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.]*

   Solution:

   The branching process approximation assumes all mutant alleles are present in independent heterozygotes, which in our model are selected with coefficient hs. The independent heterozygote assumption is valid if N is large and p is small, so that mutant alleles exist mostly in heterozygotes that don't encounter each other. The fixation probability λ for one mutant in the Wright-Fisher model is the solution of

   $$\lambda = e^{(\lambda+1)(hs+1)},$$

   which we can find using `R`'s `uniroot` function. Applying the independence assumption, the fixation probability is

   $$U_{\mathrm{BP}}(p) = 1 - \lambda^{2Np}$$

   where p is the initial allele frequency. It was legitimate to use approximation $\lambda = 1 - \frac{2s}{(1+s)^2}$, though you should state it is good because $\lambda \approx 1$ for the parameters here. It was also legitimate to use approximation $\lambda = 1 - 2s$, though you should state it is a good approximation because s is small.

   The diffusion approximation has fixation probability

   $$U_{\mathrm{D}}(p) = \frac{\int_0^p g(x)dx}{\int_0^1 g(x)dx}$$

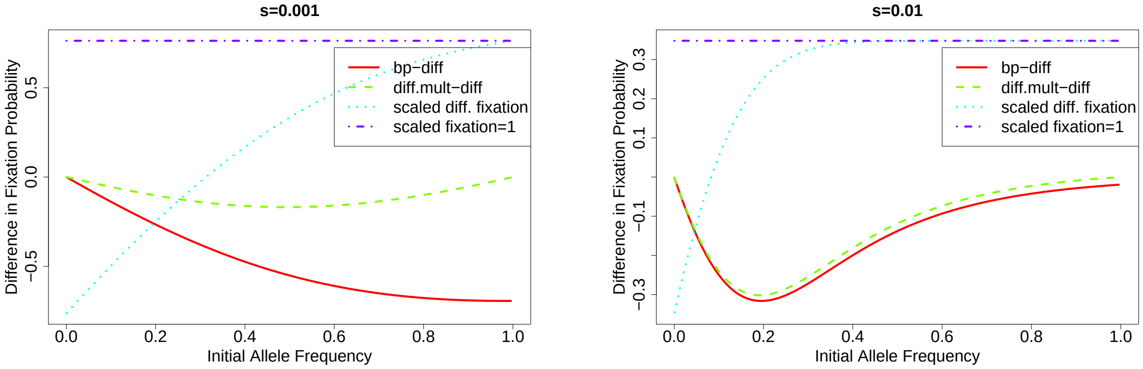
   where $g(x) = e^{-2Ns(1-2h)x^2-4Nshx}$ for partially dominant selection. The diffusion approximation is good if allele frequencies change little each generation, i.e. if N is large and s and hs are small. We can use `R`'s `integrate` function to compute the required integrals numerically.

   If multiplicative selection is assumed, i.e. $w_{\text{heterozygote}} = (1 + hs)$ and $w_{\text{homozygote}} = (1 + hs)^2$, the integrals can be compute analytically, so the fixation probability is

   $$U_{\mathrm{DM}}(p) = \frac{1 - e^{-4Nshp}}{1 - e^{-4Nsh}}$$

The above approximation is as good as $U_{\mathrm{D}}(p)$ if selection is actually multiplicative. When selection is not multiplicative, the above approximation works well if the mutant exists largely in the heterozygote, i.e. p is small.

Note, there were two challenges in deriving the above results that required you to think above and beyond the lecture notes. First, you had to extend the branching process formula to compute the fixation probability for multiple starting alleles by using the independence assumption (and remember there are $2N$ alleles in a population of size N). Second, you had to replace s in the BP and DM formulae with hs. The justification for the latter is that both approximations are valid if most (DM) or all (BP) of the selection acts on the heterozygote. Thus, the important coefficient of selection to match with the true model is that of the heterozygote. The lecture notes assumed $w_{\text{heterozygote}} = 1 + s$. Here, $w_{\text{heterozygote}} = 1 + hs$. If you used s, you were essentially modeling a different mutant with much greater selective advantage when applying the $U_{\mathrm{BP}}(p)$ and $U_{\mathrm{DM}}(p)$ approximations as compared to $U_{\mathrm{D}}(p)$. Therefore, you found the "approximations" to strongly overestimate the probability of fixation. However, you should recognize that BP approximations always *underestimate* the fixation probability unless selection is overdominant, because they fail to account for the extra selective boost awarded to homozygotes. Also, DM approximations may *underestimate* or *overestimate* fixation probabilities depending on whether the homozygous selection $(1 + hs)^2 = 1 + 2hs + h^2s^2 \approx 1 + 2hs$ is smaller or greater than the true selection $1 + s$, respectively (cusp around $h = 0.5$). For our values of s and $h < 0.5$, multiplicative-predicted selection on homozygotes was less than true homozygous selection, so fixation probabilities were underestimated.



