ○ $\frac{3}{9} \times \frac{3}{9} \times \frac{3}{9}$
- ● $1 - (\frac{6}{9} \times \frac{6}{9} \times \frac{6}{9})$
- ○ $1 - (\frac{3}{9} \times \frac{3}{9} \times \frac{3}{9})$

(c) (4 pt) One of Noor's classes has online office hours in the morning. She will attend the office hours on Tuesday and Thursday next week, bringing her mug with her. What is the chance that the mugs she has on those two days are the same color?

- ● $\frac{3}{9}$
- ○ $\frac{3}{9} \times \frac{3}{9}$
- ○ $1 - (\frac{3}{9} \times \frac{6}{9})$

4. (9 points) Comparing Chances

In the United States, 28% of adults use LinkedIn. Suppose you sample US adults randomly so that each sampled adult has chance 0.28 of being a LinkedIn user independently of all the others.

(a) (2 pt) For which sample size below is there a higher chance that the percent of LinkedIn users in the sample will be at least 25%?

- ○ 200
- ● 400

(b) (2 pt) For which sample size below is there a higher chance that the percent of LinkedIn users in the sample will be at least 50%?

- ● 200
- ○ 400

(c) (2 pt) For which sample size below is there a higher chance that the percent of LinkedIn users in the sample will be at least 25% but less than 50%?

- ○ 200
- ● 400

(d) (3 pt) Briefly explain your choices in Parts (a)-(c).

The Law of Averages (Chapter 10.1.4) says that the larger the sample size, the closer the proportion in the sample will be to 28%.

5. (10 points) A/B Test on Turtles

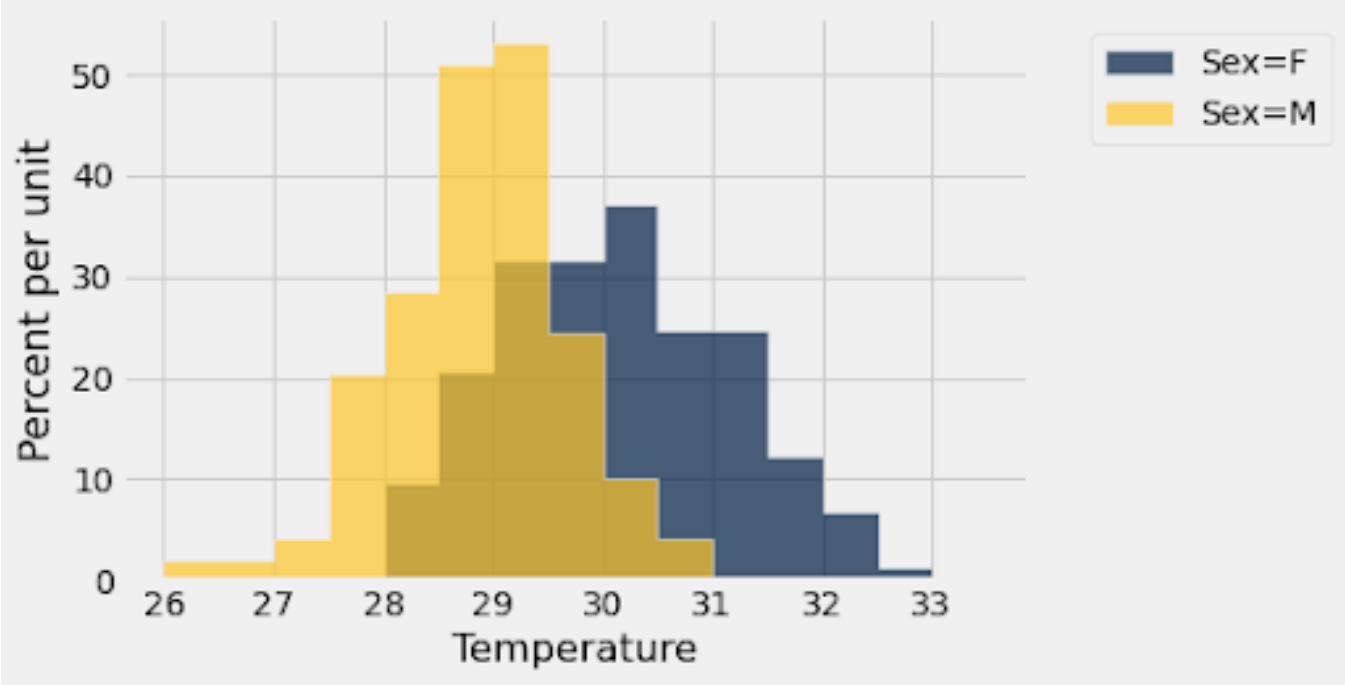
When hatching a baby turtle from an egg, we incubate the egg at some temperature. Ellen read that the temperature an egg is incubated at influences whether or not the turtle that hatches will be male or female.

