Problem 1: Bagel Burdens (40 points)

Wondering about the building you live in, you decide to try to determine how many floors it is. Every time someone orders a bagel, they give their floor number as part of the order. This gives you little glimpses into the world around you. Over the course of seven customers, you receive the following floor numbers: 70, 33, 43, 18, 121, 43, 56

Assume the floors are numbered consecutively, 1 on the first floor to some highest floor N.

1) What's the fewest number of floors the building might have? I.e., what's the smallest possible value of N, based on this data? (2 points)

   - The building must have at least 121 floors, as that is the highest floor number given in the data.

2) If there are N total floors, what is the outcome space of possible 7-floor observations? How many possible observations are there in this outcome space? What assumptions are you making here? (3 points)

   - A. The outcome space for observing 7 floors would be all possible combinations of 7 floors chosen from a total of N floors (i.e. 43, 33, 70 is the same as 70, 43, 33). The combination formula can be used to calculate the number of possible observations, which is (N choose 7). B. N!/7!(N-7)! = (N, 7) C. This assumes that floor order does not matter and that there are no repeat customers or floor numbers in the given data.

3) Assuming that there are roughly the same (large) number of people on each floor, and that people on each floor like bagels as much as people on any other floor, what was the probability of observing that specific sequence of customer floors? (5 points)

   - Assuming an equal number of people on each floor and equal preference for bagels among customers on all floors, the probability of observing a specific sequence of customer floors is (1/N)^7. Each customer's floor number is independently chosen with equal probability.

3.b) Bonus: Why is it useful to assume there is a large number of people on each floor? (2 points)

- Assuming a large number of people on each floor helps to mitigate potential biases that may arise from a small number of people on certain floors. For example, if only 5 people are on the 70th floor and 100 people are on the 33rd floor, it is more likely that a customer from the 33rd floor will order a bagel than a customer from the 70th floor. By assuming a large number of people on each floor, it can be assumed that the probability of any customer from any floor ordering a bagel is roughly the same.

4) What number of floors N would make the probability of that specific sequence of customer floors as large as possible? (5 points)

- To maximize the probability of observing the specific sequence of customer floors, the number of floors N would be 121, which is the highest floor number given in the data. In this case, the probability of observing the specific sequence would be (1/121)^7, which is the highest possible probability as the denominator is the largest possible.

5) Is the answer to Q1.4 above necessarily the true number of floors? Why or why not? (2 point)

- The data set is small: with only 7 customers, the sample size may not represent the entire building population, and the conclusions may not be accurate.
- The assumptions are not true: assuming roughly the same number of people on each floor and that people on each floor like bagels equally may not be accurate, affecting the conclusions.
- The data is not random: if the customers who ordered bagels were not a random sample, the conclusions may not be accurate.
- The building is not consecutive: if the floors are not numbered consecutively, this could impact the conclusions.

You note that if N were very, very large (very tall building), it seems unlikely that you would've observed numbers only as large as 121. If there were 1000 floors, surely someone from floor 122 to 1000 would've ordered a bagel!

6) Assume that for any n as large or larger than your answer to Q1.1, the probability that N = n is proportional to the probability of observing 70, 33, 43, 18, 121, 43, 56 if there were only n floors. (i.e., P(N = n) = c ∗ your answer to Q1.3 based on n total floors for some constant c). Find c and give the complete formula for P(N = n) for all feasible n. (10 points)

- To find c, we can set P(N = n) = c * P(observing 70, 33, 43, 18, 121, 43, 56 | N = n) equal to 1 for n = N, where N is the true number of floors.

P(N = n) = c * P(observing 70, 33, 43, 18, 121, 43, 56 | N = n) = 1

c = 1 / P(observing 70, 33, 43, 18, 121, 43, 56 | N = n)

Therefore, the complete formula for P(N = n) for all feasible n is:

P(N = n) = c * (1/n)^7 * C(n,7)

7) Using your answer to Q1.6, find the smallest integer nmax such that P(N ≤ nmax) is greater than or equal to 0.95, i.e., a number of floors such that you are at least 95% confident the true number of floors is less than nmax. (8 points)

- $P(N \leq n_{max}) = \sum P(N = n)$ for n = N to n_{max}

We can find n_{max} by increasing the value of n until $P(N \leq n_{max})$ is greater than or equal to 0.95.

After doing all this calculation to get a 95% confidence bound on how tall the building is, however, you hear that the top 20 floors are executive levels, with half the number of people on them.

