GENOME 541, Spring 2012
Problem Set #2 Solution

---

1. **[20 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) **[2 points]** What are the genotype frequencies in the sample?
   Answer: The genotype frequencies are computed as follows:

$$
\begin{aligned}
p_{AA} &= \frac{n_{AA}}{n_{Total}} = \frac{699}{3420} = 0.204, \\
p_{AB} &= \frac{n_{AB}}{n_{Total}} = \frac{541}{3420} = 0.158, \\
p_{AC} &= \frac{n_{AC}}{n_{Total}} = \frac{69}{3420} = 0.020, \\
p_{BB} &= \frac{n_{BB}}{n_{Total}} = \frac{369}{3420} = 0.108, \\
p_{BC} &= \frac{n_{BC}}{n_{Total}} = \frac{882}{3420} = 0.258, \\
p_{CC} &= \frac{n_{CC}}{n_{Total}} = \frac{860}{3420} = 0.252.
\end{aligned}
$$

   (b) **[3 points]** What are the allele frequencies?
   Answer: The allele frequencies can be computed based on the genotype frequencies.

$$
\begin{aligned}
p_A &= p_{AA} + 0.5 \times p_{AB} + 0.5 \times p_{AC} = 0.204 + 0.5 \times 0.158 + 0.5 \times 0.020 = 0.2936 \\
p_B &= p_{BB} + 0.5 \times p_{AB} + 0.5 \times p_{BC} = 0.108 + 0.5 \times 0.158 + 0.5 \times 0.258 = 0.3159 \\
p_C &= p_{CC} + 0.5 \times p_{AC} + 0.5 \times p_{BC} = 0.252 + 0.5 \times 0.020 + 0.5 \times 0.258 = 0.3905
\end{aligned}
$$

   (c) **[3 points]** Given the allele frequencies, what are the expected genotype frequencies assuming Hardy-Weinberg equilibrium?
   Answer: Assuming HWE, the expected genotype frequencies are computed as follows:

$$\bar{p}_{AA} = p_A^2 = 0.2936 \times 0.2936 = 0.0862$$
$$\bar{p}_{AB} = 2p_A p_B = 2 \times 0.2936 \times 0.3159 = 0.1855$$
$$\bar{p}_{AC} = 2p_A p_C = 2 \times 0.2936 \times 0.3905 = 0.2293$$
$$\bar{p}_{BB} = p_B^2 = 0.3159 \times 0.3159 = 0.0998$$
$$\bar{p}_{BC} = 2p_B p_C = 2 \times 0.3159 \times 0.3905 = 0.2467$$
$$\bar{p}_{CC} = p_C^2 = 0.3905 \times 0.3905 = 0.1525.$$

(d) **[2 points]** Given the expected genotype frequencies, what is the expected count for each genotype?
Answer:

$$\bar{n}_{AA} = \bar{p}_{AA} \times n_{Total} = 0.0862 \times 3420 = 294.804$$
$$\bar{n}_{AB} = \bar{p}_{AB} \times n_{Total} = 0.1855 \times 3420 = 634.41$$
$$\bar{n}_{AC} = \bar{p}_{AC} \times n_{Total} = 0.2293 \times 3420 = 784.206$$
$$\bar{n}_{BB} = \bar{p}_{BB} \times n_{Total} = 0.0998 \times 3420 = 341.316$$
$$\bar{n}_{BC} = \bar{p}_{BC} \times n_{Total} = 0.2467 \times 3420 = 843.714$$
$$\bar{n}_{CC} = \bar{p}_{CC} \times n_{Total} = 0.1525 \times 3420 = 521.55$$

(e) **[5 points]** Compute the Chi-square statistics (χ^2).
Answer:

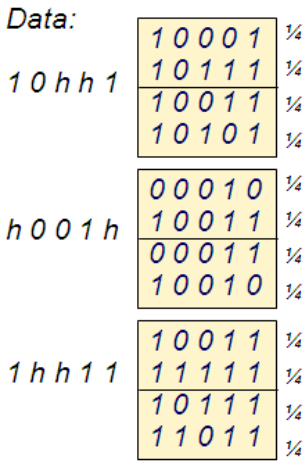
$$\begin{aligned}\chi^2 =& \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \\ =& \frac{(699 - 294.804)^2}{294.804} + \frac{(541 - 634.41)^2}{634.41} + \frac{(69 - 784.206)^2}{784.206} + \frac{(369 - 341.316)^2}{341.316} \\ &+ \frac{(882 - 843.714)^2}{843.714} + \frac{(860 - 521.55)^2}{521.55} \\ =& 554.469 + 13.751 + 652.190 + 2.237 + 1.723 + 219.701 \\ =& 1444.071\end{aligned}$$

(f) **[5 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.

Answer: The null hypothesis is that this population is at Hardy-Weinberg equilibrium. Rejecting the null hypothesis means that the HWE assumption is not correct; the expected genotype frequencies are much deviated from the actual genotype frequencies. Since $\chi^2 = 1444.071 > 5.991$, we reject the null hypothesis and this means that this population is not at HWE.