Ellen loves turtles and is wondering whether this is really right, or whether differences might just be due to chance. She collects data on 100 randomly drawn turtles. She records the incubation temperature (in Celsius) and the sex of the turtle that hatches in the table `turtles`:

Temperature	Sex
30.8	M
31.5	F
32.4	F

(. . . 97 more rows)

(a) (6 pt) Ellen decides to visualize her data before doing any inference. She creates the following histograms, using the same bins for female and male turtles. All bars of the histograms are clearly visible.



Histogram of incubation temperatures

Which of the following are conclusions that can be drawn from the histogram? Select all that apply.

☐ 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.

■ In this sample, the proportion of male turtles with incubation temperatures between 29.5 and 30 degrees is the same as the proportion of female turtles incubated between 30.5 and 31 degrees.

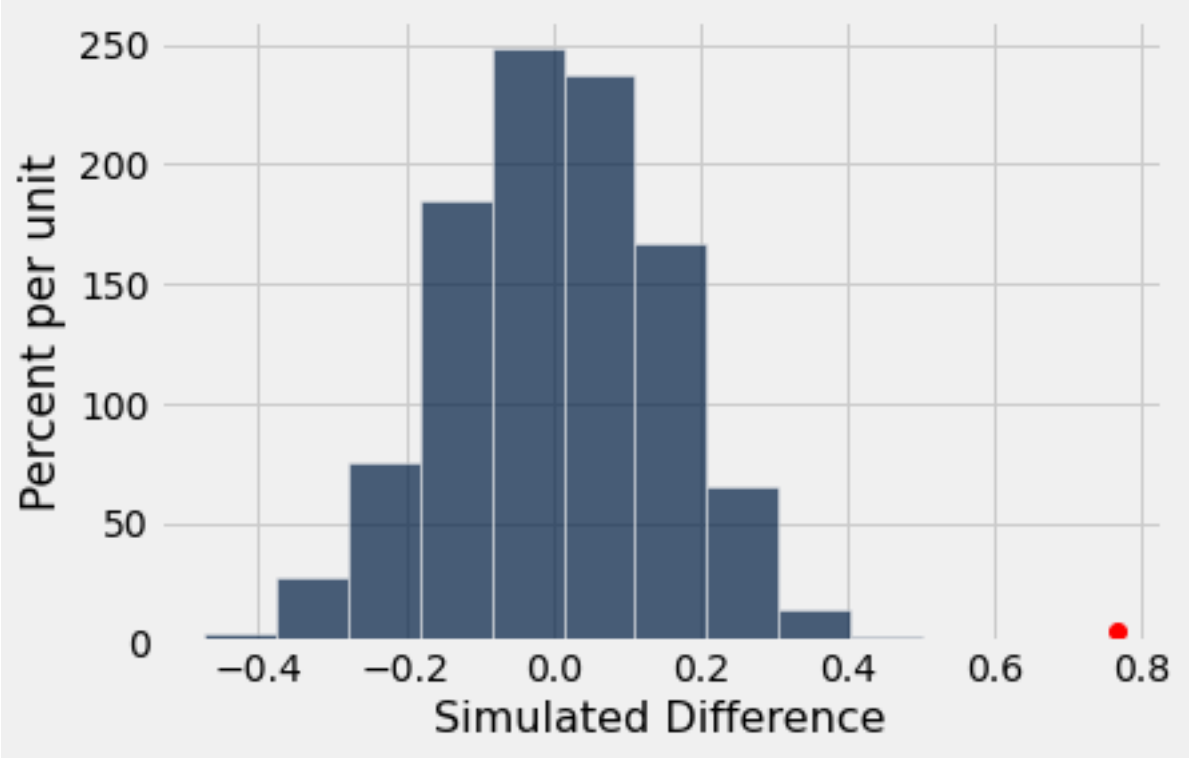
■ There was not a single male turtle in this sample incubated at a temperature above 31 degrees.