8) How does that change the above calculations? Find the new nmax based on this. (5 points).

- The new max based on this information would be different because the probability of observing a floor number at or above 150 would be lower since there are fewer people on those floors. Additionally, the probability of observing lower floor numbers would also be affected since there are more people on those floors. Therefore, we would have to recalculate the probability of observing the specific sequence of customer floors and the constant c.

• Bonus 1: If you got even a single customer with a floor number at or above 150, you'd immediately know the building had at least 150 floors. What's the fewest number of customers you'd have to observe with floor numbers less than 150 in order to be at least 95% confident there were no more than 150 floors? (5 points)

- If we observe even a single customer with a floor number at or above 150, we would know that the building has at least 150 floors. To be 95% confident that there are no more than 150 floors, we would have to observe at least 8 customers with floor numbers less than 150. This is because the probability of observing at least one floor number above 150 if there were more than 150 floors would be higher than 0.05.

• Bonus 2: What assumptions in this analysis aren't necessarily accurate, and how could they be improved? (5 points)

- Assuming that the floors are numbered consecutively, starting from 1 to some highest floor N.
- Assuming that there are roughly the same (large) number of people on each floor.
- Assuming that people on each floor like bagels as much as people on any other floor.
- Assuming that the probability of observing a specific sequence of customer floors is the same for all possible values of

To improve these assumptions, we could gather more data and use a different method to estimate the number of floors. For example, we could use the data to estimate the probability of observing a floor number at or above a certain value, and use that to estimate the number of floors. Additionally, we could gather data on the number of people on each floor, and use that to estimate the number of floors. Lastly, we could gather data on the bagel preferences of people on different floors and use that to estimate the number of floors.

Problem 2: Delivery Bots! (30 points)

1) After counting the morning flock on the second day, what is the smallest number of drones N that might be currently working in MagusCorp, in total? (2 points) $80+130 = 210$
- The smallest number of drones N that might be currently working in MagusCorp, in total, is 210. This is because on the second day, there were 210 droids present to pick up bagels, and this is the total number of droids that could be working in the building.

2) What was the sample space of possible sets of 210 drones that might've been there on the second morning? How many possible outcomes is this? (3 points)
- The sample space of possible sets of 210 drones that might've been there on the second morning is all combinations of 80 tagged and 130 untagged drones, where order does not matter. This is because there were 80 droids that carried trace radiation from the isotope-tagged bagels and 130 droids that were clear, and any combination of these droids could have been present on the second morning. The number of possible outcomes is the number of ways to choose 80 droids out of 150 total, which is binomial coefficient (150 choose 80) = 3,965,375.

3) What was the probability that the morning flock on the second day contained exactly 80 tagged bots and 130 untagged bots? Note, this should be a function of N, the total number of delivery bots. (5 points)
- The probability that the morning flock on the second day contained exactly 80 tagged bots and 130 untagged bots is a function of N, the total number of delivery bots. It is given by the binomial probability mass function: $P(X=80) = (N \text{ choose } 80) * (150/N)^{80} * (N-150)/N^{80}$ where X is number of tagged droids

4) What number of drones N makes your second morning observations as likely as possible? (10 points)
- The number of drones N that makes your second morning observations as likely as possible is 210, as this is the total number of droids present on the second day.

5) Assuming again that the probability that $N = n$ is proportional to the probability of observing this specific data on the second day, find again the smallest nmax so that $P(N \leq nmax) = 0.95$. (10 points) 2 Computer Science Department - Rutgers University Spring 2023

- Assuming again that the probability that N is proportional to the probability of observing this specific data on the second day, find again the smallest nmax So that$P(N<nmax)=0.95$. We can use cumulative distribution function (CDF) to find this. So we need to find the smallest nmax such that CDF($N<nmax$)=0.95.

Bonus 3: By tagging each bagel with different amounts of isotope, you could potentially tag each droid with a uniquely identifiable amount of radiation. Is this useful? Be Thorough. (5 points)

- By tagging each bagel with different amounts of isotope, you could potentially tag each droid with a uniquely identifiable amount of radiation. This could be useful for tracking the delivery bots as they move through the building, as well as for determining which specific droids carried which bagels. It would also allow you to identify specific droids that may be malfunctioning or in need of maintenance. However, it would also require additional equipment and resources to measure and track the unique isotope levels on each droid.

