Exercise Set #4

Do all the problems

1. (Theory and statistical inference) This problem is about the standard Kingman coalescent with infinite sites mutation. Let K_i denote the number of mutations that appear while there are exactly *i* lineages (i.e., between times T_{i+1} and T_i on the coalescent).

(a) Show that

$$P(K_i = k) = \left(\frac{i-1}{\theta + i - 1}\right)\left(\frac{\theta}{\theta + i - 1}\right)^k, k = 0,1,2, \ldots$$

The easiest way to do this is to think about the sequence of "events" (coalescence and mutations) going back in time. If the first event is a coalescence, then there would be no mutations. If the first event is a mutation, then we restart the mutation and coalescence clocks (memorylessness) and ask what the next event is. Of course, each such event has a probability. What has to happen to get exactly *k* mutations before the first coalescence when there are *i* lineages? What kind of random variable (name it) is K_i?

(b) Find an equation for $P(K_2 = k_2,\ K_3 = k_3)$, where k_2 and k_3 are nonnegative integers. What important property are you using?

(c) Since the probability in (b) is a function of θ, we can use this equation to find an estimate of θ by finding the θ that maximizes the probability for given numerical values of k_2 and k_3. Find a formula for this "maximum likelihood estimate" (MLE). [Hint: When maximizing a function, it is often easier to maximize the ln of the function, especially when the function involves products and/or powers since $\ln(AB) = \ln(A) + \ln(B)$, $\ln(A/B) = \ln(A) - \ln(B)$, and $ln(A^c) = c\ ln(A)$. When differentiating to find the max, the above trick simplifies things considerably.]

(d) If $k_2 = 1$ and $k_3 = 2$, what is your estimate of θ? Compare this to the value of Waterson's estimate for this special case.