■ For at least half the male turtles in the sample, the incubation temperature was below 29.5 degrees.

■ In this sample, males and female turtles have different distributions of incubation temperatures.

☐ None of the above

(b) (4 pt) Ellen performs an A/B test to see whether females in the population in general have higher incubation temperatures than the males, or if the observed difference in distributions is due to chance. Ellen's test statistic is the difference between average incubation temperatures, defined as "female average minus male average". She simulates the statistic 1000 times under the null hypothesis. The histogram below shows the 1000 simulated differences. The red dot shows the observed difference.



Results of simulating the test statistic

Which of the following statements is justified based on this visualization?

- ◯ Based on the test, a reasonable conclusion is that the difference observed in the sample is due to chance.
- ● Based on the test, a reasonable conclusion is that the average incubation temperature of females in the population is higher than the average for males in the population.
- ◯ Based on the test, Ellen cannot reasonably decide between her two hypotheses.

6. (12 points) Testing Hypotheses

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 reported being online almost constantly.

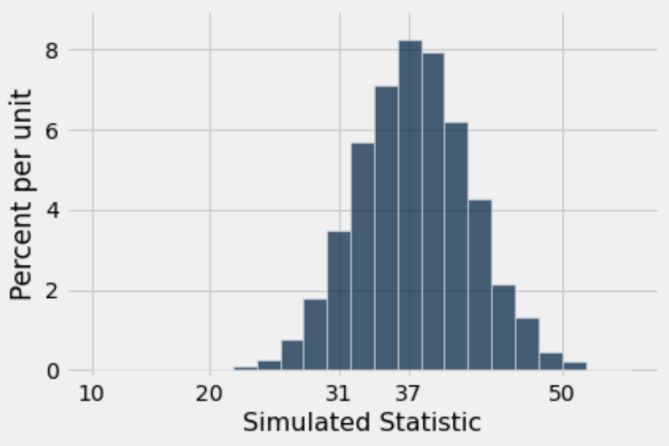
One member of the team says, "The percent of San Francisco adults who are online almost constantly is more than in the nation."

Another member of the team says, "No, it's just chance."

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

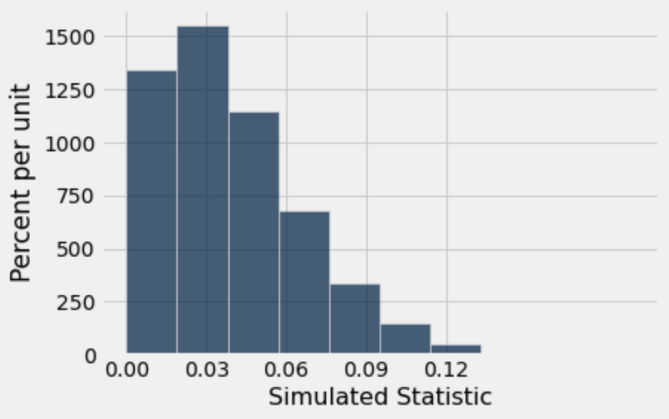
(a) (4 pt) State a clear and complete null hypothesis.

The sample of adults from San Francisco is like draws at random with replacement from individuals of whom 31% are labeled "online almost constantly."



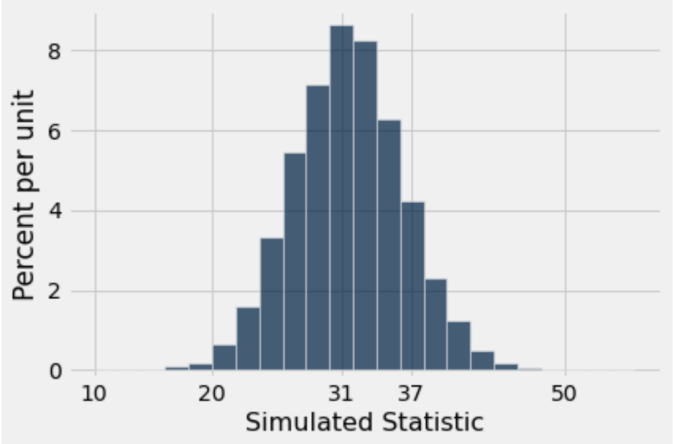
Testing Option C

(b) (3 pt) 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. One of the graphs below is the histogram of their simulated values. Which one is it, and why? [Note that in each graph, some relevant values are labeled on the horizontal axis.]



