1 Question 01

You decide to study a recent outbreak of **Flu-X**, a super-flu plaguing Lehigh University students. The Flu-X virus is quite infectious, and when it infects its victims they exhibit a curious symptom—they uncontrollably study statistics!

The first step in your study is to estimate the proportion of students infected with **Flu-X**.

(A)

What is the **population** under study?

To estimate the proportion of students infected with Flu-X, now uncontrollably studying statistics, you decide to select 100 names from the student registry at Lehigh University. You plan to observe each student and record: their major, their expected year of graduation, whether they are uncontrollably studying statistics.

(B)

What is another name for the 100 students you've selected to study?

(C)

Is randomly sampling 100 students **representative** of the population? Why or why not?

(D)

What if instead of randomly selecting and observing the 100 sampled students, you decided to send a University-wide email asking students to reply with their major, expected year if graduation, and whether they cannot stop studying statistics. Could asking students to reply to your email introduce any biases?

(E)

What are the **observations** in the above project?

(F)

Please name one **variable** that is collected

2 Question 02

Can the number of observations in a sample be larger than the number of observations in the population of interest? Why or why not?

3 Question 03

After sampling 100 students you find more than 80% are infected with Flu-X. Students all around you on campus are uncontrollably studying statistics! There is a shortage of calculators, students are babbling about statistical distributions and computing averages. Only you can stop this outbreak!

(A)

If you sampled 100 students by choosing friends you've made at LU and students who attend the same classes you do, what would this sample be called?

(B)

Other researchers may call the sample of friends and students in the same class as you a biased sample. What do they mean when they say that?

(C)

Describe, in your own words, how you might take a simple random sample of 100 students on campus.

(D)

Suppose you wanted to collect an equal number of freshman, sophomores, juniors, and seniors. To do this, you divided students by whether they're freshman, sophomores, juniors, and seniors, and then within each group (or strata) selected 25 students at random. What would the above sampling strategy be called?

Question 04

Great work so far on solving the Flu-X outbreak! But now, it is time to experiment.

After intense laboratory experiments it appears a cure may be at hand—pizza. To confirm your scientific breakthrough you decide to run an experimental study that enrolls students infected with Flu-X. With a probability of 1/2 you assign students to eat pizza versus not eat pizza.

(A)

What is it called when you assign observations to different groups (eating versus not eating pizza)?

(B)

Are there any potential advantages to enrolling and assigning students infected with Flu-X to eat or not pizza versus an observational study that records student symptoms at a local pizzeria?

Question 05

Friends of yours, other biostatisticians not yet infected with Flu-X, and yourself meet to discuss data that was collected on campus. When everyone pools together their data you find we have several variables.

Please classify the following variables as numerical continuous, numerical discrete, categorical ordinal, and categorical nominal.

1. Student's age
2. Whether they have do or do not have Flu-X symptoms
3. Study's body temperature (elevated temperature is a sign of Flu-X)
4. Whether the student is a freshman, sophomore, junior, or senior.

4 Question 06

Suppose we collected data about 100 students, entered the data into a spreadhseet, and wanted to analyze the data using Python. To do this, we will import a package—or set of functions—called Pandas (Python Data Analysis Library) and use a function from Pandas to read in our dataset from a file called ”_100StudentsData.csv”.

```
import pandas as pd # This is a package, or set of function, used to read, write
                                  , and manipulate datasets.
ourData = pd.read_csv("./\100StudentsData.csv")
```

And we print our datset

```
print(ourData)
```

and see the following

Student	Major	Year of Graduation	Flu-X
1	Poph	2025	Yes
2	Engineering	2025	No
3	Literature	2024	No
4	Economics	2023	Yes
5	Neuroscience	2022	No

Table 1: The first 5 out of 100 rows of our dataset.

We can extract individual columns from our dataset in python into **lists**—an ordered array of values. Let's extract the “Major” column as an example

```
studentMajors = ourData.Major # you can extract indivudal columns by writing <
                                  dataset.column>
```

and store it in our computer's memory.

A statistical name for a column from our dataset, like the column Major, is?