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 October 22.
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.

0. Keep it Rolling [Gradescope]

Complete these questions about supports and PMFs of random variables by answering these questions on this gradescope assignment. Please check your answers as it will be autograded so we cannot give partial credit.

1. CDF to PMF [7 points]

Let X be a discrete random variable that takes integer values from 1 to 8 (inclusive), and has the following cumulative distribution function (CDF):

$$F_X(n) = \begin{cases} 0 & \text{if } n < 1 \\ \frac{\lfloor (n+1) \rfloor \cdot \lfloor (n+2) \rfloor}{90} & \text{if } 1 \leq n \leq 8 \\ 1 & \text{if } n > 8 \end{cases}$$

Find the probability mass function (PMF) for X.

2. 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 deck with the standard 4 suits and 13 ranks, for 52 cards total) so that 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, so their net-winnings is $\$11$)

- If exactly two cards are diamonds, the player gets a payout of $\$6$
- 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 can't be paid out with standard U.S. currency).

Problems 3-5 will be much easier after learning linearity of expectation on Friday.

3. A quadruple of quadruplets [14 points]

You have 9 sets of identical quadruplets of puppies (that's 36 puppies in total!). There are 2 quadruplets of golden retrievers, 3 quadruplets of corgis, and 4 quadruplets of huskies. All four puppies in a quadruplet are identical, but each set of quadruplets is distinct from each other. For example, one set of 4 golden retrievers is all identical to each other; the other set of 4 golden retrievers are all identical to each other, but a golden retriever from the first quadruplet is not identical to a golden retriever from the other quadruplet.

A quadruplet of puppies is "perfectly matched" if they are from the same quadruplet (same species AND same group of 4). A group of 4 puppies is "adorably mismatched" if they are of the same breed but come from at least two different quadruplets (for example, two huskies from the first quadruplet and two huskies from another quadruplet are adorably mismatched; 2 huskies, 1 golden retriever and 1 corgi are not, and 4 identical corgis are not).

You (uniformly) randomly create groups of 4 puppies out of the 36 puppies. That is, you divide the puppies into 9 groups of 4. In each of these problems, clearly define any random variables you need.

(a) What is the expected number of "perfectly matched" groups of 4 puppies? [7 points]

(b) What is the expected number of "adorably mismatched" groups of 4 puppies? [7 points]

4. Plantability [11 points]

Kevin has n different plant seeds. He wants to plant 7 of them in a pot. However, not all seeds can be planted together in a pot; Kevin can only plant 7 seeds in the pot if none of the 7 plants are incompatible with each other.

For each pair of plants, there's a probability p (independently) that they are incompatible. What is the expected number of 7-plant groups that Kevin can plant in a pot? (It's ok if the same plant appears in multiple groups).

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

(b) Calculate the expected number of 7-plant groups that Kevin can grow in a pot. [7 points]

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

A biased coin has probability p of being heads each time you flip it. It is flipped n times, independently. What is the expected number of maximal "runs", where a maximal "run" is a sequence of consecutive flips that all land on the same side? For example, $HHHTTHHHHHHTHT$ has six maximal runs, where H represents a head, and T represents tails.

Use linearity of expectation, carefully define indicator RVs, and justify your work.

6. Real-World: Bayes Theorem [13 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]

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:

> 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.

[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.

(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 approximately 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]

7. 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?