Lisa Yan
CS 109

Problem Set #5
November 6, 2019

Problem Set #5
Due: 1:00pm on Friday, November 15th

With problems by Mehran Sahami, Chris Piech, and David Varodayan

Errata (Updated Friday 11/8, 9am): In Question 9, the function that you are asked to write pseudocode for has been renamed to `inferProbFlu()` to avoid naming conflicts with other provided functions.

For each problem, briefly explain/justify how you obtained your answer. In fact, most of the credit for each problem will be given for the derivation/model used as opposed to the final answer. Make sure to describe the distribution and parameter values you used where appropriate. **Provide a numeric answer for all questions when possible.**

1. 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)?

2. You are developing medicine that sometimes has a desired effect, and sometimes does not. With FDA approval, you are allowed to test your medicine on 9 patients. You observe that 7 have the desired outcome. Your belief as to the probability of the medicine having an effect before running any experiments was Beta(2, 2).
    a. What is the distribution for your belief of the probability of the medicine being effective after the trial?
    b. Use your distribution from (a) to calculate your confidence that the probability of the drug having an effect is greater than 0.5. You may use `scipy.stats` or an online calculator.

3. **[Coding]** Let X be the sum of 100 independent uniform random variables each of which are identically distributed as Uni(0, 1). Simulate 100,000 calculations of X.
    a. Use the simulations to calculate the probability that X is in the ranges: [35 to 40], [40 to 45] as so on up until [60 to 65]. Draw a bar graph of your results.
    b. Write the probability density function for an approximation of X.
    c. Use your answer to part (b) to calculate the probability that X is in the range 40 to 45. Make sure that your answer aligns with the result you reported in part (a). Round your result to two decimal places.

4. A fair 6-sided die is repeatedly rolled until the total sum of all the rolls exceeds 300. Approximate (using the Central Limit Theorem) the probability that *at least* 80 rolls are necessary to reach a sum that exceeds 300.

Continued on the next page. . .

5. Program A will run 20 algorithms in sequence, with the running time for each algorithm being independent random variables with mean = 46 seconds and variance = 100 seconds^2. Program B will run 20 algorithms in sequence, with the running time for each algorithm being independent random variables with mean = 48 seconds and variance = 200 seconds^2.

    a. What is the approximate probability that Program A completes in less than 900 seconds?
    b. What is the approximate probability that Program B completes in less than 900 seconds?
    c. What is the approximate probability that Program A completes in less time than Program B?

Better Peer Grading

6. **[Coding]** Stanford's HCI class runs a massive online class that was taken by ten thousand students. The class used peer assessment to evaluate students' work. We are going to use their data to learn more about peer graders. In the class, each student has their work evaluated by 5 peers and every student is asked to evaluate 6 assignments: five peers and the "control assignment" (the graders were unaware of which assignment was the control). All 10,000 students evaluated the same control assignment, and the scores they gave are in the file `peerGrades.csv`. You may use simulations to solve any part of this question.

    a. What is the sample mean of the 10,000 grades to the control assignment?
    b. Students could be given a final score which is the *mean* of the 5 grades given by their peers. Imagine the control experiment had only received 5 peer-grades. What is the variance of the mean grade that the control experiment would have been given? Show your work and describe any code you used to calculate your answer.
    c. Students could be given a final score which is the *median* of the 5 grades given by their peers. Suppose the control experiment had only received 5 peer-grades. What is the variance of the median grade that the control experiment would have been given? Show your work and describe any code you used to calculate your answer.
    d. Is the expected median of 5 grades different than the expected mean of 5 grades?
    e. Would you use the mean or the median of 5 peer grades to assign scores in the online version of Stanford's HCI class? Hint: it might help to visualize the scores.

Continued on the next page. . .

A/B Testing

7. **[Coding]** In this problem you are going to learn how to use and misuse 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.

   Suppose you are working at Coursera on new ways of teaching a concept in probability. You have two different learning activities `activity1` and `activity2` and you want to figure out which activity leads to better learning outcomes.

   Over a two-week period, you randomly assign each student to be given either `activity1` or `activity2`. You then evaluate each student's learning outcomes by asking them to solve a set of problems. The data (the activity shown to each student and their measured learning outcomes) are 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. Write a program to estimate the p-value (using the bootstrap method) for the observed difference in means reported in part (a). In other words: assuming that 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 in part (a)? Provide any code you used to calculate your answer.

   Scientific journals have traditionally accepted an experiment's result as "statistically significant" if the p-value is below 0.05. By definition, this standard means that 5% of findings published in these journals are in fact not true, but just false positives. The scientific community is beginning to move away from using arbitrary p-value thresholds to determine whether a result is publishable. For example, see this 2019 editorial in the journal *Nature*: `https://www.nature.com/articles/d41586-019-00874-8`.

   You are now troubled by the p-value you obtained in part (b), so you decide to delve deeper. You investigate whether learning outcomes differed based on the background experience of students. The file `background.csv` stores the background of each student as one of three labels: more experience, average experience, less experience.

   c. 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.

   d. Your manager at Coursera is concerned that you have been "p-hacking," which is also known as data dredging: `https://en.wikipedia.org/wiki/Data_dredging`. In one sentence, explain why your results in part (c) are not the result of p-hacking.

Continued on the next page. . .

