Lisa Yan
CS 109

Problem Set #5
July 30, 2018

Problem Set #5
Due: 1:30pm on Monday, August 6th

With problems by Mehran Sahami and Chris Piech

For each problem, briefly explain/justify how you obtained your answer. Brief explanations of your answer are necessary to get full credit for a problem even if you have the correct numerical answer. The explanations help us determine your understanding of the problem whether or not you got the correct answer. 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 don't need to calculate those all out to get a single numeric answer.

Unless otherwise stated, you may also use functions in a library like Python's `scipy.stats` to compute values of PMFs and CDFs; if you use these, provide your code that calls these functions and explain how you arrived at each parameter to a function or constructor.

1. Say we have two independent variables X and Y, such that $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$. Mathematically derive an expression for $P(X = k \mid X + Y = n)$.

2. Choose a number X at random from the set of numbers $\{1, 2, 3, 4, 5\}$. Now choose a number at random from the subset no larger than X, that is from $\{1, \ldots, X\}$. Let Y denote the second number chosen.

   a. Determine the joint probability mass function of X and Y.
   b. Determine the conditional mass function of X given $Y = i$. Do this for $i = 1, 2, 3, 4, 5$.
   c. Are X and Y independent? Justify your answer.

3. Say we have a coin with unknown probability X of coming up heads when flipped. However, we believe (subjectively) that the prior probability (before seeing the results of any flips of the coin) of X is a Beta distribution, where $E[X] = 0.5$ and $\text{Var}(X) = 1/36 \approx 0.0278$.

   a. What are the values of the parameters a and b (where $a, b > 1$) of the prior Beta distribution for X?
   b. Now say we flip the coin 12 times, obtaining 8 heads and 4 tails. What is the form (and parameters) of the posterior distribution of $(X \mid 12 \text{ flips resulting in 8 heads and 4 tails})$?
   c. What is $E[X \mid 12 \text{ flips resulting in 8 heads and 4 tails}]$?
   d. What is $\text{Var}(X \mid 12 \text{ flips resulting in 8 heads and 4 tails})$?

4. You are tracking the distance of a satellite from Earth by reading values from a distance measurement instrument. Before you observe the instrument reading, your belief of the distance D of the satellite was a Gaussian distribution $D \sim N(\mu = 98, \sigma^2 = 16)$. The instrument gives a reading that is true distance plus Gaussian noise G, where $G \sim N(0, 4)$. Suppose the instrument reports that the satellite is 100 a.u. from Earth.

   a. What is the PDF of your prior belief of the true distance of the satellite?

b. What is the probability density of seeing an observation of 100 a.u. from your instrument, given that the true distance of the satellite is equal to t?

c. What is the PDF of your posterior belief (after observing the instrument reading) of the true distance of the satellite? You may leave a constant in your PDF and you do not need to simplify the PDF.

5. Let $X_1, X_2, \ldots$ be a series of independent random variables which all have the same mean μ and the same variance σ^2. Let $Y_n = X_n + X_{n+1}$. For $j = 0$, 1, and 2, determine $\mathrm{Cov}(Y_n, Y_{n+j})$. Note that you may have different cases for your answer depending on the value of j.

6. Let X_1, X_2, X_3, and X_4 be a set of pairwise uncorrelated random variables (i.e., $\rho(X_i, X_j) = 0$ when $i \neq j$), which all have the same mean μ and the same variance σ^2.

   a. What is the correlation $\rho(X_1 + X_2, X_3 + X_4)$?
   b. What is the correlation $\rho(X_1 + X_2, X_2 + X_3)$?
   c. What is the correlation $\rho(2X_1, X_1 + X_2)$?

7. In class, we considered the following recursive function:

```
int recurse() {
  int x = randomInteger(1, 3);
    // randomInteger is equally likely to return 1, 2, or 3
  if (x == 1) return 3;
  else if (x == 2) return (5 + recurse());
  else return (7 + recurse());
}
```

Let Y = the value returned by `recurse()`. We previously computed $E[Y] = 15$. What is $\mathrm{Var}(Y)$?

8. Consider the following function, which simulates repeatedly rolling a 6-sided die (where each integer value from 1 to 6 is equally likely to be "rolled") until a value ≥ 3 is "rolled".

```
int roll() {
  int total = 0;
  while (true) {            // loop forever
    int roll = randomInteger(1, 6);
      // randomInteger is equally likely to return 1,...,6
    total += roll;
    if (roll >= 3) break;   // exit condition
  }
  return total;
}
```

   a. Let X = the value returned by the function `roll()`. What is $E[X]$?
   b. Let Y = the number of times that the die is "rolled" (i.e., the number of times that `randomInteger(1, 6)` is called) in the function `roll()`. What is $E[Y]$?

9. You go on a camping trip with two friends who each have a mobile phone. Since you are out in the wilderness, mobile phone reception isn't very good. One friend's phone will independently drop calls with 10% probability. Your other friend's phone will independently drop calls with 25% probability. Say you need to make 6 phone calls, so you randomly choose one of the two phones and you will use that *same* phone to make all your calls (but you don't know which has a 10% versus 25% chance of dropping calls). Of the first 3 (out of 6) calls you make, one of them is dropped. What is the conditional expected number of dropped calls in the 6 total calls you make (conditioned on having already had one of the first three calls dropped)?

10. Let X = the number of requests you receive at your web site per minute, where $X \sim \text{Poi}(12)$. Each request, independently of all other requests, is equally likely to be routed to one of N web servers. Compute the expected number of web servers that will receive at least one request each during a minute. (Hint: there are a few ways to do this problem, but one way you might approach it is to first determine the *conditional* expectation of the number of web servers that receive at least one request each during a minute, conditioned on some fixed number, k, of requests during that minute. Then use that result to compute the *unconditional* expectation of the number of web servers that receive at least one request each during a minute.)

11. **[Coding]** In this question you are going to learn how to calculate p-values for experiments that are called *A/B tests*. These experiments are ubiquitous. They are a staple of both scientific experiments and user interaction design.

    Massive online classes have allowed for distributed experimentation into what practices optimize students' learning. Coursera, a free online education platform that started at Stanford, is testing out new ways of teaching a concept in probability. They have two different learning activities `activity1` and `activity2` and they want to figure out which activity leads to better learning outcomes. After interacting with a learning activity Coursera evaluates a student's learning outcome by asking them to solve a set of questions.

    Over a two-week period, Coursera randomly assigns each student to either be given `activity1` (group A), or `activity2` (group B). The activity that is shown to each student and the student's measured learning outcomes can be found in the file `learningOutcomes.csv`.

    a. What is the difference in sample means of learning outcomes between students who were given `activity1` and students who were given `activity2`?
    b. Calculate a p-value for the observed difference in means reported in part (a). In other words: assuming the learning outcomes for students who had been given `activity1` and `activity2` were identically distributed, what is the probability that you could have sampled two groups of students such that you could have observed a difference of means as extreme, or more extreme, than the one calculated from your data? Provide any code you used to calculate your answer.
    c. The file `background.csv` stores the background of each user. Student backgrounds fall under three categories: more experience, average experience, less experience. For each of the three backgrounds, calculate a difference in means in learning outcome between `activity1` and `activity2`, and the p-value of that difference.

Yes, this problem set has only 11 problems on it. Be happy.