CSE 312: Foundations of Computing II — Winter 2021

Problem Set 5 (due Wednesday, February 17, 11:59pm)

Directions:
***Answers**: For each problem, remember you must briefly explain/justify how you obtained your answer, as correct answers without an explanation will receive **no credit**. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide. It is fine for your answers to include summations, products, factorials, exponentials, or combinations; you dont need to calculate those all out to get a single numeric answer, for instance 26^7 or $26!/7!$ or $26 \cdot \binom{26}{7}$.*

Your solutions need to be concise and clear. We will take off points for lack of clarity or for excess verbosity. Please see section worksheet solutions (posted on the course website) to gauge the level of detail we are expecting.

Please clearly indicate your final answer, in such a way as to distinguish it from the rest of your explanation.

***Groups:** This homework is to be completed in groups of 1 or 2. Specific guidelines about collaboration are available on the syllabus, but every group will be submitting their own submission. Please cite any collaboration at the top of your submission. Instructions are available here as to how to add groupmates to your submission.*

Before you start your homework, write down the list of people you collaborated with. Remember, you can only collaborate outside your group by discussing the problems at a high level. Provide names and email addresses.

***Submission**: You must upload a **pdf** of your written solutions to Gradescope under "PSet 5 [Written]". Problem 9 is a coding problem, so under "Pset 5 [Coding]" you will be uploading a .py file called cse312_pset5_gen_rvs.py. Problem 10 is an extra credit problem, so if you answer it upload your solution to "Pset 5 [Extra]". (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. We will take off points for hand-written solutions that are difficult to read due to poor handwriting and neatness.)

Please cite any collaboration at the top of your submission (beyond your group members, who should already be listed).

1. Random Grades (9 points)

Every week, 20,000 students flip a 10,000-sided fair dice, numbered 1 to 10,000, to see if they can get their GPA changed to a 4.0. If they roll a 1, they win (they get their GPA changed). You may assume each student's roll is independent. Let X be the number of students who win.

(a) [3 Points] For any given week, give the appropriate probability distribution (including parameter(s)), and find the expected number of students who win.

(b) [3 Points] For any given week, find the exact probability that at least 2 students win. Give your answer to 5 decimal places.

(c) [3 Points] For any given week, estimate the probability that at least 2 students win, using the Poisson approximation. Give your answer to 5 decimal places.

2. Instagram (15 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". What will the **expected** time (in milliseconds) for CPU demand per client be (as a function of N, p or λ) if N follows

(a) [3 Points] the "distribution" where N is the same fixed number with probability 1?

(b) [4 Points] the Poisson distribution with parameter λ?

(c) [4 Points] the geometric distribution with parameter p?

(d) [4 Points] $N = 80X + 5$, 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.

3. Binomial From Nowhere (10 points)

Consider repeatedly rolling a fair 6-sided die, each roll being independent of the others. Define the random variable Y to be the number of rolls until (and including) the first roll of a 6, and define the random variable X to be the number of 1's rolled before the first 6 is rolled. Show that $\Pr(X = j \mid Y = i)$, as j ranges over its possible values, is the probability mass function of a binomially distributed random variable and determine its parameters n and p.

4. Sample Sampling Algorithm (10 points)

Consider the following algorithm for generating a random sample of size n from the set of integers $\{1, 2, \ldots, N\}$, where $0 < n < N$.

```
1  Sample(N, n):
2     I = 0
3     chosen = {}                    // chosen is a set of distinct integers, initially an empty set
4     while |chosen| < n:
5        I += 1                      // I is counting the total number of rolls of the die
6        chosen.add(RollDie(N))      // if the roll of the die (which is random in 1 ... N)
7                                    // is not in chosen, then add it to chosen.
8     return chosen
```

(a) [7 Points] Let the random variable I_i be the number of rolls it takes from the time the chosen set has $i-1$ values to the first time a new value is added (i.e., the chosen set has i values). What type of random variable from our zoo is I_i and what is/are the relevant parameter(s)? What is I in terms of the random variables I_i? Calculate $E[\mathrm{I}]$ in terms of N and n.

(b) [3 Points] What is $\mathrm{Var}(\mathrm{I})$? You can leave your answer in summation form.

5. Explore the zoo! (20 points)

[**Coding**] Understanding the process that leads to different random variables is a great way to gain familiarity for what they mean. For each random variable, write a function that simulates its generation process. Your function should return a random sample of that rv, with the appropriate probability. The **only** function you **can and should** use to generate randomness is np.random.rand(): a function that returns a uniform random float

in the range $[0, 1]$. Note that a function from one part may call a function from a previous part if you wish. For more clarity, we are asking you to generate a random sample from a particular distribution; multiple calls to your function can and should return different values in its range, approximately matching that variable's probability mass function.

Write your code for the following parts in the provided file: cse312_pset5_gen_rvs.py

(a) $X \sim Ber(p)$: 1 with probability p and 0 with probability $1 - p$.

(b) $X \sim Bin(n, p)$: the number of heads in n independent flips of a coin with probability of heads p. Implement the function gen_bin.

(c) $X \sim Geo(p)$: the number of flips up to and including the first head, when the probability of heads is p. Implement the function gen_geo.