Learning While Helping

8. You are designing a randomized algorithm that delivers one of two new drugs to patients who come to your clinic—each patient can only receive one of the drugs. Initially you know nothing about the effectiveness of the two drugs. You are simultaneously trying to learn which drug is the best and, at the same time, cure the maximum number of people. To do so we will use the Thompson Sampling Algorithm.

> **Thompson Sampling Algorithm:** For *each* drug we maintain a Beta distribution to represent the drug's probability of being successful. Initially we assume that drug i has a probability of success: $\theta_i \sim \text{Beta}(1, 1)$.
>
> When choosing which drug to give to the next patient we **sample** a value from each Beta and select the drug with the largest **sampled** value. We administer the drug, observe if the patient was cured, and update the Beta that represents our belief about the probability of the drug being successful. Repeat for the next patient.

   a. Say you try the first drug on 7 patients. It cures 5 patients and has no effect on 2. What is your belief about the drug's probability of success, θ_1? Your answer should be a Beta.

Method	Description
`V = sampleBeta(a, b)`	Returns a real number value in the range [0, 1] with probability defined by a PDF of a Beta with parameters a and b.
`R = giveDrug(i)`	Gives drug i to the next patient. Returns a True if the drug was successful in curing the patient or False if it was not. Throws an error if $i \notin \{1, 2\}$.
`I = argmax(list)`	Returns the index of the largest value in the list.

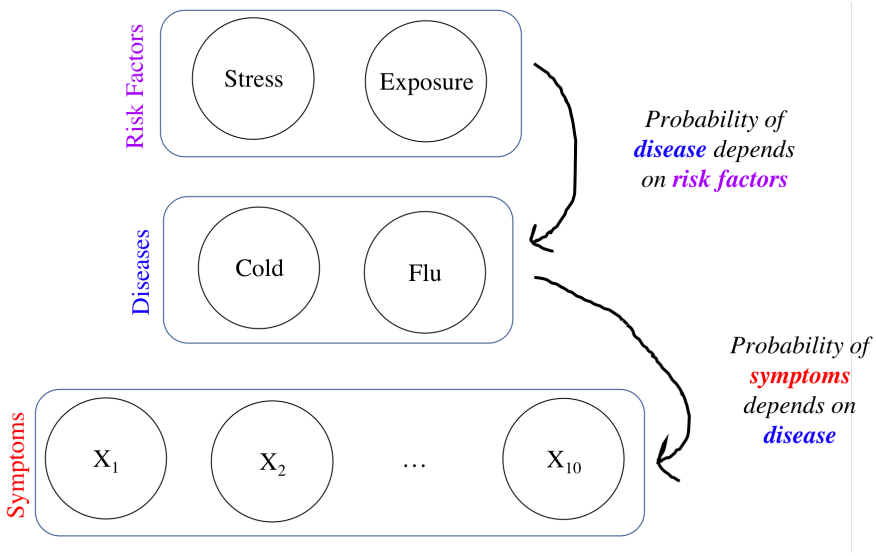
   b. Write pseudocode to administer either of the two drugs to 100 patients using Thompson's Sampling Algorithm. Use functions from the table above. Your code should execute `giveDrug` 100 times.

   c. After running Thompson Sampling Algorithm 100 times, you end up with the following Beta distributions:
   $\theta_1 \sim \text{Beta}(11, 11)$,
   $\theta_2 \sim \text{Beta}(76, 6)$.
   What is the expected probability of success for each drug?

Continued on the next page. . .

WebMD

9. We are writing a WebMD program that is slightly larger than the one we worked through in class. In this program we predict whether a user has a flu ($F = 1$) or cold ($C = 1$) based on knowing any subset of 10 potential binary symptoms (e.g., headache, sniffles, fatigue, cough, etc) and a subset of binary risk factors (exposure, stress).



- We know the prior probability for Stress is 0.5 and Exposure is 0.1.
- The functions `probCold(s, e)` and `probFlu(s, e)` return the probability that a patient has a cold or flu, given the state of the risk factors stress (`s`) and exposure (`e`).
- The function `probSymptom(i, f, c)` which returns the probability that the `i`th symptom (X_i) takes on value 1, given the state of cold (`c`) and flu (`f`): $P(X_i = 1 | F = f, C = c)$.

We would like to write pseudocode to calculate the probability of flu *conditioned on observing* that the patient has had exposure to a sick friend and that they are experiencing Symptom 2 (sore throat). In terms of random variables $P(\text{Flu} = 1 \mid \text{Exposure} = 1, X_2 = 1)$:

```
def inferProbFlu() # P(Flu = 1 | Exposure = 1 and X2 = 1)
```

a. Write pseudocode that calculates `inferProbFlu()` using **Rejection Sampling**.

b. (Reach) Write pseudocode that calculates `inferProbFlu()` without using sampling.

Note: Causality implies the following: (1) risk factors are independent; (2) all diseases are independent of one another conditioned on risk factors; and (3) all symptoms are independent of one another conditioned on knowing the state of diseases.

$$P(S = s, E = e) = P(S = s)P(E = e) \quad (1)$$

$$P(C = c, F = f | S = s, E = e) = P(C = c | S = s, E = e) \cdot P(F = f | S = s, E = e) \quad (2)$$

$$P(\text{symptoms} | F = f, C = c) = \prod_j P(X_j = k_j | F = f, C = c) \quad (3)$$

"Reach" means I don't *expect* CS109 students to be able to solve the problem. But thinking about it will be useful. Show your work. If you get stuck (> 15 mins), explain what is hard.