Homework 7: Multiple RVs and Concentration

Version 2: Posted 5/17 at 3:30 PM. We corrected a mistake in 3f, and corrected answers to the autograded problems in gradescope.

For each problem (except those marked as autograded), remember you must briefly explain/justify how you obtained your answer, as correct answers without an explanation will not receive full credit. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide.

In general, your goal in an explanation is to write enough that a student from class who has attended lecture, but not read the problem yet, could understand your approach, verify your reasoning, and believe your answer is correct. While we do not usually need to see arithmetic, you must include enough work that in principle one could rederive your answer with only a scientific calculator. For each problem, make sure to explicitly define all random variables you use, and be clear about how they are related to each other using proper notation (conditionals, summations, etc.).

Unless a problem states otherwise, you should leave your answer in terms of factorials, combinations, etc., for instance 26^7 or $26!/7!$ or $26 \cdot \binom{26}{7}$ are all good forms for final answers.

Instructions as to how to upload your solutions to gradescope are on the course web page.

Remember that you must tag your written problems on Gradescope.

Submission: There will be three submission boxes for HW7:

- A submission box for problems 1,2. These will be autograded (similar to a concept check). We won't give partial credit, but have divided the problems into more parts to find common mistakes. We hope this will save you some time writing and the TAs some time grading this week.
- A submission box for problems 3-5. These will be graded like previous homeworks have been.
- A submission box for problem 6.

You must upload a **pdf** of your written solutions to Gradescope under "HW 7". (Instructions as to how to upload your solutions to gradescope are on the course web page.) The use of latex is *highly recommended*. (Note that if you want to hand-write your solutions, you'll need to scan them. We will take off points for hand-written solutions that are difficult to read due to poor handwriting and neatness.)

Due Date: This assignment is due at 11:59 PM Wednesday May 22.
You will submit the written problems as a PDF to gradescope; to reduce grading load, a few marked problems will be entered into gradescope in boxes that allow for automatic grading (similar to a concept check).g
For the written problems, please put each numbered problem on its own page of the pdf (this will make selecting pages easier when you submit), and ensure that your pdfs are oriented correctly (e.g. not upside-down or sideways). The coding problem will also be submitted to gradescope.

Collaboration: Please read the full collaboration policy. If you work with others (and you should!), you must still write up your solution independently and name all of your collaborators somewhere on your assignment.

0. Extra Instructions :o

0.1. Evaluating Integrals

For calculations that require evaluating integrals (unless we indicate otherwise), you must

(a) Show the integral to evaluate (e.g., $\int_0^2 z \cdot 2dz$)

(b) Show an antiderivative and the values to evaluate at (e.g., $z^2|_0^2$)

(c) Plug in the values and simplify (e.g., $2^2 - 0^2 = 4$)

This is not a problem, so nothing needs to be submitted here.

0.2. Autograded Problem

To help manage grading load, we will autograde a subset of the problems on this homework. Those marked "autograded" will **not** require a pdf submission. Instead we will have a gradescope assignment (setup similar to a concept check) where you enter your answers and gradescope will automatically evaluate them. Due to limitations in gradescope, these will be graded without partial credit, but you will get instant feedback on whether your submission is correct.

1. Covariance?!?! [15 points] — autograded

Recall the 'Hungry Washing Machine' question from section. You have 10 pairs of socks, where each is a different color. Left and right socks are distinct. Your washing machine randomly eats 4 socks. Suppose that your washing machine is equally likely to eat any 4 of your 20 distinct socks. For $i = 1, \ldots, 10$, let X_i be the indicator/Bernoulli random variable that is 1 if pair i is complete after this fiasco and 0 otherwise. Let $Y = \sum_{i=1}^{10} X_i$ be the total number of complete pairs of socks that you have left.

(a) What is Cov (X_i, X_j) where $1 \leq i, j \leq 10$ and $i = j$? [4 points]

(b) What is Cov (X_i, X_j) where $1 \leq i, j \leq 10$ and $i \neq j$? [4 points]

(c) What is Var (Y)? Give your answer as a simplified fraction. [7 points]
Hint: You will want to use your answers to the previous parts.

