GENOME 541, Spring 2012
Problem Set #2
(Due Apr 28th 11:59pm)

---

1. **[40 points] Testing for Hardy-Weinberg Equilibrium: Chi-square Test**

   Suppose we are interested in determining whether an tri-allelic site is in Hardy-Weinberg equilibrium, the numbers of genotypes observed were shown in the table below. Let's use the chi-square, Goodness of Fit Test to make that decision.

n_{AA}	n_{AB}	n_{AC}	n_{BB}	n_{BC}	n_{CC}	n_{Total}
699	541	69	369	882	860	3420

   (a) **[4 points]** What are the genotype frequencies in the sample?

   (b) **[6 points]** What are the allele frequencies?

   (c) **[6 points]** Given the allele frequencies, what are the expected genotype frequencies assuming Hardy-Weinberg equilibrium?

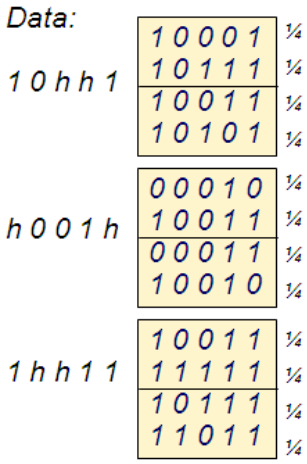
   (d) **[4 points]** Given the expected genotype frequencies, what is the expected count for each genotype?

   (e) **[10 points]** Compute the Chi-square statistics (χ^2).

   (f) **[10 points]** Suppose that you reject your null hypothesis when $\chi^2 > 5.991$, then is the population at Hardy-Weinberg equilibrium? Explain what it means to reject the null hypothesis.

2. **[60 points] Implementing the EM-based Haplotype Reconstruction**

   Let's consider the following example of a haplotype reconstruction problem. You are given the genotype data on 5 markers from 3 individuals: ({10hhh1}, {h001h}, {1hh11}). Given the initial haplotype frequencies listed below, we want to describe how each of the E-step and M-step works. We also want to implement an EM-based haplotype reconstruction algorithm.



Frequencies	
0 0 0 1 0	1/12
0 0 0 1 1	1/12
1 0 0 0 1	1/12
1 0 0 1 0	1/12
1 0 0 1 1	3/12
1 0 1 0 1	1/12
1 0 1 1 1	2/12
1 1 0 1 1	1/12
1 1 1 1 1	1/12

(a) **[8 points]** Describe what are hidden variables and what are parameters.

(b) **[8 points]** Given the haplotype frequencies listed above, describe the next E-step.

(c) **[8 points]** Write down the result of E-step that will be used in the next M-step.

(d) **[8 points]** Given the result of the E-step you described in part (b), describe the M-step.

(e) **[8 points]** Write down the result of M-step that will be used in the next E-step.

(f) **[20 points]** Based on the E-step and M-steps you described above, implement the EM-based haplotype reconstruction method. Given the genotype data ({10hhh1}, {h001h}, {1hh11}) as input, what are the final results at convergence?