See the linked solution that generated these plots.

Interpretation

(a) Both approximations underestimate the probability of fixation, BP because it assumes selection only acts on the heterozygote, and DM because its underestimates the true force of selection acting on homozogytes (see above discussion).

(b) As predicted by assumptions, both approximations get worse as initial allele frequency p increases, except when fixation probability is near 1 (see blue and purple lines in the plot). Because fixation probability $U(p) \leq 1$, the error necessarily declines as $U(p)$ approaches 1.

(c) The BP approximation becomes increasingly worse than the DM approximation as p increases. Neither is predicted to do well for large p, but BP suffers from two assumptions (independence and no homozygous selection), while DM only suffers from one (poor approximation to homozygous selection).

(d) The difference between the BP and DM approximations was greatest for $s = 0.001$ because with such weak selection, homozygotes play a greater role in determining whether the allele will fix.

Those of you who used s instead of hs in the branching process and multiplicative selection approximations have different conclusions. Replace interpretation 1a with: "Both approximations *overestimate* the probability of fixation because the heterozygote is given twice the advantage in the branching process and multiplicative models. This ovestimation is especially pronounced when $s = 0.01$ because fixation is largely determined before homozygotes emerge."

Those of you who assumed the initial number of mutants was Np rather than $2Np$ could not make the (correct) comparisons between BP and DM, interpretations 1c and 1d.

Points: 2 for hs instead of s, 3 for formulae, 2 for plots, 1 for 2N instead of N, 2 for discussion, +1 for using uniroot for exact extinction probability

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.]*

Solution:

When there is forward and backward mutation, then an equilibrium allele frequency is expected in infinite populations. In finite populations, an equilibrium *distibution* of allele frequencies is expected, somewhere around the infinite population equilibrium. Specifically, the following beta distribution is predicted for the proposed model

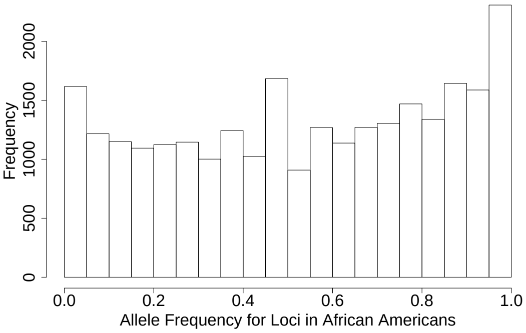
$$\phi(p) \propto p^{4N_e v-1}(1-p)^{4N_e u-1}$$

The mean and variance of the above distribution are

$$\begin{aligned} E(p) &= \frac{v}{u+v} = \frac{4N_e v}{4N_e u + 4N_e v} \\ \mathrm{Var}(p) &= \frac{E(p)[1-E(p)]}{1+4N_e u+4N_e v} \end{aligned}$$

Notice that in both equations, u and v appear multiplied by the unknown constant $4N_e$ (the first equation only identifies u and v up to a multiplicative constant).

Over time or across multiple replicate populations, the actual allele frequency p will vary because of genetic drift, but the observed p should be a draw from this Beta distribution if the model is correct. In our case, we observe p for thousands of loci in the African American population. Since we treat these loci as independent, but otherwise equivalent, it is as if we had observed p in thousands of replicate populations. (In the past, we have referred to the allele frequencies in these subpopulation as q and called what is here $E(p)$, p. Hope this is not too confusing.) The resulting empirical distribution, which looks a little U-shaped and skewed right, is shown below.



We can use method-of-moments to estimate $4N_e u$ and $4N_e v$ by substituting in the sample mean, $\bar{p} = 0.5438392$ and sample variance $s_p^2 = 0.09344654$, on the left hand side of these equations (and also on the right side of the second equation). Solving these equations yields

$$\begin{aligned} 4N_e v &= \bar{p}^2(1-\bar{p})/\sigma_p^2 - \bar{p} = 0.7548366 \\ 4N_e u &= 4N_e v(1-\bar{p})/\bar{p} = 0.8999233 \end{aligned}$$

We conclude that the forward mutation rate is just a bit larger than the backward mutation rate. Without knowing effective population size N_e, we can only speculate on the absolute magnitude of these mutation rates.

Note, the model is ridiculous because the loci are (1) not independent (some SNPs are clearly linked) and (2) not all mutating at the same rates. Further, it is not clear what is the forward vs. backward mutation rate, or in other words, what is the normal and the mutant allele.