Testing Option A

○



Testing Option B

●

○

(c) **(2 pt)** Explain your choice above. One or two sentences should suffice.

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 pt)** The 10,000 simulated values of the data scientists' test statistic are in an array called `SIM_STAT_ARR`. Write an expression that evaluates to the p-value of the test.

```
np.count_nonzero(SIM_STAT_ARR >= 37) / 10000
```

7. (8 points) A/B Testing on News

Each person in a random sample of 1000 U.S. adults was asked if they agreed with the statement, "News organizations are growing in influence." Among the sampled men, 39% agreed. Among the sampled women, 43% agreed.

Data scientists have used an A/B test to see whether or not the observed difference is due to chance.

(a) (3 pt) The null hypothesis is one of the statements below. Pick the right one.

- ○ In the sample, the percent of women who agree is the same as the percent of men who agree. The observed difference is due to chance.
- ○ In the U.S., 39% of the men agree and 43% of the women agree, due to chance.
- ● In the U.S., the percent of men who agree is the same as the percent of women who agree. The difference in the sample is due to chance.
- ○ In the U.S., the percent of women who agree is different from the percent of men who agree, due to chance.

(b) (5 pt) The data scientists are using a 1% cutoff for the p-value of the test. They run the test and the p-value comes out to be 0.5%, that is, 1 in 200.

Select **all** of the true statements below. Only one may be true, or more. Make sure you select all that are true.

- ☐ The data scientists will conclude that the data are consistent with the null hypothesis.
- ☐ There is only a 1 in 200 chance that the null hypothesis is true.
- ☐ There is a 199 in 200 chance that the alternative hypothesis is true.
- ■ The data scientists will reject the null hypothesis.
- ■ The assumptions made in the null hypothesis are used in the calculation of the p-value.
- ☐ None of the above statements is true.

8. (14 points) Simulation

The table `WELCOME_TBL` contains the results of this semester's Data 8 welcome survey. The first two rows are shown below. Each row corresponds to a student. In the column `Extraversion`, each student scored themselves on a scale of 1 (not extraverted) to 10 (extremely extraverted).

Year	Extraversion	Number of Textees	Hours of Sleep	Handedness	First Pant Leg	Sleep Position
Second	8	5	6	Right-handed	Right	Left
Second	7	8	7.5	Right-handed	Right	Left

(. . . 1000 rows omitted)

(a) (4 pt) Complete the code below to define a function `FUN_NAME` that takes a sample size as its argument. The function should sample that many times at random **without** replacement from all the students and return the maximum extraversion score of the sampled students.

```
def FUN_NAME(...):
    ...
    ...
```

```
def FUN_NAME(sample_size):
    samp = WELCOME_TBL.sample(sample_size, with_replacement=False)
    return max(samp.column("Extraversion"))
```

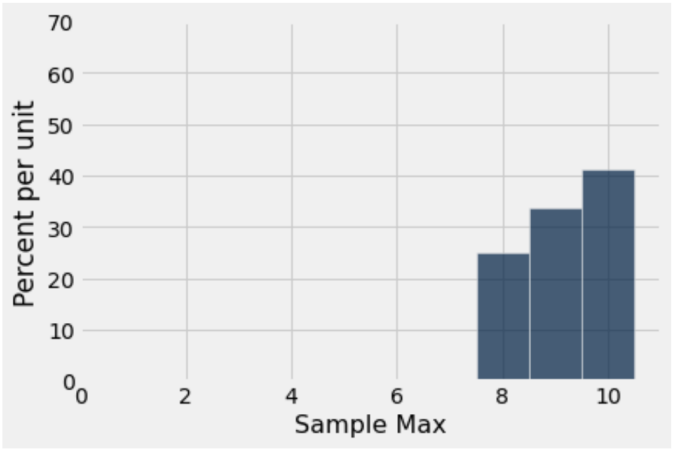
(b) (5 pt) Complete the code below so that the last line evaluates to an array of 10,000 simulated values of the maximum extraversion score in a random sample of size 25 drawn without replacement from all the students. Your code should use the function `FUN_NAME` that you defined above.

```
repetitions = ...
SIM_VALS = ...

for ... in ...:
    ...

SIM_VALS
```

