Quiz 3 - DSC 10, Winter 2025

Full Name:

PID:

Quiz Time: ◯ 3PM ◯ 3:30PM ◯ 4PM ◯ 4:30PM

Instructions:

- This quiz consists of 3 questions. You have a total of 20 minutes to complete it.
- Please write **clearly** in the provided answer boxes; we will not grade work that appears elsewhere. Completely fill in bubbles and square boxes; if we cannot tell which option(s) you selected, you may lose points.

  ◯ A bubble means that you should only **select one choice**.

  ☐ A square box means you should **select all that apply**.
- If your answer is a string, make sure to put it in quotes. If your answer is a float, make sure to include a decimal point.
- You may use one handwritten sheet of notes. No calculators and no computers.

By signing below, you are agreeing that you will behave honestly and fairly during and after this quiz.

Signature:

Version A

Please do not open your quiz until instructed to do so.

Question 1

A __(a)__ provides insight into how much a __(b)__ might vary across samples and is key for estimating the corresponding __(c)__. This type of distribution is an example of a __(d)__ because it is based on trials of an experiment, as opposed to a __(e)__, which is based on theory.

The five terms that should fill in the blanks are below. Determine where each term belongs. Each term goes in exactly one blank.

parameter	◯ (a)	◯ (b)	◯ (c)	◯ (d)	◯ (e)
statistic	◯ (a)	◯ (b)	◯ (c)	◯ (d)	◯ (e)
probability distribution	◯ (a)	◯ (b)	◯ (c)	◯ (d)	◯ (e)
empirical distribution	◯ (a)	◯ (b)	◯ (c)	◯ (d)	◯ (e)
bootstrapped distribution	◯ (a)	◯ (b)	◯ (c)	◯ (d)	◯ (e)

Question 2

The DataFrame `geisel` contains a row for each day of this academic year. The `"Students"` column contains the number of students who visited Geisel Library on that day, as an `int`.

a) Fill in the blanks below such that

- `geisel_sample` is a **simple random sample** of 50 rows of `geisel`, and
- the code prints the endpoints of a 95% bootstrapped confidence interval for the **mean** number of students at Geisel Library, based on the data in `geisel_sample`.

```
geisel_sample = __(a)___
y = np.array([]):
for i in np.arange(10000):
    x = __(b)__
    y = np.append(y, x)
print(np.percentile(y, 2.5), np.percentile(y, 97.5))
```

(a):

(b):

b) Which of the following is the best description of `x` and `y` in the code above?

- ◯ `x` represents the original sample, and `y` represents many resamples.
- ◯ `x` represents the average student count for a single day, and `y` represents the student count on all days.
- ◯ `x` represents the sample statistic from a single resample, and `y` represents the distribution of those statistics across multiple resamples.
- ◯ `x` represents the sample statistic for the original sample, and `y` represents a distribution of statistics across multiple resamples.
- ◯ `x` represents the population parameter, and `y` represents a bootstrapped distribution of sample statistics.

c) Select all true statements below.

- ☐ If `geisel_sample` had instead had 60 rows, the resulting 95% confidence interval would have been wider.
- ☐ If we made 100 95% confidence intervals based on `geisel_sample`, about 95 of them would contain the population mean.
- ☐ On about 95% of days, the number of students at Geisel Library falls between the endpoints of our confidence interval.
- ☐ It would have also been appropriate to generate a confidence interval for this parameter using the Central Limit Theorem.
- ☐ The standard deviation of `geisel_sample` should be approximately the same as the standard deviation of `geisel`.
- ☐ The data in `geisel_sample` is roughly normally distributed.

Question 3

Your friend at SDSU records the number of students who visit their library each day, for 100 days. They tell you that the average is 6,000 and that the standard deviation is 500.

a) Without knowing anything about the distribution of your friend's data, find the endpoints of the smallest interval which is guaranteed to contain at least 75% of your friend's data. Both endpoints should be given as **integers**.

$$\left[\quad \square \quad , \quad \square \quad \right]$$

b) If you then learn that your friend's data is normally distributed, approximately what percentage of the data is actually contained in the interval you found above? Give your answer as an **integer**.

$$\square \ \%$$

Reminders:

- **Write your PID** on the front page and on the top right corner of page 3.
- If your answer is a string, make sure to put it in **quotes**.
- If your answer is a float, make sure to include a **decimal point**.
- Stay in your seat until the quiz is over. No leaving early.