Assignment 3:

Instructions

Complete the following problem set showing your work. Problems may be worked out "by hand" or in "python" or with the assistance of other analytical software (e.g., Mathematica, MatLab). You may use chatGPT to assist in coding.

Solutions must be type written (e.g., in Jupyter, markdown, or latex). Upload PDF solution by question to crowdmark (link will be emailed to you) by **11:59pm** on the **Sunday** of the corresponding week (see syllabus). If you have issues with Crowdmark submission please email solutions to Rebekah Hall (rah11@sfu.ca).

All problems are equally weighted within an assignment. Students in 468 may or may not choose to attempt the challenge question for a bonus pts. Students in 795 are required to complete the challenge question.

Problem Set

1. **Museum Collections**

Museum collections are extraordinarily valuable in the study of ecology and evolution as they give us access to rare long-term (longitudinal) data that couldn't be collected in the 4 years of a typical PhD project. Consider the number of specimens of European Goldenrod *Solidago virgaurea* present in a herbarium (herbarium: museum for plants). These collections are done by many different researchers, often with years in between. However, given that a researcher is collecting data on goldenrods they are likely to submit more than one accession (accession: submission to a biological database or museum) at the same time. Hence we can model the accumulation of accessions of *S. virgaurea* using a compound Poisson process.

Suppose that research studies on goldenrod occur at a constant rate $\lambda = 0.61\ 1/\text{year}$ and that the number of accessions submitted per study is distributed according to a negative binomial distribution with parameters $r = 8, p = 0.75$.

Part A: How many independent research studies are expected to occur during a PhD (e.g., 4 years)? What is the expected time between research studies? How many accessions are submitted by the average research study? Plot the distribution of accessions/study. How many accessions would constitute a 'large' study?

Part B: What is the expected number of accessions over a 10-year period by all researchers?

Part C: What is the variance in the number of accessions submitted over this 10 year period?

2. **Stochastic SI Epidemic Dynamics**

Consider a simplified epidemiological model for the spread of an infectious disease in a small population. The model includes two compartments: susceptible individuals (S) and infected individuals (I). The disease spreads through a single type of interaction with a given infection rate.

Susceptible individuals become infected at a per-capita rate βI
Infected individuals recover (immediately become susceptible) at a per-capita rate γI

Part A: Write a system of ODEs describing the dynamics in this system. Solve them numerically for $S(0) = 99, I(0) = 1$ and $\beta = 0.001$ and $\gamma = 0.05$?

Part B: For this model the value $R_0 = \frac{N\beta}{\gamma}$ is the number of secondary infections that result from a single initial infection in an otherwise susceptible population. If $R_0 < 1$ then the disease is guaranteed to go extinct if $R_0 > 1$ the disease will spread in the deterministic model. What are four parameter combinations that have an R_0 value of $0.5, 0.9, 1.1$ and 1.5 respectively? Plot the dynamics for $I(t)$ in the deterministic model for each of these parameter sets.

Part C: Write a Gillespie algorithm to simulate the dynamics of a corresponding stochastic epidemic where $R_0 = 1.5$. Describe what the possible events are, their rates, and what is their effect on state space?

Part D: Simulate 50 trajectories for each of the four parameter sets you chose in B. How do the stochastic dynamics compare to the deterministic dynamics? Are you surprised by any of the results given R_0?

3. **Skip This Problem** **Bog Bodies and Ancestral DNA**
   The chemical conditions of peat bogs are ideal for the natural preservation of human bodies making them a rich source of ancient cadavers known as 'bog bodies' these bodies can range in age but are often from the Iron Age (1300 B.C.E. to 800 C.E.). Over this time period, the effective human population size is approximately 5000.

Part A: Consider a bog body that is 3000 years old. Draw the topology of a gene genealogy between yourself and this bog body.

Part B: Assuming that human generation times are 20 years. How long ago did you and the bog body share a common ancestor? Give the full distribution of times to common ancestry, the expected time, and the variance in times and make sure to note the units of time that each of these answers is measured in. What is the expected time to your common ancestor in units of years, generations, and coalescent time units?

Part C: Assuming an infinite sites model, what is the expected number of pairwise differences between your genome and that of the bog body? Assume a mutation rate of $\theta = 0.8$. How many segregating sites are there in this sample of two genomes (you and the bog body)?

Part D (Challenge Part for 795): Now consider a sample with three genomes, your genome, the 3000-year-old bog body, and a second 2000-year-old bog body. Draw the two possible genealogical topologies between the three samples. Do these two topologies occur with equal probability, if not what is the probability each occurs?
[Hint: What is the probability that you and the 2000-year-old body coalesce before 3000 years ago?]

Part E (Challenge Part for 795): What are the expected times to the common ancestor of a) You and the 2000-year-old body, b) You and the 3000-year-old body, and c) the 2000-year-old body and the 3000-year-old body

4. **Genetic diversity**

Consider the following sample of 5 genome sequences. There are several different formats in which DNA sequences can be reported. The following is in the style of a VCF "Variant Call Format" file that reports only sites in which two or more neucleotides are present.

seq:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1 :	A	T	T	A	A	C	G	G	A	G	G	G	C	G	C	G	T	G	T	A	T	T
2 :	A	A	T	A	A	C	G	G	A	G	G	G	A	G	C	G	T	G	A	A	T	T
3 :	T	T	A	A	A	C	G	G	A	G	G	A	A	C	C	G	T	G	T	G	A	C
4 :	A	T	A	G	G	T	C	G	C	C	C	G	A	G	G	A	A	G	T	A	A	T
5 :	A	T	A	G	G	T	C	A	C	C	G	G	A	G	G	A	T	A	T	A	A	T

Part A: Calculate the number of segregating sites, S, in the sample.

Part B: Calculate the number of pairwise differences between each,
between each pair of sequences. What is the average number of pairwise differences in this sample?

Part C: Calculate the observed site frequency spectrum, ξ.

Part D (Challenge Part for 795): Assuming genetic diversity evolves according to the infinite sites model, propose a hypothetical genealogy of this sample. What is the likely topology of this genealogy and what are the likely coalescent times?

5. **Time to the most recent common ancestor**

Consider a population from which you have sampled 4 haploid individuals.

Part A: Draw and label a coalescent history in which the coalescent times shown are proportional to the expected values.

Part B: Draw a $\pm 1SD$ error bar at each internal node indicating the variance in the TOTAL time to that coalescent event.

Part C: What is the distribution of times until there are exactly 2 lineages in the population, $\Pr(T_4 + T_3)$? Plot this distribution to double check your answers above.