# Stat 536 Homework 8

Due: 11/3/08

1. [10 pts] Consider the following data sampled from two populations that have accumulated differences in allele frequencies because of genetic drift. Recently, a new pesticide was applied to both populations and the locus under study is thought to be one of the targets of the pesticide. You are charged with proving this assertion.

| | Population 1 | | | Population 2 | | |
|---|---|---|---|---|---|---|
| | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
| Generation 1 | 13 | 192 | 835 | 29 | 350 | 661 |
| Generation 2 | 12 | 158 | 570 | 62 | 237 | 441 |

   Is there evidence of selection acting on this locus? Please provide estimates and confidence values of the relative fitness values.

2. Now you will consider data collected by an ISU graduate student studying the genetics of the tsetse fly, which causes sleeping sickness, a deadly disease if not treated early. Flies were sampled from 21 different sites in the Great Rift Valley. Flies were genotyped at 8 different microsatellite loci. The original data contain many alleles at each locus, but I have reduced this dataset by taking the most prevalent allele and combining all other alleles as one "combined" allele. In other words, all loci are biallelic. The simplified data is available on the website.

   The tsetse fly habitat is highly fragmented and there is a question whether the flies migrate between these fragmented environments. We will assume there are two forces acting on this population: (1) migration homogenizing tsetse flies at sampling sites, and (2) genetic drift differentiating sites. At equilibrium, the following relationship approximately holds

$$F_{ST} = \frac{1}{1 + 4N_e m} \tag{1}$$

   for effective population size $N_e$ and migration rate $m$. In what follows, you may also assume $F_{IS} = 0$, though there may be evidence against it in the data.

   (a) [10 pts] First, we need an estimate of $F_{ST}$. Assume loci are independent, and focus here on deriving the likelihood of a single locus. Because we believe an equilibrium has been achieved, we will assume site allele frequencies $q$ are distributed according to a Beta distribution with mean $p$ and variance $\frac{p(1-p)}{1+\theta}$, specifically

$$q \sim \frac{\Gamma(\theta)}{\Gamma(p\theta)\Gamma((1-p)\theta)} q^{p\theta}(1-q)^{(1-p)\theta}$$

   where $\Gamma(x)$ is the gamma function (see `R` functions `gamma` and `lgamma`). If we observe $m_1, \ldots, m_{21}$ copies of the most prevalent allele at this locus in each population, then show (by conditioning and integrating over unobserved $q$ as for HW7) the likelihood for the $i$th population is

$$P(m_i) = \frac{n_i!}{m_i!(n_i - m_i)!} \frac{\Gamma(\theta)\Gamma(m_i + \theta p)\Gamma(n_i - m_i + \theta(1-p))}{\Gamma(n_i + \theta)\Gamma(\theta p)\Gamma(\theta(1-p))} \tag{2}$$

where $n_i$ is the total number of alleles observed in the $i$th population. (For calculations (needed in next part), see `R` functions `factorial` and `lfactorial`.) [Hint: Note $\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.]

(b) [10 pts] Using the code template provided in the website, find the maximum likelihood estimates of $\theta$ and $p$. Note, the full data includes 21 sites each with 8 loci, so the total log likelihood will be a $21 \times 8$-term sum over log likelihoods from eq. (2). Also, maximization is somewhat tricky, so be sure to try from multiple starting values to make sure the algorithm has found the true MLEs.

(c) [10 pts] Use $\hat{\theta}$ to provide an estimate of $\hat{F}_{ST}$ and then use the equilibrium condition [eq. (1)] to estimate the total number of tsetse files migrating each generation. Interpret your results. Will strategies to eliminate the tsetse fly in certain localities, but not others, eliminate the fly in the long run? If microsatellite alleles mutate (tools to answer this question will taught in Tuesday lecture, 2008-10-28), will our estimates of migration rate be over- or under-estimated?

[Warning. I have mediocre `R` skills, so take the code template with a grain of salt. It will help you find a workable solution, but I wouldn't swear by its elegance or speed.]