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.

Solution:

First, we compute x, x', y, and y', where y and y' are the allele frequency ratios for population 2.

x	x'	y	y'
0.117	0.140	0.244	0.323

These numbers satisfy two linear equations, assuming w_{11} and w_{22} are the same in both populations.

$$\begin{aligned} w_{22} &= \frac{x^2}{x'}w_{11} + \frac{x}{x'} - x \\ w_{22} &= \frac{y^2}{y'}w_{11} + \frac{y}{y'} - y \end{aligned}$$

with solutions

$$\begin{aligned} \hat{w}_{11} &= \frac{y/y' - y - x/x' + x}{x^2/x' - y^2/y'} = 2.368 \\ \hat{w}_{22} &= x^2/x'\hat{w}_{11} + x/x' - x = 0.949 \end{aligned}$$

We can sample with replacement from genotype counts in each population independently to obtain bootstrap confidence intervals.

$$\begin{aligned} \hat{w}_{11} &\in (0.081, 11.893) \\ \hat{w}_{22} &\in (0.591, 2.247) \end{aligned}$$

These numbers suggest underdominant or recessive selection favoring allele A_1 mostly, but do not exclude other possibilities. There is more compelling evidence that A_1A_1 is positively selected than A_2A_2, but even that is not certain. In fact, the truth was $w_{11} = 2.034$ and $w_{22} = 0.847$. This example illustrates how hard it is to prove the presence of selection. $w_{11} = 2$ is pretty hefty selection, but it is not possible to even prove $w_{11} > 1$ with this quite extensive data.

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*

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

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 ith 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 ith population. (For calculations (needed in next part), see `R` functions `factorial` and `lfactorial`.) [Hint: Note $\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.]

Solution:

Recall $F_{IS} = 0$, so HWE is satisfied inside each population, that is $m_i \sim Bin(n_i, q_i)$, if we condition on the allele frequency q_i in the ith population. Below, I drop the decoration on q_i.

$$\begin{aligned}
P(m_i) &= \int_0^1 P(m_i \mid q) f(q \mid p, \theta) dq \\
&= \int_0^1 \frac{n_i!}{m_i!(n_i - m_i)!} q^{m_i}(1-q)^{n_i - m_i} \frac{\Gamma(\theta)}{\Gamma(p\theta)\Gamma((1-p)\theta)} q^{p\theta-1}(1-q)^{(1-p)\theta-1} dq \\
&= \frac{n_i!}{m_i!(n_i - m_i)!} \frac{\Gamma(\theta)}{\Gamma(p\theta)\Gamma((1-p)\theta)} \int_0^1 q^{m_i + p\theta - 1}(1-q)^{n_i - m_i + (1-p)\theta - 1} dq
\end{aligned}$$

But cursory knowledge of the beta distribution (or you can just look at the hint) tells us that

$$\int_0^1 q^{m_i+p\theta-1}(1-q)^{n_i-m_i+(1-p)\theta-1}dq = \frac{\Gamma(m_i+p\theta)\Gamma(n_i-m_i+(1-p)\theta)}{\Gamma(n_i+\theta)}$$

yielding

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

(b) *[10 pts] Using the code template provided in the website, find the maximum likelihood estimates of θ and p. Note, the full data includes 21 sites each with 8 loci, so the total log likelihood will be a 21×8-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.*

Solution:

Note that θ is assumed constant over loci and $p = (p_1, \ldots, p_8)$, where p_i is the frequency of the (population-level) dominant allele at the ith locus. See the linked code. It produces estimates

$$\begin{array}{lll} \hat{\theta} = 2.831 & \hat{p}_1 = 0.459 & \hat{p}_2 = 0.329 \\ \hat{p}_3 = 0.521 & \hat{p}_4 = 0.782 & \hat{p}_5 = 0.702 \\ \hat{p}_6 = 0.730 & \hat{p}_7 = 0.693 & \hat{p}_8 = 0.254 \end{array}$$

(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?*

Solution:

Let q_i be subpopulation allele frequencies at the ith locus. Notice that the distribution of q_i has variance $\mathrm{Var}(q_i) = \frac{p_i(1-p_i)}{1+\theta}$, with θ independent of locus i. Also, recall $F_{ST} = \frac{\mathrm{Var}(q_i)}{p_i(1-p_i)}$, so F_{ST} is also independent of locus.

$$\begin{aligned} \hat{F}_{ST} &= \frac{\mathrm{Var}(q_i)}{p_i(1-p_i)} \\ &= \frac{1}{1+\hat{\theta}} \approx 0.2610384 \end{aligned}$$

Furthermore, in the allele migration model, about $2N_em$ alleles are migrating per generation. Since tsetse flies are diploid, $2N_em$ alleles translates to about N_em individuals per generation. We rearrange 1 to estimate the number of adult migrants

per generation:

$$\begin{aligned} \hat{F}_{ST} \approx 0.261 &= \frac{1}{1+4N_e m} \\ \hat{N}_e m &= \frac{1-\hat{F}_{ST}}{4\hat{F}_{ST}} \approx 0.7077134 \end{aligned}$$

flies migrating per generation. We are near the threshold for levels of migration needed to significantly stop the diversifying effects of genetic drift within populations ($4N_e m > 1$), but the question here is a little different. If pregnant females migrate, then all it takes is one female fly to restart a tsetse fly population where it has been previously eliminated through control measures. (If pregnant flies do not migrate, then a much less likely event, the migration of a male and female, followed by mating must occur, and what follows is invalid.) We expect under one fly to migrate per generation. Let's round up to 1, then 1 fly migrates per generation. The chance that a female migrates from an off-site location to the region under tsetse control is $\frac{11}{2\times 12^2} \approx 0.04$. It will take, on average, about $24 = 1/0.04$ generations to see a fly migrate from the uncontrolled regions to the controlled region. If there are no longer any measures implemented to reduce the chance of survival of this fly, then the tsetse fly population will rebound. Note, the migration rate estimated here is a rate of migration followed by survival and successful contribution to the next generation (effective migration, if you will), so we do not need to consider the chance of survival post-migration in these calculations unless control measures are ongoing. It is unclear how long it would take for the population to rebound (although tsetse are interesting in that they proliferate quite slowly), but these calculations show it will probably be necessary to reapply control measures periodically to control the flies in the long run.

If there is mutation at these loci (and mutation is probably not negligible because these are microsatellites), then we estimate $N_e(m+u) \approx 0.71$. In this case, our estimate of the number of migrants is too high, and it would be important to know how u compares to m.