ORCA 4500: Foundations of Data Science Fall 2019

Assignment 5

Deadline: 10/18/2019 11:59pm

Instruction: Please hand in two files on courseworks.

1. A pdf document for Exercise 1 through 4, named as "Assignment5_UNI.pdf".
2. Jupyter notebook file for Exercise 5, named as "Assignment5_UNI.ipynb".

Exercise 1
The Registrar's Office at a school has the school's complete enrollment information. For every student, it has the complete list of classes in which the student is enrolled, and for every class, it has the complete list of enrolled students. There are 300 classes being offered at the school. The Registrar selects a probability sample of students as follows: she selects a sample of 10 of the 300 classes at random without replacement, and then for each selected class, she adds all the students in that class to the probability sample. (So if a student is in more than one of the selected classes, that student appears in the probability sample more than one time.) This method is called **cluster sampling** because each selected class is a cluster of students.

(a) Do all students have the same chance of entering the sample? Explain your answer.

(b) Describe three significant differences between the properties of the Registrar's sample and properties of a sample drawn at random without replacement from among all the students.

Exercise 2
Given the following information, what is the probability that a woman over 50 has cancer if she has a positive mammogram result?

(a) One percent of women over 50 have breast cancer.

(b) Ninety percent of women who have breast cancer test positive on mammograms.

(c) Eight percent of women who do not have breast cancer test positive on mammograms.

Exercise 3
Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50.

(a) What can be said about the probability that this week's production will exceed 75?

(b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

Exercise 4

If 100 fair dice are rolled, we want to calculate the approximate probability that the sum obtained is between 325 and 375 using the **central limit theorem.** In particular, let X_i denote the value of the i-th die for $i = 1, 2, \cdots, 100$, and let $\bar{X}$ be the average of the 10 dice.

1. What is $\mathbb{E}(X_i)$ and what is $\mathbb{E}(\bar{X})$?
2. What is the variance $Var(X_i)$ and what is the variance $Var(\bar{X})$? [Hint: Recall the identity for variance: $Var(X) = E(X^2) - E^2(X)$]
3. What distribution does $\bar{X}$ follow approximately due to the CLT?
4. What is the probability $\mathbb{P}(3.25 \leq \bar{X} \leq 3.75)$? You can use the online calculator here: `http://onlinestatbook.com/2/calculators/normal_dist.html`.

Exercise 5

When flipping coins, a streak of length k is a series of k consecutive flips that all come up with the same value. For example if your flips are **HHTTHTTTHTHT**, there is one streak of length 3. The longest streak in a series of flips is the streak of greatest length. So in the sequence above, the longest streak is of length 3. As another example, in the seqence **THHHHTHHHTHHT**, the longest streak is of length 4. We want to know the expected length of the longest streak when flipping n coins, for $n = 10, 100, 1000, 10000, 100000$. Perform some simulations and answer this question as best you can.

Show the answer in the submitted jupyter notebook, since it might take a while for us to run the simulation. You can either do so by keeping the output or type your answers in a "markdown" environment.