Chris Piech
CS 109

Problem Set #5
Nov 6th, 2021

Problem Set #5
Due: 1pm, Nov 15th

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. When asked to succinctly describe how you calculated your answer, you should provide a sentences to a paragraph explaining the relevant methodology as if to another student in CS109. If you would like to include equations or short pseudo code (up to 10 lines), that is fine.

Warmup

1. 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? Assume X and Y are independent.

2. (Coding) Let X be the sum of 100 independent Beta random variables each of which are identically distributed as Beta(4, 2). Simulate 100,000 calculations of X.
    a. Write pseudo code that you used to simulate X.
    b. Use the simulations to calculate the probability that X, rounded to a whole number, takes on the values in the range 60 to 75 inclusive. Draw a bar graph of your results.
    c. Use the Central Limit Theorem to come up with an approximate distribution for X. Explain your strategy in a few sentences.
    d. Use your answer to part (c) to calculate the probability that X is in the range 70.5 to 69.5. Make sure that your answer aligns with the result you reported in part (b). Round your result to two decimal places.

3. A fair 6-sided die is repeatedly rolled until the total sum of all the rolls exceeds 300. Approximate the probability that at least 80 rolls are necessary to reach a sum that exceeds 300.

4. 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?

5. 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 an approximate probability distribution for the total amount of money earned.
d. What is the approximate probability that the band makes at least $350?

Algorithmic Analysis

6. 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".

```
def roll():
  total = 0;
  while (True):
    # equally likely to return 1,...,6
    roll = randomInteger(1, 6)
    total += roll
    # exit condition
    if (roll >= 3) break
  return total
```

a. Let X be the value returned by the function roll(). What is E[X]?
b. Let Y be 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]?

7. Our ability to fight contagious diseases depends on our ability to model them. One person is exposed to llama-flu. The method below returns the number of individuals who will get infected.

```
# Get number of people infected by one individual
def num_infected():
   # most people are immune to llama-flu
   immune = bernoulli(p = 0.99)
   if immune: return 0

   # people who are not immune, spread the disease far
   spread = 0

   # they make contact with k people (up to 100)
   k = binomial(n = 100, p = 0.25)
   for i in range(k)
     spread += num_infected():

   # total infections should include this individual
   return spread + 1
```

What is the expected return value of `num_infected`?

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.

8. 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? Succinctly describe how you calculated your answer.

   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? Succinctly describe how you calculated 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. Succinctly describe how you calculated your answer.

Better Peer Grading

9. (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 control assignment grades? Succinctly describe how you calculated your answer.

   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 of those five grades? Succinctly describe how you calculated 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? Succinctly describe how you calculated your answer.

   d. What is the difference in the expected median of 5 grades and 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? Explain why. Hint: it might help to visualize the scores.