Homework 7: Concentration Inequalities

For each problem, 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: You must upload a **pdf** of your written solutions to Gradescope under "HW 7 [Written]". (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 November 19.
You will submit the written problems as a PDF to gradescope.
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.

1. Knitting Requires Concentration [16 points]

Robbie is slowly knitting a blanket, made of 200 squares. It takes an average of 3 hours for Robbie to knit a square, with a variance of 0.5 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 645 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 555 hours, but in time for basketball season. [6 points]

2. Goldfish [15 points]

You have 3000 goldfish bowls in a 60-by-50 grid. Each has one goldfish swimming inside. You will throw **60000** algae wafers (independently) towards the grid of bowls, with equal probability for each wafer to land in each bowl. Additionally, you know that every wafer lands in some bowl.

You *hope* that at the end of the process, each goldfish will eat at least 6 wafers. You want to upper bound the probability that you fail to distribute at least 6 wafers to every goldfish.

(a) Suppose a goldfish named Marmalade is swimming in a bowl at the lower-left corner of the grid. Use a Chernoff bound from class to bound the probability that marmalade does not eat at least 6 wafers. [7 points]

(b) Is the probability of Marmalade eating less than 6 wafers independent of the probability that another goldfish located at the upper-right corner eats less than 6 wafers? Briefly explain (you may give a formal or informal derivation/calculation as an explanation). [3 points]

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

3. Random Airlines [23 points]

You are tired of airlines overcharging you, and want to create your own. There are 100 cities that you would like to consider having flights into or out of. For every pair of cities, you will schedule a direct flight between the two cities independently with probability 0.2. Let Y be the total number of direct flights you schedule.

(a) What is the expectation and variance of Y? [4 points]

(b) Give an upper bound on the probability that $Y \geq 1200$ using Markov's inequality. [4 points]
Let $\mathbb{E}[Y] = a$, and write your answer in terms of a.

(c) Give an upper bound on the probability that $Y \geq 1200$ using Chebyshev's inequality. [4 points]
Let Var $(Y) = b$, and write your answer in terms of a, b.

(d) Give an upper bound on the probability that $Y \geq 1200$ using the Chernoff bound from lecture, and keep your answer in terms of a and b. [4 points]

(e) Use the CLT to approximate Y as a normal random variable, and give the appropriate parameters. Use this approximation (and continuity correction) to approximate the probability that $Y \geq 1200$. Write your answer in terms of $\Phi(\cdot)$, the standard normal random variable CDF, and a and b. [4 points]

(f) Call 7 cities "perfectly connected" if there is a direct flight between each pair within the 7 cities. Using a union bound, upper bound the probability that there exists a "perfectly connected" set of 7 cities. [3 points]

4. Real-World: Modeling Assumptions [12 points]

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.[1] Indeed, occasionally these assumptions are ways that people "lie with statistics" or provide evidence for claims that aren't actually true.

[1] 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.

In this question, you will critique the modelling assumptions made in an analysis.

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 4.1 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]

(d) Do you believe their analysis is correct? If so, choose an assumption from C and explain why you agree with it. If not, what assumption would you change? [3 points]

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

- Some think that the probability of encountering at least one shiny Pokémon does not change with seeing more Pokémon.
- 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?
- 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

5. 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?