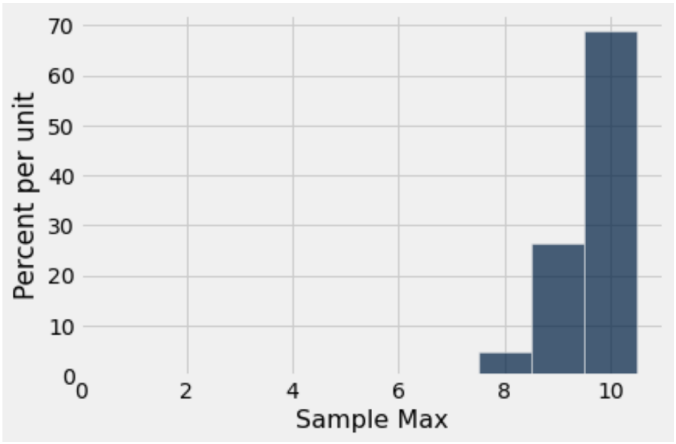
```
repetitions = 10000
SIM_VALS = make_array()

for i in np.arange(repetitions):
    one_sim_max = FUN_NAME(25)
    SIM_VALS = np.append(SIM_VALS, one_sim_max)

SIM_VALS
```

(c) **(3 pt)** A student mistypes the sample size in the previous question to be 55 instead of 25. One of the histograms below shows the distribution of the maximum values simulated by this student. The other shows the distribution of the maximum values that you simulated using a sample size of 25. Which is which?

A:



B:



- (●) A is sample of 25, B is sample of 55
- ( ) A is sample of 55, B is sample of 25

(d) **(2 pt)** Explain your answer above.

The probability of containing 10 (the population maximum) in a given sample is larger for a larger sample size.

9. (8 points) Interpreting Visualizations

A medical institute that specializes in sports medicine has recorded data on athletes with leg injuries. The variables are the distance that the athlete achieved in a test called the triple hop, and how high the athlete could jump vertically. Both distances were measured in centimeters.

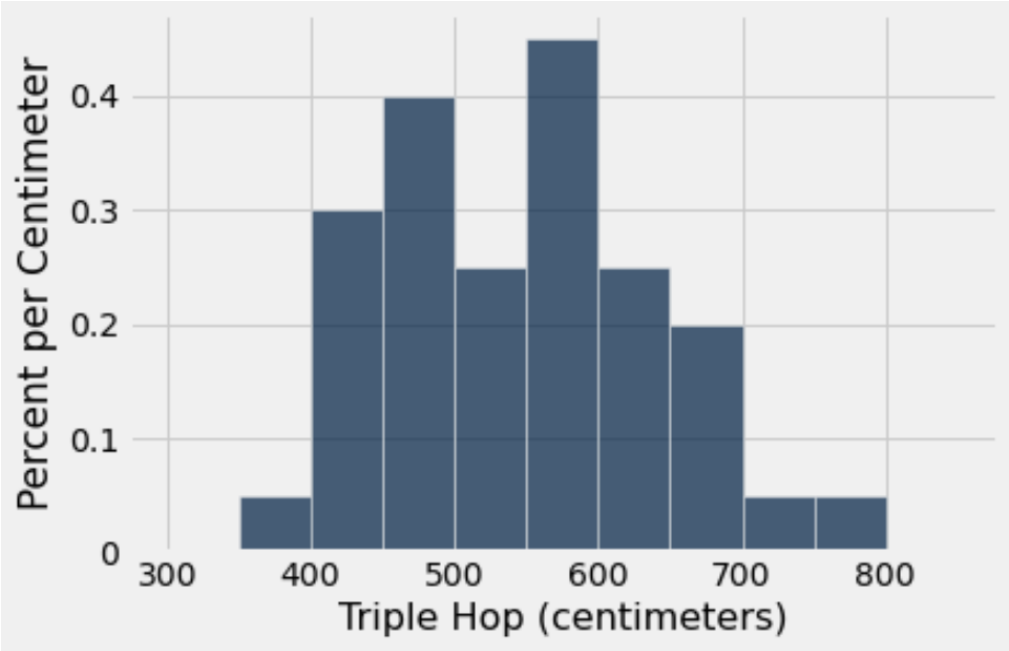
The data are in a table called `jump` that has columns labeled `Triple Hop` and `Vertical`.

