Probability distributions and expectation practice problems

1. **Geometric distribution**. The geometric distribution represents the number of trials Y until a "success", where at each trial the probability of success is p. If we want to find the probability we will succeed after y trials, we will have $y-1$ "failures", each with probability $(1-p)$ and then a success. This gives us the probability mass function (pmf):

$$P_Y(y) = (1-p)^{y-1}p$$

   (a) Verify that the total probability over all values $y \in \{1, 2, \cdots, \infty\}$ sums to 1. *Hint: what is the sum of an infinite geometric series?*

   (b) Verity that the expected value of the geometric distribution is $E[Y] = \frac{1}{p}$. *Hint: differentiate the sum of an infinite geometric series twice.*

2. **Exponential distribution**. The continuous analog of the geometric distribution is the exponential distribution. We can think of an exponential random variable X as the "waiting time" to success without discrete trials (i.e. time it takes to wait for the bus). The probability density function (pdf) for the exponential distribution with parameter λ is:

$$P_X(x) = \lambda e^{-\lambda x}$$

   Instead of summing over all possible values x, with a continuous probability distribution we need to integrate over all possible $x \in [0, \infty)$.

   (a) Verify that the total probability over all $x \in [0, \infty)$ sums to 1.

   (b) Verity that the expected value of the exponential distribution is $E[X] = \frac{1}{\lambda}$. *Hint: use integration by parts.*

Coalescent practice problems

1. Let μ be the per base, per generation mutation rate. Given that the expected time to coalescence for two lineages is $2N$ generations, how many differences do we expect between two sequences?

2. The expected value of T_i (time when there are i lineages) is:

$$E[T_i] = \frac{1}{\binom{i}{2}} = \frac{2}{i(i-1)}.$$

   Let T_{total} be the total branch length of the coalescent genealogy (sum of all branch lengths). Making use of $E[T_i]$, what is $E[T_{\text{total}}]$ (your result can include a summation)?

3. Using $E[T_i]$ again, what is the expected value of T_{MRCA}, the time to most recent common ancestor of the entire sample? Let the sample size be n. Simplify your result so it does not include a summation.