Homework 4

Due: Wednesday, July 24th, by 11:59pm

Instructions

See the instructions and FAQ for homeworks on the course website for important notes on the submission format!

Solutions submission. You must submit your solution via Gradescope. In particular:

- Submit a *single* PDF containing the solution to all of Tasks 1-6 to Gradescope under "**HW4 [Written]**".
- Each numbered task should be solved on its own page (or pages). Follow the prompt on Gradescope to link tasks to your pages.
- Do not write your name on the individual pages – Gradescope will handle that.

Task 0 – Collaborators [0 pts]

List the full names of anyone you collaborated with on this homework. If you did not collaborate with anyone, write "None" in this section.

Task 1 – On Safari in an RV [12 pts]

Here, we will get some practice with identifying and using random variables from the discrete *zoo* of *random variables* (RVs). These are all some applications of these random variables in CS-related contexts! For each problem (1) **identify the distribution** the random variable follows **and the parameters** to this distribution (2) give a **brief 1-2 sentence justification** and (3) provide the **expected value** of this random variable.

Note that it's possible for a random variable to be a scaled version of one of the random variables from the zoo (e.g., $10X$ where X follows a geometric distribution). An example is shown below:

> *You are running a buffet that has a fixed price of $40 per person. On average, you have 20 customers in one hour. You want to model the earnings you will get in one hour.*
>
> (1) The random variable we're interested in is $20X$ where X is the number of customers in the hour, and $X \sim \text{Poi}(20)$.
> (2) The number of customers follows a Poisson distribution because we're looking for the number of successes (customers) in a time interval, and each customer is independent of each other. The parameter is 20 because, the average number of customers in an hour is 20.
> (3) Plugging into the formula $\mathbb{E}[X] = 20$

a) A file has a 95% chance of being successfully downloaded. Each attempt of downloading the file is independent of any previous attempts. To determine a timeout limit, you want to model the number of tries it will take to successfully download the file.

b) You are tasked with analyzing the rate of incoming API requests to optimize server performance. An external API sends independently requests to your server at an average rate of 5 requests per minute - you want to model the number of requests received by the server in an hour.

c) 200 packets are being sent across the Internet, and each packet takes a different path. Each path fails independently with probability 0.15. You want to model the number of packets that are successfully sent across.

d) 200 packets are being sent across the Internet, but now all packets take the exact same path which fails with probability p. Thus, either all the packets get through or none get through. You want to model the number of packets the are successfully sent across.

NOTE: For questions that involve linearity of expectation, you MUST include the following steps:

- Clearly **define the random variable(s)** you are looking for the expected value of
- Show how you're **breaking that random variable into a sum of random variables**, and clearly define what those random variables are
- Explicitly show **how you're using linearity of expectation**
- Explain **how you're computing the expectation of each** of those random variables, and the final answer

Task 2 – Five-Star Drinks [12 pts]

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? Be sure to clearly define all random variables.

Task 3 – If you give a monkey a typewriter... [12 pts]

Every second, a monkey types one word uniformly at random from the 4-word set $\{$business, is, monkey, serious$\}$. The word it types is independent of words it types at other times. We let the monkey type for 1 hour (i.e., $60 \cdot 60 = 3600$ seconds), what is the expected number of times that the phrase "monkey business is serious business" appears?

Task 4 – Underwater Basket Weaving [10 pts]

In a class about "Underwater Basket Weaving" there are N submissions for an essay assignment where students write about how important the class is. To ensure academic integrity, a checker is being used to compare each pair of submissions to check for similarities. Every pair of submissions will be compared, so the checker will be run $M = \binom{N}{2} = \frac{N(N-1)}{2}$ times.
Unfortunately, we don't know how many people are going to submit the assignment, so to plan, we want to analyze the expected number of times the checker will be run (M) based on distribution that N might follow.

Answer each of the following questions (5 points each). Make sure that each of your answers are **not** in the form of a summation.

