Homework 4: 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.

Instructions as to how to upload your solutions to gradescope are on the course web page.

Remember that you must tag your written problems on Gradescope.

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 April 24th
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).

Collaboration: Please read the full collaboration policy. If you work with others (and you should!), you must still write up your solution independently and name all of your collaborators somewhere on your assignment.

1. CDF to PMF [7 points]

Let X be a discrete random variable that takes integer values from 1 to 10 (inclusive), and has the following cumulative distribution function:

$$F_X(n) = \begin{cases} 0 & \text{if } n < 1 \\ \frac{\lfloor (n+1) \rfloor \cdot \lfloor (n+2) \rfloor}{132} & \text{if } 1 \leq n \leq 10 \\ 1 & \text{if } n > 10 \end{cases}$$

Find the probability mass function (PMF) for X.

2. Keep It Rolling [10 points]

You roll a die independently 100 times. Let X represent the difference between the number of rolls that turn out to be 4 or 6 (a composite number), and the number of rolls that turn out to be 1, 2, 3, or 5 (a prime number).

(a) What is the support of X? [4 points]

(b) What is the probability that $X = 4$? [6 points]

Note that we are looking at the difference without taking the absolute value. Some examples:
If you rolled a 4, a 2, and a 1, there is 1 composite number and 2 primes. The difference would be $1 - 2 = -1$.
If you rolled a 4, a 6, and a 1, there is 1 prime and 2 composite numbers. The difference would be $2 - 1 = 1$.

3. Turnabout is Fair Play [8 points]

You want to open a casino with a new card game. In the game, a dealer will shuffle a full deck of cards (a standard deck with 4 suits and 13 ranks) so they are fully-shuffled (i.e. every ordering is equally likely). The dealer then flips over the top three cards.

To play, a player pays $\$1$, then gets a payout based on the cards flipped.

- If all three cards are diamonds, the player gets a payout of $\$12$ (not including the dollar they already paid)
- If exactly two cards are diamonds, the player gets a payout of $\$4$
- If exactly one of the cards is a diamond, the player gets a payout of z.
- If none of the cards are diamonds, the player gets no payout.

You want all the games in your casino to be **fair games** – the expected profit (payout minus cost) of your game should be $\$0$. What should the value of z be? Give your answer as a simplified fraction (don't worry if your number couldn't be paid out with standard U.S. currency)

4. Adorable Triplets [14 points]

You have 10 sets of identical triplets of pets (you are extremely lucky–that's 30 animals total!). 3 triplets of kittens, 5 triplets of puppies, and 2 triplets of bunnies. All three pets in a triplet are identical, but each set of triplets is distinct from each other. So, for example, one set of 3 bunnies is all identical to each other; the other set of 3 bunnies are all identical to each other, but a bunny from the first triplet is not identical to a bunny from the other triplet.

A triple of pets is "perfectly matched" if they are from the same triplet (same species AND group of 3). A triple of pets is "adorably mismatched" if they are of the same species but come from at least two different triples (for example, three puppies that aren't from the same triplet are adorably mismatched; but a kitty, puppy and bunny is not, and three identical bunnies are not).

You (uniformly) randomly create triples out of the 30 pets. That is, you divide the pets into 10 groups of 3. In each of these problems clearly define any random variables you need.

(a) What is the expected number of triples of perfectly matched pets? [7 points]

(b) What is the expected number of "adorably mismatched" triples of pets? [7 points]

5. Mixology Certification [11 points]

In order to impress his study group at an upcoming dinner party, Abed needs to create an extra special drink. To do this, Abed bought a set of n different cool (cool cool) ingredients[1]. He decides that he will mix exactly five ingredients to create his special drink.

Abed's study group is has strong opinions: Abed's special drink will taste good if and only if the five ingredients he mixes are pairwise compatible. By pairwise compatible, we mean that for every pair of ingredients among the five, that pair of ingredients are compatible with each other.

Each pair of ingredients is compatible (independently) with probability p. What is the expected number of 5-ingredient-special-drinks that will taste good?

(a) Carefully define random variables for this problem. [4 points]

(b) Calculate the expected number of 5-ingredient-special-drinks that will taste good. [7 points]

[1] Chocolate Milk, Coca-Cola, Sprite, Fanta, Dr. Pepper, Mountain Dew, Barq's, etc.

6. Runs in a Sequence of Coin Flips [15 points]

A coin with probability p of coming up heads is tossed independently n times. What is the expected number of maximal "runs", where a maximal "run" is a maximal sequence of consecutive flips that are the same? For example, the sequence HHHTTHTHHH has 5 maximal runs: HHH, TT, H, T, and HHH. Use linearity of expectation, carefully define indicator rvs, and justify your work.

7. 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[2] – 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.[3]

Collaboration Policy: For the work in 7.3, 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.

7.1. 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 $\frac{1}{100{,}000{,}000}$.

7.1.1. The Prosecution

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]

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

[3] 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.

(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]

7.1.2. The Defense

The defense attorney argues as follows:

> 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

(a) 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?

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

7.2. Your Analysis

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 above. If you use neither bullet, you must incorporate some other information and have something different from either of the analyses in the last subsection. 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.

(a) What is your prior? Briefly explain where it comes from. [1 point]

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

(c) 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]

7.3. 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"[4].

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]

[4] 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.

(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]

(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]

7.3.1. Some Ideas

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

- We saw in class that routine 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?

7.3.2. Sample Solutions

We will post sample solutions on Ed so you have a sense of what to expect.

8. Feedback [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.

- 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?
- Do you have any thoughts about specifically "Real-World: Bayes Theorem" that you would like to share with us?
- Any other feedback for us?