Triple Hop	Vertical
443	59
481	62

(. . . 86 more rows)

(a) (3 pt) The histogram below shows the distribution of the triple hop distances, drawn using the following code.

```
jump.hist('Triple Hop', bins=np.arange(300, 900, 50))
```

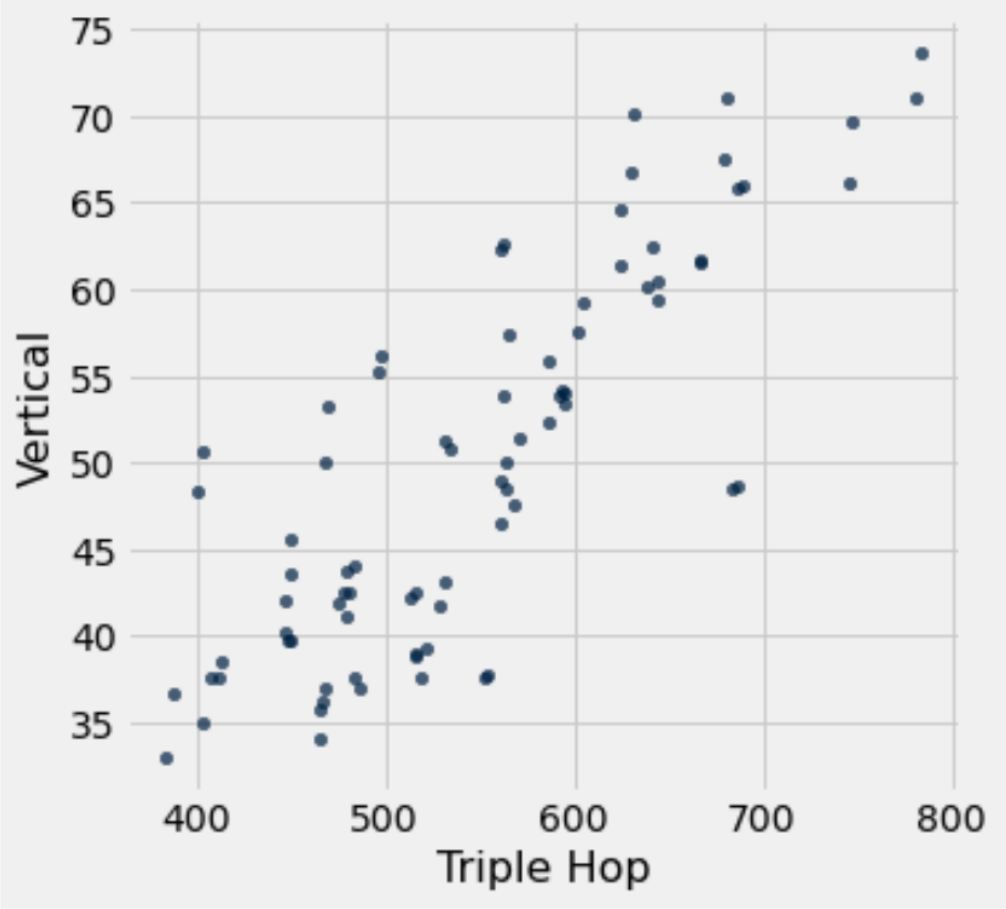


Histogram of triple hop distances

Complete the sentence with the correct option.

The percent of athletes whose triple hop distances were at least 400 centimeters but less than 500 centimeters is equal to

- ○ 0.7%
- ○ 7%
- ○ 30%
- ● 35%
- ○ 40%
- ○ some value that is none of the above or cannot be computed based on the information given



Scatter plot of athlete data

(b) (5 pt) The scatter plot below has a point for each of the athletes. Pick **all** the conclusions that can be drawn from the scatter plot. Make sure you pick all that apply.

- [x] More than half the athletes jumped less than 60 centimeters vertically.
- [x] Most of the athletes whose triple hop distances were longer than average also jumped higher than average.
- [ ] If athletes were to increase their triple hop distances then they would be able to jump higher.
- [ ] If athletes were to increase the heights of their vertical jumps, they would be able to triple hop longer distances.
- [ ] None of the above conclusions can be drawn from the scatter plot.

10. (0 points) Final Words

(a) (0 pt) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question 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.

No more questions.