Stat 536 Homework 4

Due: 9/29/08

1. In this question, you will derive the EM algorithm for estimating haplotype frequencies when phase is unknown. Then, you will use the resulting estimates to test linkage equilibrium for a sample dataset, comparing it to the test for composite linkage disequilibrium.

   (a) Consider a diallelic system with alleles A and a at locus 1 and B and b at locus 2. The missing information in phase unknown data is that the four alleles in the n_{AaBb} individuals with genotype $AaBb$ may have be arranged as $|^A_B\ |^a_b$ or $|^A_b\ |^a_B$ and therefore, the haplotype counts n_{AB}, n_{Ab}, n_{aB}, and n_{ab} cannot be directly observed. Assuming HWE, derive the conditional probabilities $P(|^A_B \mid AaBb)$, $P(|^A_b \mid AaBb)$, $P(|^a_B \mid AaBb)$, and $P(|^a_b \mid AaBb)$ needed by the EM algorithm.

   (b) Implement the algorithm to estimate MLE haplotype frequencies for the following phase-unknown data.

n_{AABB}	n_{AaBB}	n_{AABb}	n_{AaBb}	n_{aaBB}	n_{AAbb}	n_{Aabb}	n_{aaBb}	n_{aabb}
1	8	10	32	6	6	17	10	10

   (c) Test linkage equilibrium using the EM-derived haplotype frequencies and assuming no other allele associations exist.

   (d) Why wouldn't an EM algorithm help if there is HWD or other kind of cross-chromosome association, i.e. why is the assumption of HWE required to make the EM algorithm work?

   (e) Test linkage equilibrium using the composite linkage disequilibrium coefficient Δ_{AB}, *still* disregarding trigenic and quadrigenic association. What might explain the differences between this result and the result for part 1c if any. (Note: you may compute the variance of Δ_{AB} using multiple methods, some easier than others.)

2. Consider the following outcross data from homozygous mothers and use it to estimate the outcross proportion ρ. Estimate the variance of ρ using the delta method approximation

$$\mathrm{Var}(\hat{\rho}) = \sum_{u=1}^{4} \left(\frac{\partial \rho}{\partial h_u} \right)^2 \mathrm{Var}(h_u)$$

   To get the required derivatives, recognize that MLE $\hat{\rho} = \rho(h_1, \ldots, h_4)$ is some (unknown) function of the heterozygous counts. Although $\rho(\cdot)$ is only defined through an implicit equation, we can use the chain rule to take derivatives of this expression with respect to h_u, noting that h_u appears explicitly in the formula and also inside $\rho(\cdot)$. The resulting equation can be solved for the derivatives we seek. Treat allele frequencies as known.

Mother Genotype A_uA_u	Total Offspring n_u	Heterozygous Offspring h_u	Allele Frequency p_u
A_1A_1	100	85	0.10
A_2A_2	400	256	0.34
A_3A_3	400	286	0.25
A_4A_4	100	68	0.31