Problem Set 1

This problem set covers material from Week 1, dates 2/13- 2/16. Textbook problems can be found at the end of the corresponding subsection.

Instructions: Write or type complete solutions to the following problems and submit answers to the corresponding Canvas assignment. Your solutions should be neatly-written, show all work and computations, include figures or graphs where appropriate, and include some written explanation of your method or process (enough that I can understand your reasoning without having to guess or make assumptions). A general rubric for homework problems appears on the final page of this assignment.

Tuesday 2/13

Mini probability review!

1. Let $X \sim f_X(x)$, where for some $a > 0$,

$$f_X(x) = \begin{cases} ax^{a-1} & 0 < x < 1 \\ 0 & o.w. \end{cases}$$

   (a) Using the definition of expectation, find $\mathbb{E}[X]$.

   (b) Confirm your answer in (a) by identifying the well-named distribution of X, for which a formula for expectation exists.

2. Let $X_1, \ldots, X_{25} \overset{\text{iid}}{\sim} \text{Poisson}(2)$, and define $Y = \sum_{i=1}^{25} X_i = X_1 + \ldots + X_{25}$.

   (a) What distribution does Y follow (state name and parameters)?

   (b) What is $P(Y = 40)$?

   (c) Express $P(Y < 40)$ as a sum. Then use `R` to find this probability.

   (d) Using Central Limit Theorem, Y is approximately Normally distributed. What is the mean of variance of this Normal distribution?

   (e) Using the Normal approximation, approximate $P(Y < 40)$ using $\Phi()$ notation.

   (f) Use `R` and the Normal approximation to estimate $P(Y < 40)$. *For this problem, a separate .Rmd output is not necessary; you can just hand-write the code you used if you prefer.*

Thursday 2/15

3. For each of the following, identify the *random variables* that are being observed, the *statistical model* under consideration, the *parameter(s)* of this model, and the *parameter space.*

   (a) An ecologist would like to learn about the water temperature of Otter Creek. However, it is known that water temperatures can fluctuate and the thermometer used to obtain measurements is a bit faulty. For this reason, it is good practice to take repeated measurements, and the ecologist decides to obtain 5 independent measurements. Suppose that the true temperature θ of the water is unknown, but the variance in the measurements obtained is 0.1 degrees. Further suppose that the thermometer used to obtain the measurements is known to be positively biased by 0.2 degrees (i.e. the average value of the errors is 0.2). Assume that the errors are otherwise identically distributed normal random variables.

(b) A species of invasive insect is making its way across Vermont. The insects injure pine trees by crawling inside the trunk and eating the insides, resulting in a weaker tree. However, some trees are more resistant to the insects than others! Suppose each pine tree is able to resist the insects with unknown probability p, independent of other trees. Healthy pine trees that are able to resist the insects will produce pine cones independently, where the number of pine cones is distributed Poisson with unknown mean λ_1. However, if a tree is unable to resist the insects, the number of pine cones it produces is distributed Poisson with a different unknown mean λ_2. An ecologist studies a sample of 100 pine trees in Vermont and for each tree, observes whether or not the tree is infected and counts the number of pine cones it produced.

4. A jar contains 20 marbles, each of varying color and mass (but otherwise same shape/size). Marbles are drawn uniformly at random with replacement, one-by-one, from the jar. Every time we draw a marble, we record its mass before putting it back into the jar. Suppose we draw three marbles. Let M_1, M_2, and M_3 be the masses of the three marbles drawn. Further suppose c is an unknown positive constant, and that the jar contains exactly one marble of each of the following masses: $\{c, 2c, 3c, \dots, 19c, 20c\}$.

   (a) What is/are the random variable(s) of interest?

   (b) State the implied model for the data, parameter(s) for this model, and the parameter space.

   (c) Give examples of three different statistics which, by definition, could be defined using just the data M_1, M_2, M_3. Which of these statistics do you think would be most useful for estimating the value of c? Why?

Friday 2/16

5. Textbook 7.2: Problem 2

6. Suppose the proportion θ of defective items produced by a large manufacturing company is unknown. We use the following prior for θ:

$$p(\theta) = 2(1-\theta) \qquad 0 < \theta < 1$$

Now we obtain the data as follows: items are randomly selected from the company until exactly three are found defective. Items from the lot are selected at random one by one until exactly three defectives have been found. If we find that we must select a total of eight items in this experiment, what is the posterior distribution of θ at the end of the experiment? *Hint: What is the sampling model here?*

General rubric

Points	Criteria
5	The solution is correct *and* well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect.
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification.
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information).
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.
Notes:	For problems with code, well-written means only having lines of code that are necessary to solving the problem, as well as presenting the solution for the reader to easily see. It might also be worth adding comments to your code.