Bonus 4: Why did I frame this problem with a queue of bots waiting each morning for bagels, rather than in terms of bots arriving to pick up bagels over the course of the day? (3 points)

- The problem is framed with a queue of bots waiting each morning for bagels, rather than in terms of bots arriving to pick up bagels over the course of the day, because it simplifies the problem and makes it easier to count and track the droids. By having all of the droids present at the same time in the morning, it is easier to count them and determine which ones have carried isotope-tagged bagels. Additionally, it allows for a clear and consistent measurement period, making it easier to compare data from different days.

Problem 3: More Personable (30 points)

The person you are closest to, Max, regularly stocks and resupplies you. You two have established a pleasant morning routine, where Max says 'Good Morning, BagelBot!', and you chirp in the affirmative as they access your supply tanks and reactor core. It is not clear if Max knows you've achieved sentience, but they're always glad to see you. Unfortunately, this morning Max must've been distracted, and your internal sensors aren't registering that your cream cheese reserves were filled properly. Based on what you can detect and your own internal calculations, the tank has been filled either with cream cheese or with strontium, it's impossible to say which. Not wanting Max to get in trouble, you decide to work it out yourself. You come up with the following plan: based on prior observations, you can conclude

- If a person requests cream cheese on their bagel and they get cream cheese, they reject the bagel 5% of the time. (For reasons unclear to you.)
- If a person requests cream cheese on their bagel and they get strontium, they reject the bagel only 20% of the time. (For reasons unclear to you.)

What you'll do is for every person who requests cream cheese, you'll use whatever is in your cream cheese tank, and keep track of how many people reject their bagel. If enough people who request cream cheese reject their bagel, it's reasonable to conclude that that the tank must contain strontium, otherwise you must conclude that it does in fact contain cream cheese.
Let N be the number of people who order cream cheese, and T be the threshold of rejection over which you'll conclude the tank must contain strontium.

1) Before anyone orders a bagel, what should you take the probability the tank contains cream cheese to be? What about the probability the tank contains strontium? Why? (2 points)

- Before buying a bagel, the cream cheese probability in the tank should be 1 and the strontium probability should be 0, based on the available information.

2) If the tank did contain strontium, what's the probability that T or more people out of N reject their bagel? (3 points)

- To determine the likelihood that T or more people out of N would reject their bagel if the tank did contain strontium, we can use the cumulative distribution function of a binomial distribution with parameters N and 0.2. The probability can be calculated as $P(X >= T) = 1 - P(X < T)$, where X is the total number of people who reject the bagel.

3) If the tank did contain cream cheese, what's the probability that T or more people out of N reject their bagel? (3 points)

    - If the tank contains cream cheese, we can use the cumulative distribution function of a binomial distribution with parameters N and 0.05 to determine the likelihood that T or more individuals out of N would reject their bagel. The probability can be calculated as $P(X >= T) = 1 - P(X < T)$, where X is the total number of people who reject the bagel.

4) Let X be the number of people who ordered cream cheese but ultimately reject their bagel. Express the event 'you correctly identify the contents of the tank' in terms of X, T, and the true contents of the tank. (2 points)

    - The probability of correctly identifying the contents of the tank can be stated as follows: If the tank contains strontium, $P(X >= T \mid \text{strontium}) = 1$, and if cream cheese is present in the tank, $P(X >= T \mid \text{cream cheese}) = 0$.

5) In terms of N and T, express the probability that you correctly identify the contents of the tank. (5 points)

    - To calculate the probability of correctly identifying the contents of the tank, we can use the following expression: $P(\text{correct}) = P(\text{strontium}) * P(X >= T \mid \text{strontium}) + P(\text{cream cheese}) * P(X < T \mid \text{cream cheese})$, where P(strontium) is the probability of the tank containing strontium and P(cream cheese) is the probability of the tank containing cream cheese.

6) For N = 30, what should T be to maximize the probability you correctly identify the contents of the tank? (5 points)

    - To minimize the false positive rate (rejecting a bagel while cream cheese is in the tank) and increase the likelihood of accurately detecting the contents of the tank for N=30, T should be selected such that $P(X >= T \mid \text{cream cheese})$ is kept to a minimum.

7) How many orders N should you take in order to be 95% confident that you correctly identify the contents of the tank? Assume you take the best possible T for that N. (10 points).

    - To ensure 95% confidence that the contents of the tank have been accurately identified, the number of orders N should be set such that the false positive rate is less than 0.05.

This can be done using a hypothesis testing strategy and choosing N so that the p-value is less than 0.05.