18.05 Problem Set 5, Spring 2025

(due on Gradescope Monday, Mar. 17 at 9:00 PM ET)

Problem 1. (20 pts: 5, 5, 5, 5 pts.) **Random walks**

Xeno was deep in thought when he found himself lost in the MIT tunnel that runs under the Math Department. He remembered from his probability course that if he took enough random steps, he would eventually end up back in the math department with high probability, so long as he stayed in the straight tunnel.

Suppose that each second, Xeno either takes one step forward or one step back with equal probability. Xeno's position after n seconds can be modeled as a sum of random displacements $X_1, \ldots, X_n$.

You can visualize Xeno's walks here for intuition (we'll learn about confidence intervals later):

`https://demonstrations.wolfram.com/SimulatingTheSimpleRandomWalk/`

(a) Compute the probability that Xeno is more than 100 steps away **in either direction** from his starting point at 10,000 seconds. Hint: express the answer in terms of a value you can compute with `pbinom` in R.

(b) Use the CLT and `pnorm` to estimate the same probability. How good is the estimate?

(c) Using CLT estimation as in (b), how much time must pass for there to be at least a 90% chance that Xeno is currently more than 100 steps away from where he started?

(d) Now suppose instead that Xeno either takes one step forward, takes one step back, or rests in place with equal probability. His position can no longer be modeled exactly with the binomial. Instead, use the CLT to estimate the same probability as in (b), but with this 3-option setup. Did the probability go up or down? Why does this make sense?

Note: one take-away from this problem is that, in symmetric random walks, the distance from the starting point tends to grow like a constant times the square root of the number of steps. This comes up, for example, in physics (Brownian motion and other diffusion processes) and finance (the movement of stock prices).

Problem 2. (20 pts: 5,5,5,5 pts.) **Fat tails**

Over the past 20 years, the high temperature in Boston on February 29 has averaged 39.9 degrees (all temperatures in Fahrenheit), with a standard deviation (σ) of 12.0 degrees. (The raw temperature data come from NOAA, at `https://www.ncdc.noaa.gov/cdo-web/`.) In this problem, you study the probability of extreme events based on different underlying distributions for the daily high temperature.

For this problem, computations can be done in R.

(a) Consider the uniform distribution with range $[a, b]$.

(i) Find the range, so that the mean is 39.9 and the standard deviation is 12.0.

(ii) Find the probability of an event at least one sigma from the mean (whether above or below the mean)—called a 1σ event;

(iii) Find the probability of a 2σ event.

(iv) Find the probability of a day whose temperature is greater than or equal to 76 degrees.

(b) Now consider the normal distribution, with mean $\mu = 39.9$ and standard deviation $\sigma = 12.0$.

(i) No problem here. We just want the numbering to be parallel to the other parts.

(ii)-(iv) Find the same values as in part (a).

(c) The Laplace distribution has range $(-\infty, \infty)$ and pdf $f(x) = \frac{1}{2b}\mathrm{e}^{-|x-\mu|/b}$. (This is sometimes called the double exponential, but that name is applied to other distributions as well.)

(o) Graph the pdf for $\mu = 0$, $b = 1$. A hand-drawn qualitative sketch is fine. We're just looking for you to show the most obvious features.

(i) Find the values of μ and b that give the same mean and standard deviation as above. For this part, also find the cdf $F(x)$.

You must write down the integrals you would need to compute, and you get pride points for doing the integral yourself, but you can look the Laplace distribution up on Wikipedia to get the cdf, mean and variance.

(ii)-(iv) Find the same values as in part (a).

(d) How many standard deviations above the mean is 76 degrees? Rank the three distributions from best to worst at predicting the probability of a 76 degree day in February in 2030. Briefly justify your answer.

Hence the moral of this problem: knowing the mean and standard deviation of a quantity is often insufficient to predict the frequency of extreme events (100-year floods, etc.). You need to know more about the underlying distribution itself, which may require knowing more about the underlying geophysics, chemistry, or biology. In the solutions, we will show you graphs of these distributions zoomed in around 3σ above the mean. If you do that yourself, you will see that they look very different.

Problem 3. (15: 5, 5, 5 pts.) **Maximum likelihood estimates**

(a) A game is played with 3 identical looking urns containing colored balls. Urn 1 contains 5 red, 5 green, 2 blue balls; urn 2 contains 3 red, 8 green, 4 blue balls; urn 3 contains 7 red, 7 green, 3 blue balls.

