Data 100, Summer 2022

Homework #6 Written Question

Total Points: 12

Submission Instructions (Written Only)

Question 1 in this document is a written problem and should be submitted as a separate PDF to the Written portal of Gradescope. You must submit this assignment to Gradescope by **Thursday, July 28th at 11:59 PM Pacific**. While Gradescope accepts late submissions, you will not receive **any** credit for a late submission if you do not have prior accommodations (e.g. DSP).

You can work on this assignment in any way you like: (1) Download this PDF, print it out, and write directly on these pages (we've provided enough space for you to do so); (2) If you have a tablet, you could save this PDF and write directly on it; (3) Use some form of LaTeX. Overleaf is a great tool; or (4) Write your answers on a blank sheet of paper.

Regardless of what method you choose, the end result needs to end up on Gradescope, as a PDF. If you wrote something on physical paper (like options 1 and 3 above), you will need to use a scanning application (e.g. CamScanner) in order to submit your work.

When submitting on Gradescope, you **must correctly assign pages to each question** (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our tutors. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. If you have any questions about the submission process, please don't hesitate to ask on EdStem.

Collaborators and Content Warning

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others please include their names at the top of your coding submission.

This assignment includes an analysis of daily COVID-19 cases by U.S. county through 2021. If you feel uncomfortable with this topic, please contact your GSI or the instructors.

1 Random Variables: COVID-19

In this homework, we analyze COVID-19 case counts over time. Before diving into the coding notebook, we will perform some initial analysis using properties of random variables, like expectation and variance.

Consider the COVID-19 infection and recovery cases from a start date of March 1, 2021. Define day i in the range $[0, 365]$, inclusive, to be the number of days since this start date, and define the following:

- C_i, a random variable for the number of people currently sick/infected with COVID-19 on Day i. Let $p_{c,i}$ be the probability of a person being sick/infected with COVID-19 Day i; a value; assume $p_{c,i}$ is known for all i.
- R_i, a random variable for the number of newly recovered cases on Day i. Let $p_{r,i}$ be the probability of recovery for an infected person on Day i; a value; assume $p_{r,i}$ is known for all i.
- n is the population of the world; a value.

Assume that (1) everyone who has recovered from COVID-19 will not be reinfected and/or re-recover, and (2) the population of the world n is a fixed value for all days i. In practice neither of these assumptions hold, but they greatly simplify our computations below.

(a) (2 points) Consider the random variable C_i for a particular day i. Note that C_i is not a Binomial random variable because trials are not independent; two people in close proximity to an infected individual are more likely to get infected together.

However, C_i can still be defined as a sum of n Bernoulli random variables as follows: For each of the n individuals in the world, count 1 if the individual is infected (with probability $p_{c,i}$), and 0 otherwise. Compute the expectation $E[C_i]$.

(b) (2 points) Define "aggregate case load" as the total sum of cases for all days, including repeat values of people that are sick for multiple days. That is, say there are only two people in our simplified example: Person A and Person B. If Person A is sick for Day 1/2 and Person B is sick for Day 2/3/4/5, the "aggregate case load" from Day 1-5 is 6.

Using the above definitions, write an expression to define a new variable C in terms of existing variables, such that C represents the aggregate case load from March 1, 2021 to March 1, 2022.

(c) (2 points) Compute the expectation of C, the variable you defined in part (b), $\mathbb{E}[C]$ in terms of $p_{c,i}$ for all i in the range $[0, 365]$ and the population of the world, n.

(d) (2 points) Suppose we derive the variance of C as follows:

$$\text{Var}(C) = \sum_{i=0}^{365} \text{Var}(C_i)$$

What is a potential issue with this approach?

(e) (2 points) Assume that for someone to be considered recovered on Day i (i.e. part of R_i), they must test negative twice on that day. We can define T_i, the number of tests taken by recovered people on Day i as $T_i = 2R_i$.

Which of the following are true? Justify your answer.

- ◯ A. We can compute the variance as $\text{Var}(T_i) = 2\text{Var}(R_i)$ **only if** R_i is a Binomial random variable.
- ◯ B. We can compute the variance as $\text{Var}(T_i) = 4\text{Var}(R_i)$ **only if** R_i is a Binomial random variable.
- ◯ C. We can compute the variance as $\text{Var}(T_i) = 2\text{Var}(R_i)$ regardless of what kind of random variable R_i is.
- ◯ D. We can compute the variance as $\text{Var}(T_i) = 4\text{Var}(R_i)$ regardless of what kind of random variable R_i is.

(f) (2 points) For a particular Day i, assume that the number of recovered individuals R_i is a Binomial random variable, where n is the population of the world and $p_{r,i}$ is the probability of an individual recovering on Day i. If $p_{r,i} = 0.5$, calculate $\text{Var}(T_i)$ in terms of n.