(d) $X \sim NegBin(r, p)$: the number of flips up to and including the r-th head in independent coin tosses, when the probability of heads is p. Implement the function gen_negbin. (**Note:** We did not cover this in class. You can read about it in Section 3.5.3.)

(e) $X \sim HypGeo(N, K, n)$: the number of kit kats you get when you grab n random candies from a bag consisting of N total candies, only K of which are kit kats. Implement the function gen_hypgeo. (**Note:** We did not cover this in class. You can read about it in Section 3.6.3.)

(f) $X \sim Poi(\lambda)$: the number of events in a minute, where the historical rate is λ events per minute. Implement the function gen_poi.

(g) Given an arbitrary list (or numpy array) of probabilities, like $ps = [0.1, 0.3, 0.4, 0.2]$, sample an index with the appropriate probability. That is, return 0 with probability 0.1, 1 with probability 0.3, 2 with probability 0.4, and 3 with probability 0.2. Implement the function gen_arb.

6. A Math Problem (8 points)

For this exercise, give exact answers as simplified fractions. Compute $\mathbb{E}[X]$ and $\text{Var}(X)$ if X has probability density function given by ...

$$f(x) = \begin{cases} c(1 - x^4) & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}.$$

Determine the value of c as part of your answer.

7. Exponential Darts (10 points)

You throw a dart at a circular target of radius r. Let X be the distance of your dart's hit from the center of the target. You have improved and your aim is such that $X \sim \text{Exponential}(4/r)$. (Note that it is possible for the dart to completely miss the target.)

(a) [6 Points] As a function of r, determine the value m such that $\Pr(X < m) = \Pr(X > m)$. Then, for $r = 10$, give the value of m to 3 decimal places.

(b) [4 Points] What is the probability that you miss the target completely? Give your answer to 3 decimal places.

8. The Classic Flea Problem (16 points)

A flea of negligible size is trapped in a large, spherical, inflated beach ball with radius r. 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) [5 Points] the cumulative distribution function F.

(b) [5 Points] the probability density function f.

(c) [3 Points] the expected value.

(d) [3 Points] the variance.

Reminder: the volume of a sphere of radius r is $\frac{4}{3}\pi r^3$.

9. Normal, normal, normal (11 points)

(a) [3 Points] Suppose that X is normally distributed with mean 50 and standard deviation 10. Calculate the probability that $25 < X < 75$.

(b) [3 Points] Apparently IQ is roughly normally distributed, with a mean of 100 and a standard deviation of about 15 (note: evidence shows IQ doesn't do a good job at actually measure intelligence, it only measure how good you are at doing that one test). What fraction of people would be classified as genius (IQ of 140 or above)?

(c) [5 Points] The height of American adult females is approximately normally distributed with a mean of 64 inches (5' 4") and a standard deviation of 2.7 inches. Approximately what fraction of American females are 5' tall or less? If we form a basketball team by picking 5 American adult females at random, calculate the probability that at least one of them is over 6 feet (72 inches) tall.

By the way, the above claims about the mean and variance of various distributions are pretty much made up.

10. Extra credit: Which one is real? (5 points)

Below are two sequences of Heads and Tails, each (supposedly) representing 300 independent flips of a fair coin. One of these sequences was truly randomly created, and one was typed by a human. Both sequences have exactly 149 heads. Which one is more likely to be the "real" random sequence? In your write-up, you should justify your reasoning with evidence and valid results, e.g. from running your code on the two sequences. There are multiple valid and correct approaches. To be eligible for full credit, you are required to turn in your detailed analysis along with your Python code.

(a) (Sequence 1)

```
TTHHTHTTHTTTHTTTHTTTHTTHTHHTHHTHTHHTTTHHTHTHTTHTHHTTHTHHTHTT
THHTTHHTTHHHTHHTHTTHTHTTHHTHHHTTHTHTTTHHTTHTHTHTHTHTTHTHTHHH
TTHTHTHHTHHHTHTHTTHTTHHTHTHTHTTHHTTHTHTTHHHTHTHTHTTHTTHHTTHT
HHTHHHTTHHTHTTHTHTHTHTHTHTHHHTHTHTHTTHTHHTHTHTTHTTTHHTHTTTHT
HHTHHHHTTTHHTHTHTHTHHHTTHHTHTTTHTHHTHTHTHHTHTTHTTHTHHTHTHTHH
```

(b) (Sequence 2)

```
HTHHHTHTTHHTTTTTTTTHHHTTTHHTTTTHHTTHHHTTHTHTTTTTTHTHTTTTHHHH
THTHTTHTTTHTTHTTTTHTHHTHHHHTTTTTHHHHTHHHTTTTTHTHTTHHHHTHHHHH
HHTTHHTHHTHHHHHHHTTHTHTTTHHTTTTHTHHTTHTTHTHTHTTHHHHHTTHTTTHT
HTHHTTTTHTTTTTHHTHTHHHHTTTTHTHHHHHHTHTHTHTHHHTHTTHHHTHHHHHHT
HHHTHTTTHHHTTTHHTHTTHHTHHHTHTTHTTHTTTHHTHTHTTTTHTHTHTTHTHTHT
```