Quiz 3 - DSC 10, Fall 2024

Full Name:

PID:

Quiz Time: ◯ 4PM ◯ 4:30PM ◯ 5PM ◯ 5:30PM ◯ 7PM

Instructions:

- This quiz consists of 2 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.
- No aids are allowed (no 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.

Welcome, brave space explorers, to the alien planet of Statistica (aptly named for the statistical techniques you will use in order to better understand the planet's animal populations).

Question 1

The DataFrame `space_reptiles` contains 1000 rows of information about **all** space reptiles living on Statistica, which we'll think of as a population. For each reptile, we have its `"length"` in meters, `"age"` in years, and `"number_of_eyes"`. The first five rows of `space_reptiles` are shown on the right.

	length	age	number_of_eyes
0	27.5	250	2
1	23.8	120	4
2	29.1	275	6
3	26.3	300	8
4	21.7	190	10

a) Fill in the blanks in the `sample_of_reptiles` function. The function has two parameters, `"sample_size" (int)`, which will be a positive integer, and `"column" (str)`, which will be the name of one of the columns in `space_reptiles`. The function should take a sample of reptiles from `space_reptiles`, **with replacement**, of the specified size, and return the average value in the given column for the sample.

```
def sample_of_reptiles(sample_size, column):
    return space_reptiles.sample(__(x)__).__(y)__
```

x:

y:

b) True or False: The function call `sample_of_reptiles(1000, "length")` is an example of **bootstrapping**.

◯ True ◯ False

c) Calculate the **variance** of the data in the first five rows of the `"number_of_eyes"` column of `space_reptiles`: 2, 4, 6, 8, 10. Give your answer as an **integer**.

variance =

d) Suppose the next row in the `"number_of_eyes"` column contains 6. If we add this value to our dataset and then recompute the variance, it would...

◯ decrease because the new value is less than the greatest value in the data set.
◯ decrease because the new value is equal to the mean.
◯ remain the same because the new value is equal to the median.
◯ increase because the data set has more values than it did before.
◯ increase because the new value is a positive number.

PID: _____

Question 2

Statistica's forests are filled with tall creatures called whingdingdillies. You have a large random sample of 400 whingdingdillies. In this sample, the mean height is 30m and the standard deviation is 4m. Suppose that whingdingdilly heights are normally distributed.

a) What are the endpoints of a CLT-based 95% confidence interval for the mean height of whingdingdillies? Each value should be a single number.

left endpoint = ☐ right endpoint = ☐

b) Determine the values of the variables `v` and `w` in the code below so that `wdd_prop` evaluates to the approximate proportion of whingdingdillies with heights between 30m and 33m. Each value should be a single number.

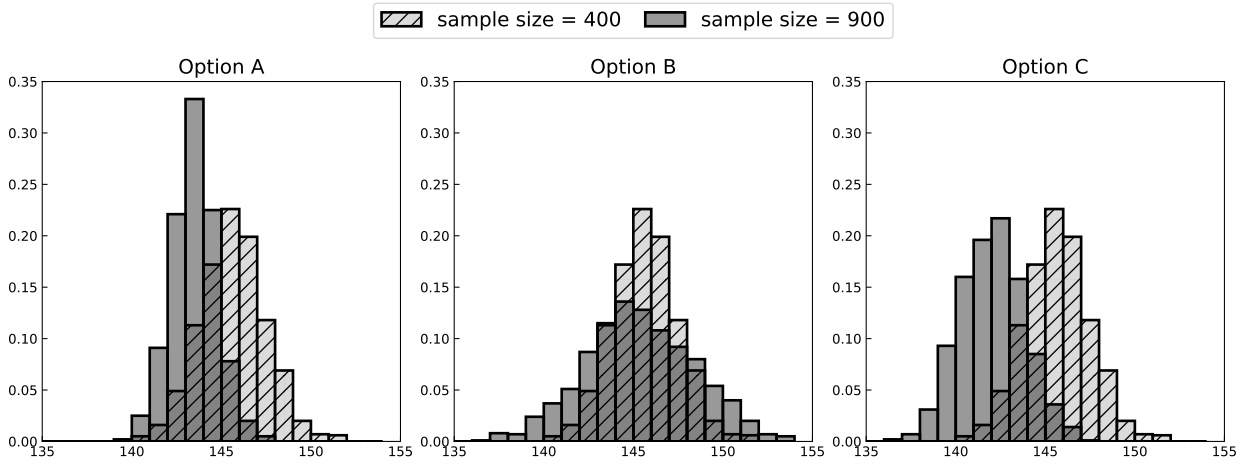
```
wdd_prop = stats.norm.cdf(v) - stats.norm.cdf(w)
```

`v =` ☐ `w =` ☐

c) Above, we stated an assumption that whingdingdilly heights are normally distributed. For which part(s) of this question did we **need** that assumption?

◯ (a) only ◯ (b) only ◯ both (a) and (b) ◯ neither (a) nor (b)

d) After a frightening encounter, you discover that whingdingdillies can run very fast. You collect a sample of 400 whingdingdilly speeds, then use this sample to generate a bootstrapped distribution of resample mean speeds. Afterwards, you wonder how your bootstrapped distribution would have looked if you had instead been able to collect a random sample of size 900. Which of the following overlaid histograms shows two bootstrapped distributions of resample mean speeds, based on samples of size 400 and 900?



◯ Option A ◯ Option B ◯ Option C

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.