(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

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

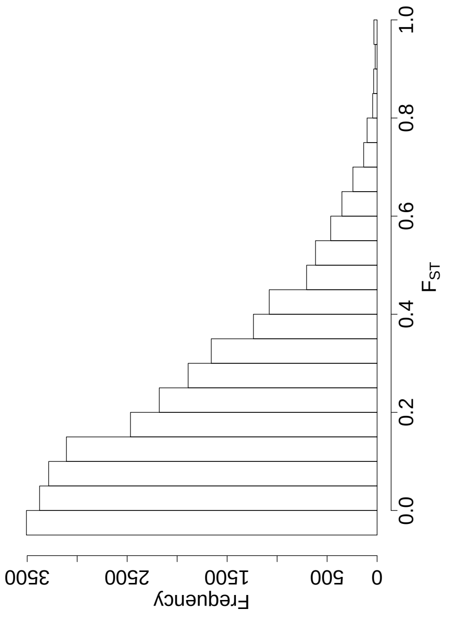
n_i is the sample size in the ith population, p_i is the allele frequency in the ith population, $\bar{p}$ is the weighted sample mean allele frequency across populations, and

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

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

Solution:

Please see the linked code, which produced the following empirical distribution of F_{ST} statistics across loci.



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

Solution:

If selection acts to maintain a particular allele in all populations, then the allele frequencies will be more similar across populations than expected and F_{ST} will be unusually small. If selection acts differently across populations, such that one allele is selected in one population, but the other allele is selected in another, then populations will be more differentiated than expected and F_{ST} will be unusually large.

(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.)

Solution:

Please see the linked code. The code checks all the SNPs identified in each of the genes and checks if the corresponding F_{ST} is in the tail of the empirical distribution. The critical quantiles of the F_{ST} distribution are $F_{ST}(\text{lower}) = -0.04274372$ and $F_{ST}(\text{upper}) = 0.6431111$. The list of F_{ST} values for each identified SNP are shown below. Those starred are significant and the p-value is calculated in the last column.

Gene	RS ID	F_{ST}	p-value
22900	1966625	0.946*	0.0015
8863	228644	-0.037	
8863	228666	-0.046*	0.0188
8863	228692	0.046	
8863	228699	0.079	
8863	228727	0.205	
8863	707463	0.032	
8863	875994	0.008	
59269	525382	0.037	
59269	616366	0.536	
59269	710231	0.209	
59269	716018	0.266	
59269	1004870	-0.041	
59269	1109256	0.005	
59269	1535505	0.659*	0.0216

Perhaps the most compelling evidence is for selection in CARD8. The only SNP in this region shows substantial evidence of differentiating selection so that populations vary more than unexpected at this locus. The other genes overlap with 7 SNPs each, but in both cases, only one of those SNPs has an unusual F_{ST} value. In the case of PER3, F_{ST} is unusually small, in fact negative. Technically, in the island model, $F_{ST} \geq 0$, so the estimated F_{ST} is not interpretable in that context. However, some kind of extreme selection, perhaps intense competition within subpopulations, could explain the negative correlation of alleles within subpopulations. Yet, the fact that only one of the 7 SNPs in PER3 and HIVEP3 is significant suggests that (1) either the selection is extremely local near the target SNP *and* the other SNPs are not linked to the selected part or (2) some assumption is violated. The latter (2) is far more likely because these SNPs are within a gene, where r is necessarily small, so all or many should probably experience selection if it happened. Furthermore, the p-values are not very small, and see critique of the p-values below.

The criticisms of the p-value are

- The p-value is invalid because the selected genes were included in the estimation of the empirical distribution. This is a minor criticism. We tested only a tiny fraction of all genes. An important point in our testing strategy is that these genes were identified for testing *before* or *independently* of their F_{ST} values. If we had decided to test them because of their F_{ST} values, we would be hard-pressed to calculate the true p-value. Obvious, but critical: A full 5% of the SNPs will have F_{ST} values that lie in the 5% tails.
- The more important reason why these p-values are suspect is because the null hypothesis only applies to a portion of all SNPs, those that happen to lie in unselected regions. What portion are unselected, we don't know, so it is not clear that the estimated empirical distribution is a good estimate of the null distribution under no selection.

- The more subtle criticism is that these SNPs are included in the database because they show variability. In fact, they were identified by taking a relatively small sample of humans from some population. If there was variability in this small sample, the SNPs were included in the bigger study. The end result is that many segregating SNPs with small allele frequencies will not have been detected as SNPs. Thus, the empirical distribution of allele frequencies p is truncated at the tails, and this truncation also impacts the F_{ST} distribution so that even under the best scenario, it does not actually represent the null distribution of unselected alleles.