Homework 7: Tail Bounds and MLE

In general, your goal in an explanation is to write enough that a student from class who has attended lecture, but not thought about 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.

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.

Remember that you must "select pages" to identify which problem is on which page as part of your gradescope upload.

Submission: You must upload a **pdf** of your written solutions to Gradescope under "HW 6 [Written]". The use of $\LaTeX$ is *required*.

Due Date: This assignment is due at **11 PM Tuesday August 15**.
You will submit the written problems as a PDF to Gradescope. 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.

Note that there will be **no leeway** with the submission time. Any submission to Gradescope after 11:00:00 PM on the deadline day will be considered late. We recommend submitting a few minutes before the deadline in order to avoid technical issues with submissions.

Further, note that the **last possible day** that this assignment can be submitted **if you have sufficient late days** will be **August 16th at 11:00pm**. That is, a maximum of one late day can be used on this assignment.

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

Full submission instructions are in the course syllabus.

0. Extra Instructions :o

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.

1. Viva la correlation [15 Points]

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**. Hint: You will want to use your answers to the previous parts. [7 points]

2. My name is Jared, I'm 19 [16 points]

You read a book, consisting of 50 chapters. It takes an average of 2 hours to read a chapter, with a variance of 0.2 hours. The time to read each chapter is independent.

You should treat time as continuous for this problem.

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

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

(c) You have 150 hours to read between now and when you need to finish the book for your book club. Use Markov's Inequality to bound the probability that you finishes the book before your club meets. Hint: you'll need to take a complement. [6 points]

(d) You will have 120 hours to read between now and the start of your book club meetings. Use Chebyshev's inequality to bound the probability that you finish the book after at least 80 hours, but in time for your book club meetings. [6 points]

3. We continue to yearn [15 points]

You have 2500 urns, that you place into a 50-by-50 grid. You will throw **40,000** balls (independently) toward the grid of urns, with equal probability for the ball to go in each urn.

You hope that at the end of the process, each urn will have at least 4 balls inside. You want to upperbound the probability that you will fail to get at least 4 balls into every urn.

(a) Use a Chernoff bound from class to bound the probability that the urn in the lower-left corner of the grid does not have at least 4 balls inside. [7 points]

(b) Is the probability of the lower-left urn having less than 4 balls independent of the probability that the urn in the upper-right corner has less than 4 balls? 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 urn has fewer than 4 balls. Give the best bound you can from your answers in (a) and (b). The bound may be trivial; sometimes, concentration inequalities are uninformative. [5 points]

4. Gotta Estimate 'em All! [15 points]

Every 312 student is surveyed on their preferred starter Pokemon between Bulbasaur, Charmander, and Squirtle. Suppose that each student independently chooses Bulbasaur with probability θ_B, Charmander with probability θ_C and Squirtle with probability $1 - \theta_B - \theta_C$. (Thus, $0 \leq \theta_B + \theta_C \leq 1$). The parameters θ_B, θ_C are unknown. Suppose that $x_1, ..., x_n$ are n independent, identically distributed samples from this distribution. Let n_B be the number of x_i's equal to Bulbasaur, let n_C be the number of x_i's equal to Charmander and let n_S be the number of x_i's equal to Squirtle. What are the maximum likelihood estimates for θ_B and θ_C in terms of n_B, n_C, and n_S?

In doing this problem, you do not need to do a second-derivative test (or any other test) to confirm you have a maximizer. You may assume any critical point you find is a maximizer.

For this problem, except where noted, all of your work must be understandable without reference to any calculator (including Wolfram-Alpha). You may check your answers using calculators, but your explanation may not rely on them.

(a) Write the Likelihood function. [3 points]

(b) Write the log-likelihood. [2 points]

(c) Give two equations describing the maximizer of the log-likelihood. Be sure to show work (e.g., taking of partial derivatives) on how you got these equations. [8 points]

(d) You should now have a system of two (linear) equations in two unknowns. Find the solution to this system, and use it to write the MLEs for θ_B and θ_C. You do **not** have to show your work for this part, and may use online algebra solvers if you wish. [2 points]

5. To be continued [24 points]

You do not need to check second order conditions in the following.

(a) Let $x_1, x_2, \ldots, x_n$ be independent samples from an exponential distribution with unknown parameter λ. What is the maximum likelihood estimator for λ? [12 Points]

(b) Given $\theta > 0$. Suppose that $x_1, \ldots, x_n$ are i.i.d. realizations (aka samples) from the model

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the maximum likelihood estimate for θ. [12 Points]

6. A journey, not the estimation [5 points]

Let $X_1, X_2, \ldots, X_n$ be random samples drawn from the continuous distribution Uniform$(0, \theta)$, for $\theta > 0$. We stumble upon a possible estimator $\hat{\theta}$ for θ, where $\hat{\theta} = 2 \cdot \frac{\sum_{i=1}^{n} X_i}{n}$. Is $\hat{\theta}$ an unbiased estimator of θ?