Homework 4: Discrete Random Variables

For each problem, remember you must briefly explain/justify how you obtained your answer, as correct answers without an explanation will not receive full credit. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide.

In general, your goal in an explanation is to write enough that a student from class who has attended lecture, but not read the problem yet, could understand your approach, verify your reasoning, and believe your answer is correct. While we do not usually need to see arithmetic, you must include enough work that in principle one could rederive your answer with only a scientific calculator.

Unless a problem states otherwise, you should leave your answer in terms of factorials, combinations, etc., for instance 26^7 or $26!/7!$ or $26 \cdot \binom{26}{7}$ are all good forms for final answers.

Submission: You must upload a **pdf** of your written solutions to Gradescope under "HW 4 [Written]". Instructions as to how to upload your solutions to gradescope are on the course web page. The use of latex is *highly recommended*. (Note that if you want to hand-write your solutions, you'll need to scan them.

Due Date: This assignment is due Wednesday July 30th
You will submit the written problems as a PDF to gradescope. Please put each numbered problem on its own page of the pdf (this will make selecting pages easier when you submit), and ensure that your pdfs are oriented correctly (e.g. not upside-down or sideways).

Academic Integrity: Please read the full academic integrity policy. If you work with others (and you should!), you must still write up your solution independently and name all of your collaborators in the separate question on gradescope.

1. Dice Try! [Gradescope]

Complete this problem on the gradescope assignment. Please check your answers as it will be autograded so we cannot give partial credit.

2. CDF to PMF [7 points]

Let X be a discrete random variable that takes integer values from 1 to 12 (inclusive), and has the following cumulative distribution function (CDF):

$$F_X(n) = \begin{cases} 0 & \text{if } n < 1 \\ \frac{\lfloor (n+2) \rfloor \cdot \lfloor (n+3) \rfloor}{210} & \text{if } 1 \leq n \leq 12 \\ 1 & \text{if } n > 12 \end{cases}$$

Find the probability mass function (PMF) for X.

3. Explore The Mine [15 points]

On Friday July 30th we will have a guest lecture from Cole Medeiros who worked alongside Eric Barone (creator of Stardew Valley and a UW Tacoma CS graduate) to create the Stardew Valley board game based on the popular video game Stardew Valley. Probability and statics plays a big role in designing games, so in this problem you will use concepts you learned in class to analyze a small (simplified) portion of the game.

A few details about the game:

- The game can be played with 1-4 people, but for this problem we will assume it is being played as single player.
- The game will be played with the standard season deck meaning there are 16 days to play in total.
- Each day you can choose two actions to preform, they can either be two adjacent actions or the same action. One such action is the Explore The Mine action which we will focus on for this problem.

Details about the Explore The Mine action:

- You can assume that a player will choose the Explore The Mine for $\frac{1}{4}$ of their total actions during the game.
- A player will roll 2 fair six-sided dice where each die has two sides for each of the three symbols: heart, Junimo, and Stardrop (see Figure 1). Each figure is equally likely to come up.
- The result of the die roll will tell the player what item they will get from the map card (see Figure 2). You can assume that the map card will not change during the game.
- The results of the die rolls indicate a row and column, but the player is equally likely to choose which die to use for which.
- The player then receives the item at that location.

Figure 1: From left to right we have the heart, Junimo, and Stardrop.

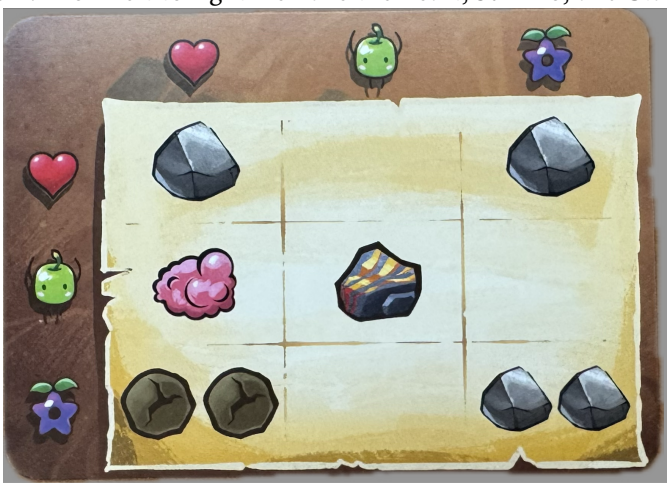

Figure 2: Mine Map Card with rocks, bug meat, ore, and geodes.

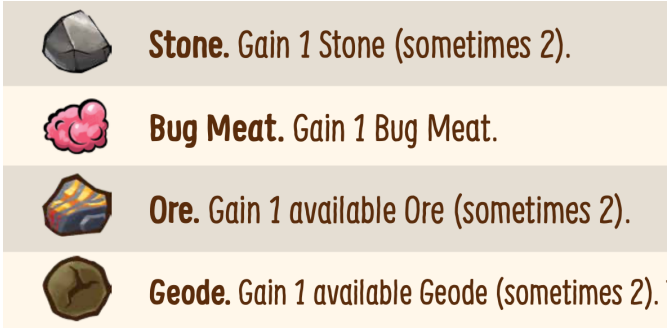

Figure 3: Mine map key.