2. Carrots and Crops [15 points] — autograded

Peter Rabbit is in a large garden with 4 paths, and each path is equally likely to be chosen.

- The first path will exit the garden[1] after a number of hours which is Geometric with parameter $\frac{1}{8}$.
- The second path will exit the garden after a number of hours which is Poisson with parameter 2.
- The third path will exit the garden after a number of hours which is Exponential with parameter $\lambda = \frac{1}{5}$.
- The fourth path leads to a path which brings him back to where he started after 3 hours.

Use the law of total expectation to compute the expected number of hours until Peter Rabbit exits the garden.

3. Whale Watching [24 points]

Avery is going whale watching for H hours, where H is a random variable, equally likely to be 1, 2 or 3. The number of whales W she sees is random and depends on how long she is in the area for. We are told that

$$\Pr(W = w \mid H = h) = \frac{c}{h}, \qquad \text{for } w = 1, \ldots, h,$$

for some constant c.

(a) Compute c. You may want to use one of the axioms of probability.

(b) Find the joint distribution of W and H using the chain rule.

(c) Find the marginal distribution of W.

(d) Find the conditional distribution of H given that $W = 1$ (i.e., $Pr(H = h \mid W = 1)$ for each possible h in 1,2,3). Use the definition of conditional probability and the results from previous parts.

[1] Peter Rabbit gets quite distracted stealing and eating the delicious carrots in the garden.

(e) Suppose that we are told that Avery saw either 1 or 2 whales. Find the expected number of hours she spent whale watching conditioned on this event. Use the definition of conditional expectation and conditional probability theorems.

(f) Avery decides to live-stream video of each whale she sees. The amount of money Avery earns for live-streaming each whale sighting is a random variable with mean 3, and the money each time is independent of the other times and of the number of whales Avery sees. What is the expected amount of money Avery earns?
Warning: you might be tempted to skip some steps and assert that expected amount of money earned is 3 times the expected number of whales seen. Even though the answer is intuitive, its formal derivation is a lot more involved. We expect you to show each step you used to get to that expression. Your work must involve the law of total expectation, conditioning on the number of whales seen. We would encourage you to also use indicator random variables, but other simple random variables may suffice.

4. Knitting Requires Concentration [16 points]

Robbie is slowly knitting a blanket, made of 250 squares. It takes an average of 2 hours for Robbie to knit a square, with a variance of 0.4 hours. The time to knit each square is independent.

You should treat time as continuous for this problem.

(a) What is the expectation of the total time to knit the blanket? [2 points]

(b) What is the variance of the total time to knit the blanket? [2 points]

(c) Robbie will have 700 hours to knit between now and when he needs the blanket to be finished to stay warm at a football game. Use Markov's Inequality to bound the probability that Robbie finishes the blanket before the game. Hint: you'll need to take a complement. [6 points]

(d) Robbie will have 535 hours to knit between now and the start of basketball season. Use Chebyshev's inequality to bound the probability that Robbie finishes the blanket after at least 465 hours, but in time for basketball season. [6 points]

5. Bananas [15 points]

You have 2000 students standing in a 40-by-50 grid. You will throw **50,000** bananas (independently) towards the grid of students, with equal probability for each banana to be collected by each student. Additionally, every banana is collected by a student (i.e. there are no bananas left uncollected).

You hope that at the end of the process, each student will have at least 6 bananas. You want to upper bound the probability that you fail to distribute at least 6 bananas to every student.

(a) Suppose your friend is standing at the lower-left corner of the grid. Use a Chernoff bound from class to bound the probability that your friend does not collect at least 6 bananas. [7 points]

(b) Is the probability of your friend collecting less than 6 bananas independent of the probability that the student in the upper-right corner collects less than 6 bananas? Briefly explain (you may give a formal derivation/calculation as an explanation or an informal one). [3 points]

(c) Bound the probability that at least one student collects fewer than 6 bananas. Give the best bound you can from your answers in (a) and (b). [5 points]

6. Real-World: Modeling Assumptions [25 points] — separate submission box

The tools of this class are useful to computer scientists, but many of them are useful beyond just "classic" computer science. In order to use the powerful tools of probability, we need to make assumptions to let our mathematical tools model the real world. Things like "this coin is perfectly fair" or "the coin flips are all independent" are usually not perfectly true.[2] Indeed, occasionally these assumptions are ways that people "lie with statistics" or provide evidence for claims that aren't actually true.

