Stat 536 Homework 9

due: 11/10/08

1. [10 pts] In this problem you will compare the various methods we discussed for calculating the theoretical fixation probability *in the Wright-Fisher model*. You will use three methods: (1) branching process approximation, (2) diffusion approximation with multiplicative selection, and (3) diffusion approximation for dominant selection with h set at the true value. Assume population of size $N = 1000$ and consider two selection levels: $s = 0.001$ and $s = 0.01$, both with partial dominance and $h = 0.1$. For this type of selection, method (3) should be the best, so for the other two methods, calculate the difference between that method and method (3). Plot these differences *against initial allele frequency, varying between 0 and 1*, in two plots, one for $s = 0.001$ and the second for $s = 0.01$. Discuss why the first two methods fail for some initial allele proportions.

   [Hints: See `R` functions `plot` and `lines` for plotting. See `integrate` for numerical integration, required for partial dominance selection under the diffusion approximation.]

2. The SNP Consortium (TSC) was founded to locate Single Nucleotide Polymorphisms (SNPs) and estimate allele frequencies of these SNPs in three major human populations. SNPs are loci that display an altered nucleotide in at least 1% of the human population and are biallelic. You will analyze a selection of SNP data in order to determine if certain sites along the genome are under selective pressure. Your dataset is located on the website and contains the fields:

   - `tsc_id`: unique identifier of SNP
   - `tsc_chrom`: chromosome location of SNP
   - `tsc_chrom_pos`: position along chromosome of SNP
   - `rs_id`: another identifier of some SNPs
   - `pEastAsian`: allele frequency in east Asian populations
   - `pEuropean`: allele frequency in European population
   - `pAfricanAmerican`: allele frequency in African American population

   (a) [10 pts] Assuming SNP loci are independent, there is no selection acting on any SNP locus, migration from other populations is negligible, and the same mutation rate applies to each locus, use the observed distribution of allele frequencies p in the African-American population to estimate the relative forward and reverse mutation rates u and v. [This is a somewhat silly exercise because of the implausibility of the assumptions, but it is a starting point from which you might imagine other more sophisticated approaches.]

   (b) [5 pts] For each locus, estimate one F_{ST} for each SNP locus using the Method-of-Moments estimator

$$\tilde{F}_{ST} = \frac{MSP - MSG}{MSP + (n_c - 1)MSG}$$

where

$$MSP = \frac{1}{s-1}\sum_{i=1}^{s} n_i(p_i - \bar{p})^2 \qquad MSG = \frac{1}{\sum_{i=1}^{s} n_i - 1}\sum_{i=1}^{s} n_i p_i(1-p_i)$$

n_i is the sample size in the ith population, p_i is the allele frequency in the ith population, $\bar{p}$ is the weighted sample mean allele frequency across populations, and

$$n_c = \frac{1}{s-1}\sum_{i=1}^{s} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$$

You may assume $n_i = 42$ for all three populations, all loci.

(c) [5 pts] Argue how selection at a locus can affect F_{ST}.

(d) [10 pts] A theoretical distribution for F_{ST} can be obtained by simulation or using asymptotic theory (we once discussed the chi-square distribution in this context), however all approaches assume a specific population model, whose assumptions may not apply to the three human populations under study. Instead, you will use the fact that demographic forces affect all loci equally, while selection affects only certain loci. Therefore, strongly selected loci would tend to fall as outliers in the observed distribution of F_{ST}. For three interesting genes, you will determine whether the SNP data suggests selection has acted on these genes. The genes you will consider are:

i. CARD8 has been linked to rheumatoid arthritis and Crohn's disease, both of which may confer selective disadvantages.

ii. PER3 is a gene involved in circadian rhythms and variation at this locus has been hypothesized to relate to cancer susceptibility since loss of rhythm control can lead to loss of control over cell replication.

iii. HIVEP3 is a gene encoding a protein that binds in the HIV and other virus promoter regions and thus may be under selective pressure because it can affect susceptibility of humans to viruses.

Each is linked to a list of SNPs found in these genes. Choose at least one of these SNPs that is also in your dataset to test for selection on the locus. If any of these genes appears to be under selective pressure, discuss what kind of selection may be acting on the allele. What is wrong with the p-value you estimated? (There are two main criticisms, one more subtle than the other. I only require your provide one, but as a hint to the second one, consider the definition of SNP carefully.)