(a) With the map card in mind (see Figure 2), what is the expected number of stones that a player will receive over the duration of the game from the mine? [8 points]

(b) How many times would a player need to do the Explore The Mine action if they want the average number of stones that they receive from the mine over the duration of the game to be at least 6? [2 points]

(c) Design a Mine Map Card such that:

- The Mine Map Card contains at least one of each stone, bug meat, ore, and geode.
- Stone, ore, and goedes can have up to two of the same type per location on the card.
- The expected number of stones is equal to $\frac{40}{9}$ over the duration of the game.
- The expected number of bug meat is equal to $\frac{16}{9}$ over the duration of the game.
- The expected number of ore is equal to $\frac{8}{9}$ over the duration of the game.
- The expected number of goedes is equal to $\frac{24}{9}$ over the duration of the game.

[5 points]

4. Volleyball [10 points]

A set in volleyball ends when a team has:

- at least 25 points *and*
- at least 2 more points than their opponent.

So, for example, a set at 25-24 is not over (no one has a two point lead), but a set at 27-25 is over.

Suppose a set is tied 24-24, and each point is won by your team (independently) with probability p. What is the expected number of points played before one team or the other wins the set? (Hint: You could calculate this with an infinite sum, in which case you may use wolframalpha to find a closed form. But a clever use of a variable from the zoo will save you quite a bit of work, and make it so an infinite summation is unnecessary.)

5. Instagram [16 points]

A photo-sharing startup offers the following service. A client may upload any number N of photos and the server will compare each of the $\binom{N}{2}$ pairs of photos with their proprietary image matching algorithms to see if there is any person that is in both pictures. Testing shows that the matching algorithm is the slowest part of the service, taking about 100 milliseconds of CPU time per photo pair. Hence, estimating the number of photos uploaded by each client is a key part of sizing their data center. The people in charge say that their gut feeling is that $N = 10$. You (the chief technical officer) say, "but N is a random variable". For each of these possible distributions for N: What will the **expected** time (in milliseconds) for CPU demand per client be? (Your answer may be a function of N, p and/or λ, as appropriate).

Hint: Many of these computations can be done cleverly by using linearity of expectation along with the formula for variance and values from the zoo!

(a) the "distribution" where N is the same fixed number with probability 1?

(b) the Poisson distribution with parameter λ?

(c) the geometric distribution with parameter p?

(d) $N = 80X + 5$, where X is a Bernoulli random variable with parameter p?

In each case, include as part of your answer the expected value of N and the variance of N. Make sure your answer is **not** in the form of a summation for this problem.

6. Real-World: Bayes Theorem [25 points]

The tools of this class are useful to computer scientists, but many of them are useful beyond just "classic" computer science. In this assignment you'll consider an application of Bayes' Rule in the real-world.

We will consider the use of DNA evidence in criminal trials. A full discussion of DNA evidence would require a discussion of many issues[1] – for this assignment, we are going to limit ourselves to just how information about DNA tests should be communicated to juries.

This assignment is a mix of technical tasks (appropriately applying theorems) and non-technical ones (considering tradeoffs between various real-world effects and groups). The technical aspects can be "right" or "wrong", but the non-technical aspects are unlikely to be simply "right" or "wrong" – we won't have to **agree** with the non-technical aspects of your analysis to consider them a good analysis. Our evaluation will be based on how well they connect to the technical aspects, as well as the depth of reasoning demonstrated.[2]

Collaboration Policy: For the work in section 6.2, you are to conduct your own search and analysis for this assignment. While you may get feedback from other students on your writing, you cannot just use the results of another student's search.

6.1. Bayes in Court

DNA evidence has been used in court cases for decades. Over time some common patterns of (dubious) argumentation have emerged, which you'll analyze in this problem.

Consider the following scenario:

A crime is committed somewhere in Seattle. No witnesses were at the crime, but there was blood left at the scene which had DNA extracted from it. The DNA was run against the 13 million DNA samples on file with the FBI. There was one match: a person who lived in Tacoma at the time of the crime.

You know the following facts about the DNA test that was run:

- The false positive rate of the test is $\frac{1}{10{,}000{,}000}$.
- The false negative rate of the test is $\frac{1}{100{,}000{,}000}$.

The prosecutor argues as follows

> The DNA match with the blood on the scene is strong. There is only a $\frac{1}{10{,}000{,}000}$ chance that the defendant is innocent (after all, the test only has a $\frac{1}{10{,}000{,}000}$ rate of failure) – certainly not a reasonable amount of doubt. You must vote to convict.

