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. The probability mass function (pmf) of Y, the number of imperfections per 10 meters of a synthetic fabric (in rolls of a fixed width), is given by

y	0	1	2	3	4
$p_Y(y)$	0.41	0.37	0.16	0.05	0.01

(a) Graph the pmf of Y and the cdf of Y side by side (like in the notes).
(b) Find the probability that there are at least 2 imperfections.
(c) Find the probability that there are at most 2 imperfections.
(d) Compute $E(Y)$ and var(Y). Place an "$\times$" on the pmf indicating where $E(Y)$ is.
(e) I observed Y for 10 consecutive 10-meter pieces of fabric (on an assembly line). The observations were 1, 1, 0, 0, 4, 0, 2, 2, 0, 1, 0. Do you think these observations are consistent with this model? Explain why or why not. Give a set of 10 observations that are grossly inconsistent with this model.

2. One quality characteristic for cabinet manufacturers is easy sliding drawers. A drawer is considered to be "easy sliding" if it does not get stuck when opened. Historically, about 2 percent of the drawers get stuck. Suppose that 20 drawers are randomly selected from a large lot and are tested. The lot will be shipped if all 20 experience "easy sliding." Let Y denote the number of drawers that get stuck (out of 20).
(a) Treating each drawer 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) What is the probability that the lot will be shipped?
(c) Find the probability that at least two drawers will get stuck.
(d) Graph the pmf of Y and the cdf of Y side by side (like in the notes).
(e) Find the mean, variance, and standard deviation of Y.

3. Screening for infectious diseases in a blood-bank setting is a major part of ensuring blood safety. At a local clinic, subjects' blood donations are tested for infection. Suppose that 5 percent of all blood donations are infected in some way (e.g., HIV, HCV, syphilis, etc.). For simplicity, assume that all subjects are independent.
(a) Let Y denote the number of subjects tested to find the first infected blood donation. Find $P(Y > 3)$. Interpret what this probability means in words.
(b) In part (a), calculate $P(Y > 5|Y > 2)$. How does this answer compare with $P(Y > 3)$? Why do you think this is?
(c) Suppose that, during a given day, there are 30 donations. Find the probability that no more than two of these donations are infected.
(d) Of the 30 donations in part (c), 10 donations come from African American (AA) donors and 20 come from non-AA donors. If I pick 3 donations at random, what is the probability that exactly one is from an AA donor?

4. A recent geological study in western Texas indicates that exploratory oil wells strike with probability 0.20 (i.e., oil is found).
(a) Treating each well as a "trial," suppose that drilling wells in this region obeys the three Bernoulli trial assumptions. State what this would imply (i.e., just give the assumptions for this situation).
(b) What is the probability that the 1st successful well is found on the 4th well drilled?
(c) What is the probability that it will take more than 4 wells to find the 2nd successful well?

5. Let Y denote the number of calls received per day by the USC Campus Police. Suppose that Y has a Poisson distribution with $\lambda = 6.5$.
(a) 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) What is the probability that on a given day there are exactly 5 calls? at least 5 calls? at most 5 calls?
(c) The **mode** of a discrete random variable Y is the value of y that maximizes the pmf $p_Y(y) = P(Y = y)$. Find the mode of Y in this example.
(d) Suppose that the daily cost (in dollars) to respond to Y calls is given by

$$g(Y) = 150 + 100Y + 5Y^2$$

Find the expected daily cost.
(e) Suppose that in a given week, there are 30 calls received. Twelve of the calls involved (in some form or another) illegal consumption of alcohol and 18 did not. If administration picks 5 cases (calls) to review at random, what is the probability that at least 4 of these cases will involve illegal consumption of alcohol?

6. Discovery of a new drug involves screening large chemical libraries to identify active compounds. These libraries are usually quite large, with thousands or perhaps millions of compounds, and the proportion p of active compounds is often quite small. We will assume the probability that each individual compound is active is p and that each compound is independent.
(a) Discuss how the Bernoulli trial assumptions apply here; talk about individual drug compounds (not generic "trials"). Also, identify a practical situation where the Bernoulli trial assumptions would not hold in this context.

For the remainder of this problem, we will assume that the Bernoulli trial assumptions hold. From a discovery point of view, the obvious question is,

> "How should we search an entire library of compounds to find those compounds that are active?"

Suppose that there are $N = mk$ compounds in a chemical library and that we would like to determine the active/nonactive status of each compound. Consider the following two ways to do this:

(i) Each compound can be tested separately. This will require N tests.

(ii) Form m pools of k compounds by assigning each compound to exactly one pool. Test the pools. If a pool tests negative, all k compounds in it are negative (and only 1 test

is needed). If a pool tests positive, each of the k compounds will subsequently be tested separately (therefore, $k+1$ tests will be required for the k compounds).

Important: We will assume that the test that classifies compounds and compound pools is perfect; there are no mistakes in classification.

(b) What is the probability that a single pool of k compounds will test positive?
(c) Let Y denote the number of tests needed to screen the entire library under (ii). Find an expression for $E(Y)$, the expected number of tests. Your answer here should depend on m, k, and p.
(d) In terms of minimizing the expected number of tests to be performed on the N compounds, which plan, (i) or (ii), would be preferred if p is close to 0? Justify your answer using the expression derived in part (c).
(e) For concreteness (only for this part), suppose that $N = 100000$, and consider the following choices of (m, k) under plan (ii):

$$m = 20000, k = 5 \quad m = 10000, k = 10 \quad m = 5000, k = 20 \quad m = 2000, k = 50$$

For each choice, graph $E(Y)$ as a function of p over the range $0 < p < 0.20$. Try to identify regions of p where each (m, k) combination above would be preferred; i.e., where $E(Y)$ is smallest.
(f) We made an assumption that there are no mistakes in classification (i.e., positive pools test positively; negative pools test negatively). Provide two realistic scenarios where such an assumption may be violated.