The rules are that the game master picks one urn at random and draws 3 balls without replacement. The contestant does not know which urn was chosen, but is told the colors of the 3 balls in the order they were chosen. They then have to guess which urn the balls came from.

Suppose that, in order, the colors are red, green, red. Which urn has the maximum likelihood of producing the scenario just described?

(b) The gamma distribution is another important family of distributions. For instance, the sum of independent identically distributed exponential distributions is a gamma distribution.

For an integer n, the gamma distribution with shape parameter n and rate parameter λ has range $(0, \infty)$ and pdf

$$f(x|\lambda, n) = \frac{\lambda^n}{(n-1)!}x^{n-1}\mathrm{e}^{-\lambda x}.$$

Suppose we have data x_1, x_2, …, x_m drawn from a gamma distribution with $n = 10$ and

unknown λ. Find the MLE for λ.

Suggestion: The likelihood can be compactly written in terms of the sum S and the product P of the data.

(c) I'm always late for dinner. Exactly how late follows a uniform(0,b) distribution where b is not known. (Apparently there is a time beyond which even I won't be late.) My partner wanted to figure out the value of b so they recorded my lateness in minutes for 4 days. The results were 2.5, 19.75, 12.0, 7.0. What is the maximum likelihood estimate for b based on this data.

Hint: this is not a calculus problem.

Problem 4. (20 pts: 5,5,5,5 pts.) **Monty Hall: Sober and drunk**
Recall the Monty Hall problem: Monty hosts a game show. There are three doors: one hides a car and two hide goats. The contestant Aviva picks a door, which is not opened. Monty then opens another door which has a goat behind it. Finally, Aviva must decide whether to stay with her original choice or switch to the other unopened door. The problem asks which is the better strategy: staying or switching?

To be precise, let's label the door that Aviva picks by A, and the other two doors by B and C. Hypothesis H_A is that the car is behind door A, and similarly for hypotheses H_B and H_C.

(a) In the usual formulation, Monty is sober and knows the locations of the car and goats. So if the contestant picks a door with a goat, Monty always opens the other door with a goat. And if the contestant picks the door with a car, Monty opens one of the other two doors at random. Suppose that sober Monty Hall opens door B, revealing a goat. So the data is: 'Monty showed a goat behind B'. Our hypotheses are 'the car is behind door A', etc. Make a Bayes table with prior, likelihood and posterior. Use the posterior probabilities to determine the best strategy.

(b) Now suppose that Monty is drunk, i.e. he has completely forgotten where the car is and is only aware enough to randomly open one of the two doors not chosen by the contestant. It's entirely possible he might accidentally reveal the car, ruining the show.

Suppose that drunk Monty Hall opens door B, revealing a goat. Make a Bayes table with prior, likelihood and posterior. Use the posterior probabilities to determine the best strategy. (Hint: the data is the same but the likelihood function is not.)

(c) Based on Monty's pre-show behavior, Aviva thinks that Monty is sober with probability 1/4 and drunk with probability 3/4. Repeat the analysis from parts (a) and (b) in this situation.

(d) Now assume Monty is sober with probability p and drunk with probability $1-p$. What is the smallest value of p where staying would be the best option?

Problem 5. (20 pts: 5, 5, 5, 5 pts.) **Bayesian dice**

Recall the setup of Studio 5 on Bayesian dice. In this scenario:

- We have our five types of Platonic dice: 4, 6, 8, 12, 20 sided.
- There is a prior distribution of the quantity of each die.

- One die is chosen at random and rolled repeatedly.
- We used Bayesian updating to figure out which die was chosen.

(a) Suppose the pool of die consists of 1 die of each type. Using the Bayesian update table format, compute the final posterior given the following sequence of rolls: 1, 7, 1, 11. Which die is most likely in the end (the MLE)?

(b) In (optional) Problem 2 of Studio 5, the data was censored so that we only learn the data of whether a roll is 1 or not 1 (which we'll denote by 0). Suppose we only had access to the censored version of the data in (a), namely: 1, 0, 1, 0. Using the Bayesian update table format, compute the final posterior. Now which die is most likely in the end? Briefly explain why it makes sense that it is the same or different as in part (a).

(c) With the same prior, starting over, suppose that in 100 rolls of the chosen die, there are exactly 5 times that it comes up 1. Compute the final posterior. Which die is most likely in the end?

(d) Suppose instead the initial pool was 96 4-sided dice and 1 each of 6-, 8-, 12-, and 20-sided. Given the same data as in (c), compute the final posterior. Which die is most likely in the end? Briefly explain why it makes sense that it is the same or different as in part (c).