CSCI 2244 – Homework 3

Out: Friday, September 13, 2019
Due: Friday, September 20, 2019, 11:59pm

This homework consists of written exercises and coding problems. You *must* type your solutions. See the "Assignments" section in the syllabus for advice about doing this. You should submit your homework via Canvas. In particular, you should upload a zip file called:

`FirstName_LastName_Homework3.zip`

Please use your full first name and last name, as they appear in official university records. The reason for doing so is that the TAs and I must match up these names with the entries in the gradebook.

This zip file should contain 3 files:

- `written.pdf` – containing your answers to all the tasks in section 1, and the results of running your code as requested by the task in section 2.
- `bloodtest.py` – containing the code you wrote for section 2. You should start with the starter code file by this name on the course website.

1 Written Exercises

As we will see next week, sometimes it is possible for the expected value of a random variable to be undefined or not exist. The issue is that the definition of expected value involves an infinite series, which might diverge. However, for all of the problems below, you can assume that all of the mentioned expected values and variances exist.

Task 1.1 (3 pts). Let X be a random variable, and let a be a constant real number. Show that $\mathsf{Var}[a + X] = \mathsf{Var}[X]$

Task 1.2 (3 pts). Let S be a sample space, and let $X : S \to \mathbb{R}$ and $Y : S \to \mathbb{R}$ be two random variables on S. Suppose that for all $s \in S$, $X(s) \leq Y(s)$. Show then that $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Task 1.3 (2 pts). Let X and Y be independent random variables. Show that

$$\mathsf{Var}[XY] = \mathbb{E}[X^2]\mathbb{E}[Y^2] - (\mathbb{E}[X]\mathbb{E}[Y])^2$$

You may use the fact that if X and Y are independent, then $f(X)$ and $g(Y)$ are independent, for any functions f and g.

Task 1.4 (2 pts). Let X be a $\mathsf{Binomial}(10, .2)$ random variable and Y be a $\mathsf{Binomial}(5, .7)$ variable. Use Markov's inequality to bound the probability that $X + Y \geq 7$.

Task 1.5 (4 pts). A university wants to anonymously survey students to find out if they drink underage. Students are skeptical about whether the survey is truly anonymous. To reassure them, the university proposes the following: while filling out the survey, each student will first flip a coin that returns heads with probability r:

- If the coin is tails, the student will answer the question truthfully.
- If the coin is heads, the student will just answer "yes" to the question.

The idea is that as long as $r > 0$, the students have some plausible deniability: a student could always say that they answered "yes" because their coin came up heads. Assuming $r < 1$ and that students indeed answer honestly when the coin is tails, the university is still able to estimate the overall rate of underage drinking.

Let n be the number of students surveyed. We assume that each student drinks underage with probability p, independent of all others. Let Z be the total number of "yes" answers to the survey.

What is the expected value and variance of Z? (Hint: represent Z as a sum/product of random variables and then use what you know about Bernoulli random variables.)

(Ungraded: do you see how the university can get an estimate of p from Z?)

Task 1.6 (5 pts). A public health organization needs to run a blood test on N people to see if they have a certain disease. One way to do this would just be to run N separate tests. Another way is to divide the N people into groups of k people, mix the blood samples for the people in a group, and run the test on the mixed sample. If the test is negative, then none of the k people in the group have the disease. If it is positive, then someone in the group has the disease, so we then run k separate tests for each person to be able to inform them of their results. In that case, the group requires $k + 1$ total tests.[1]

Assume that each person has the disease with independent probability p.

- What is the probability that the mixed sample for a group of k people will test positive?
- Let X be the total number of tests that need to be run for N people when using groups of size k. What is the expected value of X? (You should assume that k divides N with no remainder).
- Your answer to the previous part should involve the expression $(1 - p)^k$ somewhere. When $|pk|$ is very small, you can approximate $(1 - p)^k$ by $(1 - pk)$. Substitute this approximation into the expected value of X, and then show that the approximated value is minimized when $k \approx \frac{1}{\sqrt{p}}$. Ignore the fact that $\frac{1}{\sqrt{p}}$ might not be an integer. (Hint: take the derivative with respect to k).

[1] This is assuming that the test still works when samples are mixed, that the test is perfectly sensitive and there are no false-positives, and so on, which may be unrealistic. This protocol was originally proposed for use in WWII to cheaply screen drafted US soldiers for disease.

2 Coding

Task 2.1 (3 pts). Write a function called `bloodtest(N, k, p, numtrials)` which uses Monte Carlo simulation to estimate the expected value you figured out in task 1.6. That is, you should simulate generating N random patients, assigning them each the disease with probability p, and then calculating how many tests are needed if you split them into groups of k, repeating this `numtrials` times, to estimate the average number of tests. The function should then return this estimated value. Your code only has to work correctly under the assumption that k divides N evenly.

Run the code with $N \in \{1000\}$, $p \in \{.01, .02, .04\}$ and $k \in \{5, 10, 20\}$, using `numtrials` $= 3000$ for each combination. Report your results.