DATA 252 / DATA 551: Homework 7

- This homework is due by Monday 30, 2020 at the beginning of class. **You need to submit your answers on Moodle in a pdf document**. In addition, there will be a short quiz at the beginning of class on Monday 30, which might contain contents from this homework, in addition to contents from the most recent lecture.

1. In class, we used bootstrapping to study the *median* incubation period. Recall that the median is the $50th$ percentile. We might be interested in estimating other percentiles, like the $90th$ percentile, which tells us that 90% of patients will develop symptoms before this time. Use the same dataset (`ncov_simple`) and answer the following questions.

   (1) What is the $90th$ percentile of the incubation period among patients in the dataset? Recall, in R, you can use the `quantile` function.

   (2) Obtain $B = 5000$ bootstrap samples; record the $90th$ percentile of each bootstrap sample. What is the SD of the $90th$ percentile based on your bootstrap samples?

   (3) Compared to the median (we have estimated its SD in class), does the $90th$ percentile have more or less variation?

   (4) Obtain a 95% confidence interval for the $90th$ percentile of the incubation period.

   (5) Provide your code here (either by copy-pasting or using a screenshot).

2. Suppose we want to use bootstrapping to study the *maximum* incubation period.

   (1) What is the maximum incubation period among patients in the dataset?

   (2) Obtain $B = 5000$ bootstrap samples and construct a 95% confidence interval for the maximum incubation period.

   (3) Briefly explain why bootstrapping does not work well for estimating the maximum incubation period.

   (4) Provide your code here (either by copy-pasting or using a screenshot).

The following is modified from the last question on Exam 1 (some numbers have been changed). Re-do this question. The grade (out of 10 points) you receive here will be added to your exam grade. If your current exam grade is 95 and above (good job!), you can skip this part, and will receive 5 points extra credit automatically. This part will not be counted in the homework grade. **You must work on this question independently, without communicating with any other student.**

Suppose the incubation period, T, of a certain disease can be modeled by an exponential distribution with a mean of 4.5 days.

(1) What is the chance that a given person develops symptoms within 5 days of contracting the disease (i.e., $P(T \leq 5)$)?

(2) In real life, we usually cannot observe the actual incubation periods T for each person, but can only observe whether or not a person develops any symptoms in, say, 5 days. Simulate 100 patients who have contracted the disease, with a value of 1 representing that the patient has symptoms within 5 days, and a value of 0 representing that the patient has no symptom within 5 days (i.e., you'll be generating a vector of length 100 with entries 0 or 1). Copy-paste this vector here.

(3) Based on your answer in (2), how many patients, out of the 100 patients, have developed symptoms within 5 days?

(4) Run a simulation to estimate the probability that, among 100 people who have contracted the disease, 80 or more people will develop symptoms within 5 days.

(5) Suppose the number of new cases contracting the disease today follows a Poisson distribution with a mean of 100 people. Modify your simulation in (4) to estimate the probability that, among people who have contracted the disease today, 80 or more people will show symptoms within 5 days.

(6) Generate a histogram of the distribution of the number of people that will show symptoms within 5 days, under the setup in (5). Take a screenshot of the histogram to include here.

(7) Provide your code here (either by copy-pasting or using a screenshot).