Let T be the event of a positive test, and G be the event that the defendant is guilty.

(a) In terms of G and T, what probability or conditional probability is the prosecutor describing with their phrase "the chance the defendant is innocent, knowing about the test"? [1 point]

(b) What probability or conditional probability does the $\frac{1}{10{,}000{,}000}$ come from? [1 point]

(c) Describe the prosecutor's error concisely (2-3 sentences). [2 points]

The defense attorney argues as follows:

[1] Among others: under what circumstances DNA samples be taken from people and/or stored in databases.

[2] For example, trying to calculate a probability and getting 1.2 for an answer would involve a technical mistake. Saying "Witnesses shouldn't say the DNA evidence is reliable, because I saw an episode of CSI where it wasn't reliable." is not good reasoning for this assignment because it does not connect to the technical aspects of the problem. Saying "DNA evidence should be allowed as long as the Bayes factor is at least 100" relates to technical aspects and is considered good analysis whether or not we agree with you on "Bayes factor at least 100" being the right place to draw the line between allowable or not.

> The test isn't as good as it sounds. If we ran the test on all 330,000,000 people in the country, we'd have 33 innocent people come up with positive tests. The true probability of my client being guilty is only about $1/34$.

Recreate the Bayes' Rule application that the defense attorney is using

(d) What prior is being used and what is the assumption being made by the defense? Recall the "prior" is the probability of the event you're hoping to analyze *prior to* running the test. Your answer here should include both a number and where it came from. [2 points]
Hint: What is the sample space that the defense is referring to?

(e) Now use Bayes' rule to confirm that (starting from that prior), the calculation is correct. [2 points]

Now choose a new prior. What is **your** estimate of the probability the defendant is guilty? You can use either (or both) of the bullets below. If you use neither bullet, you must incorporate some other information and have something different from what the prosecutor and defense attorneys said. Since this is **your** estimate, there are many possible answers! We aren't grading whether we get the same answer, we're grading whether you have a correct application of reasonable assumptions.

- The 13 million DNA samples in the database are not from a random section of the population, but they do come from people across the whole U.S.
- The Seattle metro area has about 4 million people.

(f) What is your prior (i.e., the probability the defendant was guilty before you ran the test)? Briefly explain where it comes from. [1 point]

(g) Do a Bayes Rule calculation to give your estimate of the guilt of the defendant.[2 points]

(h) Name at least one limitation of your estimate (something you haven't accounted for that you would have liked to, or more information you would have liked about the scenario)? (2-3 sentences) [2 points]

6.2. Make Another Argument

In this part, you'll use an application of Bayes Rule to make an argument about whatever real-world scenario you would like.

Your scenario can be close-to-home (say something about an RSO you're involved in), a political issue, or anything else, as long as it's based in the "real-world"[3].

You are allowed (and encouraged!) to do your own research toward this question, but can also fall back on reasonable estimates.

(a) Define events A and B on which you'll apply Bayes' Rule (along with any other events you need). [2 points]

(b) State probabilities (or probability estimates) for three of the four quantities you need to use Bayes' rule and apply Bayes' rule. [1 point]

(c) For those estimates, either cite a source for the numbers that you think is reliable or give a justification for your estimate. [1 point]

(d) Apply Bayes' rule using the probabilities from part (b) [4 points]

[3] We will be quite lenient about what counts as real world – the hope is that you will pick something you care about. If it's just the probability that the second and third card of a deck of cards are the same value, it's probably not "real-world." But if you're an avid poker player, and you want to use Bayes' Rule to analyze a particular game scenario, that would definitely count.

(e) What is your takeaway from this calculation? This needs to be more than just restating what your calculation in part b found. [2 points]

(f) Discuss at least one limitation of your calculation/application (e.g. factors that didn't go into your estimates, or assumptions you are making that might not be correct). [2 points]

6.2.1. Some Ideas

We hope you'll think of something on your own! If you can't, here are some you might think about:

- Medical tests can lead to false positives/negatives. Some tests you might consider looking into are over-the-counter pregnancy tests, colon cancer tests, and paternity tests.
- How hard should Captchas and other "I'm not a robot" tests be to stop the robots from random guessing, but allow through fallible humans?
- How reliable is the rain prediction in weather apps for Seattle?

7. Feedback + Collaboration [1 point]

Answer these questions on the separate Gradescope box for this question.

Please keep track of how much time you spend on this homework and answer the following questions. This can help us calibrate future assignments and future iterations of the course, and can help you identify which areas are most challenging for you.

- Which students did you collaborate with for this homework?
- Is the work that you are submitting your own and does not violate the academic integrity policy outlined in the syllabus?
- How many hours did you spend working on this assignment (excluding any extra credit questions, if applicable)? Report your estimate to the nearest hour.
- Which problem did you spend the most time on?
- Any other feedback for us?