2. **[20 points] 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.
Answer: A hidden variable z_{ij} is defined for each individual i and a haplotype state j.

$$z_{ij} = \begin{array}{l} 1, \text{ if the individual } i \text{ has the haplotype pairs } j \\ 0, \text{ otherwise} \end{array}$$

For example, z_{11} is 1 if the 1st individual having the genotype {10hh1} has the halotypes {10001/10111} and 0 otherwise (if the individual's haplotypes are {10011/10101}). Thus, there are 6 hidden variables for 3 individuals: z_{11}, z_{12}, z_{21}, z_{22}, z_{31}, and z_{32}. Parameters p_i's are haplotype frequencies given to 9 haplotypes that appear in this population.

(b) **[8 points]** Given the haplotype frequencies listed above, describe the next E-step.
Answer: In the E-step, given the current parameters (haplotype frequencies), we estimate the values on the hidden variables z_{11}, z_{12}, z_{21}, z_{22}, z_{31}, and z_{32}, trying to resolve ambiguity on the haplotypes of all 3 individuals.

$$z_{11} = \frac{\text{probability that the individual 1 has haplotypes } \{10001/10111\}}{\text{probability of } \{10001/10111\} + \text{probability of } \{10011/10101\}}$$
$$z_{12} = \frac{\text{probability that the individual 1 has haplotypes } \{10011/10101\}}{\text{probability of } \{10001/10111\} + \text{probability of } \{10011/10101\}}$$
$$z_{21} = \frac{\text{probability that the individual 2 has haplotypes } \{00010/10011\}}{\text{probability of } \{00010/10011\} + \text{probability of } \{00011/10010\}}$$
$$z_{22} = \frac{\text{probability that the individual 2 has haplotypes } \{00011/10010\}}{\text{probability of } \{00010/10011\} + \text{probability of } \{00011/10010\}}$$
$$z_{31} = \frac{\text{probability that the individual 3 has haplotypes } \{10011/11111\}}{\text{probability of } \{10011/11111\} + \text{probability of } \{10111/11011\}}$$
$$z_{32} = \frac{\text{probability that the individual 3 has haplotypes } \{10111/11011\}}{\text{probability of } \{10011/11111\} + \text{probability of } \{10111/11011\}}$$

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

$$z_{11} = \frac{2 \times 1/12 \times 2/12}{2 \times 1/12 \times 2/12 + 2 \times 3/12 \times 1/12} = 0.4$$
$$z_{12} = \frac{2 \times 3/12 \times 1/12}{2 \times 1/12 \times 2/12 + 2 \times 3/12 \times 1/12} = 0.6$$
$$z_{21} = \frac{2 \times 1/12 \times 3/12}{2 \times 1/12 \times 3/12 + 2 \times 1/12 \times 1/12} = 0.75$$
$$z_{22} = \frac{2 \times 1/12 \times 1/12}{2 \times 1/12 \times 3/12 + 2 \times 1/12 \times 1/12} = 0.25$$
$$z_{31} = \frac{2 \times 3/12 \times 1/12}{2 \times 3/12 \times 1/12 + 2 \times 2/12 \times 1/12} = 0.6$$
$$z_{32} = \frac{2 \times 2/12 \times 1/12}{2 \times 3/12 \times 1/12 + 2 \times 2/12 \times 1/12} = 0.4$$

(d) **[8 points]** Given the result of the E-step you described in part (b), describe the M-step.
Answer: In the M-step, given the z_{ij}'s estimated in the previous E-step, we estimate the haplotype frequencies by partial counting.

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

$$00010: \quad p_1 = \frac{2 \times z_{21}}{12} = 0.125$$
$$00011: \quad p_2 = \frac{2 \times z_{22}}{12} = 0.0417$$
$$10001: \quad p_3 = \frac{2 \times z_{11}}{12} = 0.0667$$
$$10010: \quad p_4 = \frac{2 \times z_{22}}{12} = 0.0417$$
$$10011: \quad p_5 = \frac{2 \times (z_{12} + z_{21} + z_{31})}{12} = 0.325$$
$$10101: \quad p_6 = \frac{2 \times z_{12}}{12} = 0.1$$
$$10111: \quad p_7 = \frac{2 \times (z_{11} + z_{32})}{12} = 0.1333$$
$$11011: \quad p_8 = \frac{2 \times z_{32}}{12} = 0.0667$$
$$11111: \quad p_9 = \frac{2 \times z_{31}}{12} = 0.1$$

(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?

Answer: At convergence, the haplotype states are as follows.

Data	Haplotypes	Conditional probabilities
$10hhh1$	$10001, 10111$	0
	$10011, 10101$	1
$h001h$	$00010, 10011$	1
	$00011, 10010$	0
$1hh11$	$10011, 11111$	1
	$10111, 11011$	0

The haplotype frequencies are as follows.

$$
\begin{aligned}
00010: &\quad p_1 = 0.1667 \\
00011: &\quad p_2 = 0 \\
10001: &\quad p_3 = 0 \\
10010: &\quad p_4 = 0 \\
10011: &\quad p_5 = 0.5 \\
10101: &\quad p_6 = 0.1667 \\
10111: &\quad p_7 = 0 \\
11011: &\quad p_8 = 0 \\
11111: &\quad p_9 = 0.1667
\end{aligned}
$$