STAT 499

Final Project

November 25, 2020

Cystic fibrosis is progressive genetic disease that causes lung infections and breathing difficulties. Individuals are afflicted with cystic fibrosis when they have 2 recessive alleles in the CFTR gene. You are a statistical geneticist interested in estimating population allele frequencies for the dominant allele A and the recessive allele a. You randomly sample 200 research subjects, twelve of which have diagnosed cystic fibrosis. Use an expectation-maximization algorithm to estimate frequencies p_A and p_a. Below are some questions to guide your analysis.

1. EM algorithms are often used when there is missing data. What data are you missing?
2. Write out the complete data likelihood.
3. Write out the incomplete data likelihood.
4. Code an EM algorithm in R.
   (a) Describe the E-step.
   (b) Describe the M-step.
   (c) At each iteration, compute the complete data likelihood. End your algorithm when the likelihood stops increasing. Keep track of how many iterations it takes to converge.
   (d) Report your estimates for the allele frequencies.

Hint 1: Use the print function to confirm that the complete data likelihood increases or stays the same at each iteration.
Hint 2: The invariance property of maximum likelihood estimators implies that $\hat{p}_a = .245$.
Hint 3: Assume initial frequencies $p_A = p_a = 1/2$. Try other initializations if time permits.