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

a) What is the expected value of M if N has a Poisson distribution with parameter λ? (Your answer will be in terms of λ.)
Hint: it may be useful to recall that for any random variable $E(X^2) = Var(X) + [E(X)]^2$.

b) What is the expected value of M if N follows the following PMF:

$$p_N(k) = \begin{cases} 0.4 & \text{if } k = 10 \\ 0.35 & \text{if } k = 8 \\ 0.25 & \text{if } k = 5 \\ 0 & \text{otherwise} \end{cases}$$

Task 5 – Sample Sampling Algorithm [18 pts]

Time to use our 312 knowledge to do some algorithm analysis! Consider the following algorithm for generating a random sample S of size n from the set of integers $\{1, 2, \ldots, N\}$, where $0 < n < N$.

```
Sample(N, n):
  I = 0     S ← ∅                    // S is a set of distinct integers, initially an empty set
  while |S| < n do
    I = I + 1                        // I is counting the total number of iterations of the loop
    x ← Roll(1, N)                   // x is the outcome of rolling a fair N-sided dice
    S ← S ∪ {x}                      // if x is already in S it doesn't change
  return S
```

Let I be the number of die rolls until S is returned. Also, let I_i be the random variable which describes the number of rolls it takes from the time the set S has $i-1$ values to the first time a new value is added after that (i.e., the set S has i values).
Answer the following questions (6 points each)

a) What type of random variable from our zoo is I_i and what is/are the relevant parameter(s) for that random variable?

b) What is I in terms of the random variables I_i? Calculate $\mathbb{E}[I]$, expressing the result as a summation that depends on both N and n.

c) What is $\text{Var}(I)$? You can leave your answer in summation form.

Task 6 – Flee the Sphere [16 pts]

Note: Make sure you fully understand the "Throwing a Dart" problem from section 5 before doing this question!
A flea of negligible size is trapped in a large, spherical, inflated beach ball with radius r. (Recall that such a ball has volume $\frac{4}{3}\pi r^3$.) At this moment, it is equally likely to be at any point within the ball. Let X be the distance of the flea from the center of the ball. For X, find ...

a) (4 points) the cumulative distribution function $F_X(k)$.

b) (4 points) the probability density function $f_X(k)$.

c) (4 points) the expected value $\mathbb{E}[X]$.

d) (4 points) the variance $\text{Var}(X)$

Task 7 – Bloom filters [Coding] [10 pts]

Google Chrome has a huge database of malicious URLs, but it takes a long time to do a database lookup (think of this as a typical Set). They want to have a quick check in the web browser itself, so a space-efficient data structure must be used. In this problem, you will be implementing a **probabilistic data structure** for this.
To solve this task, we have set up a corresponding edstem lesson here. **Implement the functions** add **and** contains **in the** BloomFilter **class of** bloom.py**.** Press the Mark button above the terminal to run the unit tests we have written for you. Passing these unit tests is not enough. We have written a number of different tests for the Gradescope autograder. Your score on Gradescope will be your actual score - you have unlimited attempts to submit.

Task 9 – For Fun Problem - only if you're interested

This is a challenge problem for those of you that would enjoy proving a cool probabilistic fact. You will not be turning anything in for this. It's just for fun.

You are shown two envelopes and told the following facts:

- Each envelope has some number of dollars in it, but you don't know how many.
- The amount in the first envelope is different from the amount in the second.
- Although you don't know exactly how much money is in each envelope, you are told that it is an integer number of dollars that is at least 1 and at most 100.
- You are told that you can pick an envelope, look inside, and then you will be given a one-time option to switch envelopes (without looking inside the new envelope). You will then be allowed to keep the money in envelope you end up with.

Your strategy is the following:

1. You pick an envelope uniformly at random.
2. You open it and count the amount of money inside. Say the result is x.
3. You then select an integer y between 1 and 100 uniformly at random.
4. If $y > x$, you switch envelopes, otherwise you stay with the envelope you picked in step (a)

Show that you have a better than 50-50 chance of taking home the envelope with the larger amount of money in it. More specifically, suppose the two envelopes have i and j dollars in them respectively, where $i < j$. Calculate the probability that you take home the envelope with the larger amount of money.