Note: This homework assignment covers Chapter 3.

Disclaimer: If you use R, include all R code and output as attachments. Do not just "write in" the R code you used. Also, don't just write the answer and say this is what R gave you. If my grader can't see how you got an answer, it is wrong. I want to see your code and your answers accompanying your code (like in the notes).

1. Patient responses to a generic drug to control pain is scored on 5-point scale (1 = lowest pain level; 5 = highest pain level). The distribution of a patient's score Y has been estimated from historical information and is given by

y	1	2	3	4	5
$p_Y(y)$	0.38	0.27	0.18	0.11	0.06

(a) Graph the pmf of Y and the cdf of Y side by side (like in the notes).
(b) Compute $E(Y)$ and var(Y). Place an "×" on the pmf indicating where $E(Y)$ is.
(c) A pharmacist plans to observe patients (i.e., "trials") in succession, regarding each patient's score as independent. Let X denote the number of patients, out of 25, who respond with a 1 or 2. Graph the pmf of X and calculate $E(X)$.
(d) In part (c), would you find it to be unusual if 20 or more patients responded with a 1 or 2? Explain.

2. An environmental engineer working for the EPA is tasked with observing water specimens from lakes in northeast Georgia. In this region, each water specimen has a 10 percent chance of containing a particular organic pollutant.

(a) Treating each water specimen as a "trial," suppose the three Bernoulli trial assumptions hold. State what this would imply (i.e., just give the assumptions for this situation).
(b) Let Y denote the number of specimens that contain the pollutant out of the next 15 analyzed. Plot the pmf and cdf of Y side by side (like in the notes). Place an "×" on the pmf indicating where $E(Y)$ is.
(c) In part (b), find the probability that

- there is exactly 1 polluted sample

- there are 3 or more polluted samples (this event will trigger an EPA intervention).

Note: For practice, I want you to do part (c) "by hand," that is, show all the calculations with pencil and paper. Use R to check your work.

(d) In part (b), if the engineer found 10 or more of the 15 specimens to be polluted, what might be true?

3. On a recent trip to the SC DMV, I asked an employee to estimate what percentage of SC drivers arrive to renew their driver's license with one that is currently expired. She responded that about 30 percent of all such renewals were of this type.

(a) Suppose you observe Y, the number of SC DMV customers seeking renewal to find the first one with an expired license. What is the distribution of Y? Plot the pmf and cdf of Y side by side (like in the notes). Place an "$\times$" on the pmf indicating where $E(Y)$ is.
(b) Let W denote the number of SC DMV customers seeking renewal to find the 3rd one with an expired license. What is the distribution of W? Plot the pmf and cdf of W side by side (like in the notes). Place an "$\times$" on the pmf indicating where $E(W)$ is.
(c) Obviously, in parts (a) and (b), you are assuming that Bernoulli trial assumptions hold. State what these are in this application (e.g., think of each customer seeking renewal as a "trial.")
(d) In parts (a) and (b), find the probability that

- among the first 6 customers seeking renewal, none have expired licenses.

- you have to observe 10 or more customers seeking renewal to find the 3rd one with an expired license.

Note: For practice, I want you to do part (d) "by hand," that is, show all the calculations with pencil and paper. Use R to check your work.

4. When I looked at the class roll for this class in early January, there were 26 engineering majors and 24 majors that were not engineering. Suppose I selected 8 students at random (and without replacement) from a class of this composition.

(a) Let Y denote the number of engineering majors in the sample. Graph the pmf of Y and the cdf of Y side by side (like in the notes). Place an "$\times$" on the pmf indicating where $E(Y)$ is.
(b) Find the probability that

- the sample of 8 contains all engineering majors.

- the sample contains more non-engineering majors than engineering majors.

Note: For practice, I want you to do part (b) "by hand," that is, show all the calculations with pencil and paper. Use R to check your work.

5. An automobile manufacturer is concerned about a fault in a braking mechanism of a particular model. The fault can, on rare occasions, cause a catastrophe at high speed. Suppose that the distribution of Y, the number of cars per year that will experience the catastrophe, is Poisson with mean $\lambda = 2.2$.

(a) Graph the pmf of Y and the cdf of Y side by side (like in the notes). Place an "×" on the pmf indicating where $E(Y)$ is.

(b) Find the probability that

- no cars will experience the catastrophe

- 5 or more cars will experience the catastrophe.

Note: For practice, I want you to do part (b) "by hand," that is, show all the calculations with pencil and paper. Use R to check your work.

(c) The cost (to the manufacturer) associated with catastrophe is potentially enormous. Risk management experts have estimated that the yearly cost associated with these catastrophes (in 1000s of dollars) is described by the function $C = 150 + 1000Y + 0.1Y^2$. Find the expected cost per year.

6. The **mode** of a discrete random variable Y is the value of y that gives the largest probability $p_Y(y)$; i.e., the mode is the "most likely value" of Y.

(a) Find the mode of Y in Problem 5. Interpret what this value means.

Another discrete distribution that is useful in modeling counts is the *logarithmic series distribution*. A random variable Y has this distribution if its probability mass function (pmf) is

$$p_Y(y) = \begin{cases} \dfrac{-(1-p)^y}{y \ln p}, & y = 1, 2, 3..., \\ 0, & \text{otherwise.} \end{cases}$$

The value p satisfies $0 < p < 1$. Regardless of what the value of p is, the mode of this distribution is 1. Note also that observing $y = 0$ is not allowed under this model.

(b) Plot the pmf and cdf of the logarithmic series random variable Y when $p = 0.2$. Note that while Y can take on arbitrarily large values, you will have to pick an upper bound for graphing purposes. Make sure you choose this upper bound to be large enough so that you can see the entire distribution.

(c) Here are the number of goals per game in the 2013-2014 English Premier League season:

Goals	0	1	2	3	4	5	6	7	8	9	10+
Frequency	27	73	80	72	65	39	17	4	1	2	0

Note that there were 380 games total. For example, 30 of the games ended in a 0-0 draw, 79 of the games ended 1-0, one game had a total of 9 goals, etc. Would the logarithmic series distribution be a good distribution to model the number of goals scored per game? Why or why not?