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.**

   Solution:

$$\begin{aligned} P(|^A_B \mid AaBb) = P(|^a_b \mid AaBb) &= \frac{P(|^A_B\ |^a_b)}{P(AaBb)} = \frac{2p_{AB}p_{ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} \\ P(|^A_b \mid AaBb) = P(|^a_B \mid AaBb) &= \frac{P(|^a_B\ |^A_b)}{P(AaBb)} = \frac{2p_{Ab}p_{Ab}}{2p_{AB}p_{ab} + 2p_{Ab}p_{aB}} \end{aligned}$$

   (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?**

   Solution:

   Under HWE, the full likelihood reduces to a multinomial with haplotype probabilities and haplotype counts observed. This likelihood has MLEs that are simple sample proportions, i.e. very easy to maximize. With HWD, you would have to propose some model for the disequilibrium. Is it association between alleles at locus 1 (D_A), association between haplotypes, or what? Unless you want to make assumptions about the underlying population, you cannot come up with an explicit distribution. Furthermore, if you assumed a particular model, say the one with correlation f, then additional work may be needed to actually estimate the new parameter f. In this case, there is another hidden state: whether the alleles are correlated (with 2008-09-23 lecture: IBD) or not and the EM algorithm would need

to be extended to estimate both haplotype frequencies and f. There is a way, if you want to test yourself!

(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.)**

Solution:

The EM algorithm assumes no cross-chromosome associations within individuals, so if there are such associations ($D_A > 0$ or $D_B > 0$ or $D_{A/B} > 0$ or trigenic> 0 or quadrigenic> 0), the resulting haplotype frequency estimates will be wrong. Tests relying on these estimates may fail to detect LD or any of the other disequilibria. The composite disequilibrium Δ_{AB} does not make such an assumption and may be able to pick up such associations. Unfortunately, it may pick up associations other than traditional disequilibrium association D_{AB}, for example, perhaps $D_{AB} = 0$, but $D_{A/B} \neq 0$. Such a case was intended with this dataset, but I made a mistake in its analysis. It turns out, there was no evidence of any kind of association ($D_A = 0, D_b = 0, D^*_{AB} = 0, \Delta_{AB} = 0$, where the star indicates under assumption of HWE). See linked code. One bit of new information provided by this code and not given in lecture is an approximate variance formula for Δ_{AB} assuming $D_{A/B} = D_{AAB} = D_{ABB} = D_{AABB} = 0$. You were expected to get variance by numerical methods (also coded in solution).

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**

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

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	**Total Offspring**	**Heterozygous Offspring**	**Allele Frequency**
A_uA_u	n_u	h_u	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**

Solution:

Our goal is to apply the delta method, which gives

$$\text{Var}(T) \approx \sum_{i=1}^{n} \left(\frac{\partial T}{\partial \theta_i}\right)^2 \text{Var}(\theta_i) + \sum_{i=1}^{n} \sum_{j \neq i} \frac{\partial T}{\partial \theta_i} \frac{\partial T}{\partial \theta_j} \text{Cov}(\theta_i, \theta_j)$$

for an estimator T that is a function of random variables $(\theta_1, \ldots, \theta_n)$.

All we have is an implicit function that defines $\hat{\rho}$:

$$F(h, \hat{\rho}) = \sum_{u=1}^{4} \frac{(n_u - h_u)(1 - p_u)}{1 - (1 - \hat{\rho})(1 - p_u)} - \frac{h}{1 - \hat{\rho}} = 0$$

where $h = h_1 + \cdots + h_4$. Clearly, although we do not know the function, the estimator $\hat{\rho}$ is a function of the data $(h_1, \ldots, h_4)$, the number of observed mothers $(n_1, \ldots, n_4)$ and the known allele frequencies $(p_1, \ldots, p_4)$. Henceforth, we will drop the hat on ρ and treat it as a function $\rho(h_1, \ldots, h_4)$. Note, we also discard dependence on everything but the random variables; sample sizes and allele frequencies are considered known.

Implicit differentiation gives us

$$\frac{\partial \rho}{\partial h_i} = -\frac{\frac{\partial F}{\partial h_i}}{\frac{\partial F}{\partial \rho}}$$

The necessary partial derivatives are

$$\frac{\partial F}{\partial h_i}(h_1, \ldots, h_4, \rho) = -\frac{(1 - p_i)}{1 - (1 - \rho)(1 - p_i)} - \frac{1}{1 - \rho}$$

and

$$\frac{\partial F}{\partial \rho}(h_1, \ldots, h_4, \rho) = \sum_{u=1}^{4} \frac{-(n_u - h_u)(1 - p_u)^2}{[1 - (1 - \rho)(1 - p_u)]^2} - \frac{h}{(1 - \rho)^2}$$

Note, these partial derivatives are still functions of the data $(h_1, \ldots, h_4)$ and an unknown parameter ρ. When we use them in the delta formula, we plug in MLEs for parameters, i.e. $\hat{\rho}$, and data, i.e. $(1 - \hat{\rho})\,(n_1(1 - p_1), \ldots, n_4(1 - p_4))$. (Note: MLEs for data are the data that are most likely to be observed given the model and parameter MLEs, aka expected values in this case and as stated in earlier lectures.) Let's make the sustitutions now.

$$\frac{\partial F}{\partial h_i}(h_1, \ldots, h_4, \rho) = -\frac{(1 - p_i)}{1 - (1 - \hat{\rho})(1 - p_i)} - \frac{1}{1 - \hat{\rho}}$$

and

$$\frac{\partial F}{\partial \rho}(h_1, \ldots, h_4, \rho) = \sum_{u=1}^{4} \frac{-n_u[1 - (1 - \hat{\rho})(1 - p_u)](1 - p_u)^2}{[1 - (1 - \hat{\rho})(1 - p_u)]^2} - \frac{\sum_{u=1}^{4} n_u(1 - \hat{\rho})(1 - p_u)}{(1 - \hat{\rho})^2}$$

Also, to apply the delta formula, we need $\text{Var}(h_i)$ and $\text{Cov}(h_i, h_j)$ for $i \neq j$. However, the offspring are sampled independently from each mother, so $\text{Cov}(h_i, h_j) = 0$. Also, $h_i \sim \text{Bin}(n_i, (1 - \rho)(1 - p_i))$, so

$$\text{Var}(h_i) = n_i(1 - \hat{\rho})(1 - p_i)[1 - (1 - \hat{\rho})(1 - p_i)]$$

where we have again substituted in MLEs for the unknown ρ.

Now, we have all parts needed to use the delta formula, and the final calculations are in the linked code.