Chris Piech
CS109

PSet #5
Nov 8, 2017

Problem Set #5
Due: 3pm on Friday, Nov 17th

For each problem, **explain/justify how you obtained your answer** in order to obtain full credit. 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.

Warmup:

1. 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 effect is greater than 0.5. You may use scipy.stats or an online calculator.

2. [Coding] Let X be the sum of 100 independent uniform random variables each of which are identically distributed as Uniform(0, 1). Simulate 100,000 calculations of X.
    a. Use the simulations to calculate the probability that X, rounded to the ones digit, is 30, 31, 32 and so on up until 60. Draw a bar graph of your results.
    b. Use the Central Limit Theorem to come up with a distribution for X.
    c. Use your answer to part (b) to calculate the probability that X is in the range 47.5 to 48.5. Make sure that your answer aligns with the result you reported in part (a). Round your result to two decimal places.

3. An amateur university band passes around a pot for donations after a concert. There are 50 people in the audience. Each person gives money independently with the same distribution (IID). Each individual has a:
    0.10 probability that they give $0
    0.20 probability that they give $1
    0.35 probability that they give $5
    0.30 probability that they give $10
    0.05 probability that they give $20
    a. What is the expected amount of money that each person gives?
    b. What is the variance of the amount of money that each person gives?
    c. Give a probability distribution that approximates the total amount of money that the band earns.
    d. What is the approximate probability that the band makes at least $350?

4. Let X be the outcome of a dice roll. Let $Y = X^2$. What is the covariance between X and Y?

5. Let $X \sim N(\mu = 1, \sigma^2 = 2)$ and $Y \sim N(\mu = 1, \sigma^2 = 2)$. What is the distribution for $2X + Y$?

6. You roll 6 dice. How much more likely is a roll with: [one 1, one 2, one 3, one 4, one 5, one 6] than a roll with six 6s? Think of your dice roll as a multinomial.

Probabilistic Program Analysis:

7. 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); // 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]?

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

```
int Recurse() {
   int x = randomInteger(1, 3);  // 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 found that E[Y] = 15. What is Var(Y)?

9. Program A will run 20 algorithms in sequence, with the running time for each algorithm being independent random variables with mean = 50 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 = 52 seconds and variance = 200 seconds^2.

   a. What is the approximate probability that Program A completes in less than 950 seconds?

   b. What is the approximate probability that Program B completes in less than 950 seconds?

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

A/B Testing

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 – and promise to be able to scale more personalized educational experiences. Coursera, a free online education platform that started at Stanford, is testing out a set of 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.

10. Bad testing. For both of the experiments bellow, point out a flaw in the experimental design:
    a. For the duration of two weeks Coursera tests activity1. Then for subsequent two weeks Coursera tests activity2. Coursera compares student learning outcomes.
    b. For the duration of two weeks, Coursera gives all users in the western hemisphere activity1 and all the users in the eastern hemisphere activity2. Coursera decides which activity is more useful in general based on the difference in learning outcomes.

11. A/B testing. 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? Describe any code you used to calculate your answer.
    c. 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. Describe your methodology.

Better Peer Grading using Probability

12. [Coding] Stanford's HCI class runs a massive online class that was taken by ten thousand students. The class used peer assessment to evaluate student's 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 un-aware 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. Imagine 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**.

Bounding Midterms

13. From past experience, we know that the midterm score for a student in CS106Z is a random variable with mean = 75. Assume that exam scores can be real values (i.e., fractional points can be given), but scores cannot be negative.

    a. Use Markov's Inequality to give an upper bound for the probability that a student's midterm score will be greater than or equal to 85.
    b. Now, say we are given the additional information that the variance of a student's midterm exam score in CS106Z is 25 (and you can use this information for parts (c) as well). Use Chebyshev's Inequality to provide a bound on the probability that a student's midterm score is between 65 and 85, inclusive.
    c. According to the Central Limit Theorem, how many students would have to take the midterm in order to ensure, with at least 90% probability, that the class average would be within 5 of 75?

An Origin Story

14. [Coding] We observed that the distribution of grades in our midterm looked Beta, however we don't have a story that explains why! In this problem we are going to code a probabilistic model that can explain how the dynamics of exams can result in the observed distribution.

    Item Response Theory states that the number of points that a student *i* with an "exam-ability" A_i receives when they solve a midterm question *j* with "difficulty" b_j worth n_j points is:

    $$n_j \cdot f(A_i - b_j) \qquad \text{where} \qquad f(x) = \frac{1}{1 + e^{-x}}$$

    Assume that points are rounded to the nearest whole number and a student's overall score is the sum of the points they receive on each problem. Both the difficulty and the point value of each problem are known (thus they are constants and not random variables) however the exam-ability of each student is a random variable. Difficulties are in the range 0 to 10 where 10 is the hardest possible midterm question and 0 is the easiest. Assume that each exam-ability A_i is an IID sample from a distribution. We just have no idea what that distribution is!

    a. Come up with a first hypothesis for the distribution of exam-abilities (eg uniform) of students in a CS106Z class and chose initial values for the parameters. You may change your answer later! Don't chose Beta. It is good for exam scores, not student abilities ☺.

    b. Program your exam-ability distribution. Simulate N = 200 student abilities and calculate each student's resulting ***percentage*** score on a midterm (the test was out of 118 so divide total score by 118). Draw a histogram with the number of students who score between (0 and 10], (10 and 20], (30 and 40] and so on where each bucket is exclusive of the lower limit and inclusive of the upper limit. The difficulty and point values of all the problems are included in midtermProblems.csv.

    c. The true distribution for the class midterm is provided in trueDistribution.csv. Consider the bucket counts in your histogram from part (b) to be outcomes of a multinomial distribution, where the probability of each outcome is given by the true distribution. What is an equation for the probability of your histogram counts given the true probabilities? What is an equation for the log of that probability?

    d. Try different values of your parameters to find ones that maximize the log probability equation from part (c). Report the parameters that maximize the log probability (and the corresponding score). If you are having trouble getting a log probability greater than -18 you might want to try a different choice of distribution in part (a).

    Hint: Try to only change one parameter at a time. For each parameter setting, it may help to simulate 100 exams and look at the average log probability. Logs of probabilities should be negative. Larger log values are negative numbers closer to 0.

    e. Final grades are a sum of grades from each problem, but the overall distribution is not Normal. How come the Central Limit Theorem didn't apply?

    f. Since midterms are different each quarter (with different difficulty levels) it is hard to compare the exam scores of one class to the exam scores of another class. How could you use your new model to compare students from different quarters?

Students in the Stanford Fall 2017 CS109 class had a notably higher average ability than the typical CS109 class (with about the same variance).