In this question, you will critique the modelling assumptions made in an analysis and see if other modelling assumptions would lead to a different result.

6.1. Find a Model C/NC

Find an analysis (e.g., via a blog post/article) that uses probability and statistics tools you're familiar with from this course. By "analysis," we mean any estimate of a "real-world" probability, along with the assumptions that lead to that number. You might want to look at the examples in section 6.3 for what we mean.

We expect most of the answers to this section will be short (2-3 sentences), but you are free to write more if your resource is more complicated.

(a) Provide a link to (or somehow let us access) the analysis you're critiquing. [3 points]

(b) What is the fundamental claim of the analysis? I.e., what conclusion do they draw at the end of their analysis? [3 points]

(c) What modelling assumptions do they use? (For example, do they assume some occurrences are independent? Do they assume a set of events all have equal probability? Do they assume they know the probability? Do they use a variable from the zoo?) [3 points]

6.2. Improve the model

Now, see if their modelling assumptions are reasonable or if other ones would lead to a different conclusion. We expect parts a,b,d will be a few sentences each (though you can write more if you have more to say).

(a) Identify and explain at least one *weakness* of the modelling assumptions they have made (e.g. a potential dependence on events that are supposed to be independent). [4 points]

(b) Now create your own model for the same problem. You might do this by coming up with different probability estimates for the events, or by using a different random variable (e.g. a binomial distribution instead of a Poisson), or by incorporating some outside knowledge about the problem that you think sheds more light. Briefly describe what your model will be, and how it differs from the previous one. [4 points]

(c) Under your new model, calculate the probability of the event your source calculated. [4 points]

(d) Does the calculation change significantly? If it does, does the conclusion of the analysis change? [4 points]

6.3. Some Ideas

You are free (and encouraged!) to find your own examples outside this list if you have a topic you are passionate about, but if you can't think of anything, you may use any of these as starting points. In many cases, there are already critiques of poor statistical/probability analyses online – it's ok to look at these critiques, as long as you tell

[2] e.g., if you flip a coin repeatedly, the result of the last flip is probably how the coin will appear on your hand before you flip it, which will make the results not quite independent.

us if you're using any and still do the new probability calculation independently and put everything in your own words.

- The probability of having a baby girl/boy might not be a 50/50 chance like we might think or even 49/50. Estimating the probability of having a baby girl/boy might depend on the father's history. See here for more details.
- Richard Lustig, 7-Time Lottery Winner, Gives Tips On Winning The Powerball Jackpot. You might find that his advice doesn't make the best assumptions. See article here.
- Is this an accurate estimate of the probability of being struck by lightning in the US? Could being in different states have an impact on this estimate?
- Some think that the probability of encountering at least one shiny Pokémon does not change with seeing more Pokémon.
- Every year millions of people predict the outcomes of the NCAA men's basketball tournament. It is commonly said that the probability of a perfect bracket is $\frac{1}{2^{63}}$, (since there are 2^{63} ways the 63 games could play out) and therefore no one will ever predict a perfect bracket. (Yes, Robbie has a blog post about this one) Here is a source using that number
- A video-game streamer named "Dream" livestreamed multiple consecutive speedruns of Minecraft where they had incredible luck in a few parts of the game, across the speedruns. So lucky, that a speedrunning organization declared that Dream had to be using a modified version of the game, and that the runs were therefore invalid. The analysis that lead to the rejection of the runs.

7. Feedback [1 point]

Answer these questions on the separate Gradescope box for this question.

Please keep track of how much time you spend on this homework and answer the following questions. This can help us calibrate future assignments and future iterations of the course, and can help you identify which areas are most challenging for you.

- How many hours did you spend working on this assignment (excluding any extra credit questions, if applicable)? Report your estimate to the nearest hour.
- Which problem did you spend the most time on?
- Do you have any thoughts about specifically "Real-World: Modeling Assumptions" that you would like to share with us?
- Any other feedback for us?