Midterm

1 Short Answers (30 pts)

a. What is genetic drift? Name two consequences of genetic drift.

Solution:

Genetic drift is the random fluctuation in allele and genotype frequencies that results because populations are finite. Consequences of genetic drift are that (1) the allele frequency changes over time (and across replicate populations) (2) alleles are lost, leaving the remaining allele *fixed* in the population, (3) there is no directional change in allele frequency.

b. What are the HWE assumptions? Which of these assumptions (that we've covered in class), when violated, can lead to changes in the allele frequency?

Solution:

The assumptions of HWE are (1) no mutation, (2) no migration, (3) no selection, (4) infinite population size, (5) discrete, nonoverlapping generations, (6) no difference in genotype proportions between the sexes, and (7) random mating. Violations of (1), (2), (3), (4), and (7) can all change the allele frequencies, although finite population size (4) can not change it systematically. We had not (officially) talked about migration or selection, so you need not have mentioned them. [Note, violation of (6) does not change the overall allele frequency, although the allele frequency within the sexes will change.]

c. Provide two critiques of the effective population size estimator below. Here $\hat{p}_0, \hat{p}_2$ are the estimated allele frequencies in two generations separated by two reproduction events.

$$\hat{N}_e = \frac{1}{2\left[1 - \left(1 - \frac{\hat{p}_2 - \hat{p}_0}{\hat{p}_0(1-\hat{p}_0)}\right)^{1/2}\right]}$$

Solution:

(1) The estimator is undefined if $\hat{p}_0 \in \{0, 1\}$. (2) The estimator formula is derived assuming p_0 and p_2 are known, but in fact we must substitute estimates (MLEs in this case). In particular, $(\hat{p}_0 - \hat{p}_2)^2$ is serving as an estimator of $\mathrm{Var}(p_2)$, where the variance is caused by genetic drift. Because $\hat{p}_0 \neq p_0$ and $\hat{p}_2 \neq p_2$, the variance we actually estimate *includes* sampling variance, which is not related to N_e.

2 Hardy Weinberg Equilibrium (20 pts)

a. Consider genotype counts n_{11}, n_{12}, and n_{22} and corresponding allele counts n_1 and n_2. What are the following joint distributions under the HWE hypothesis?

Suppose p is the population frequency of allele 1. HWE means alleles are randomly assigned to genotypes, so sampling a random genotype once is the same as sampling two random alleles.

Solution:

$$\begin{aligned} P(n_{11}, n_{12}, n_{22}) &= \frac{(n_{11}+n_{12}+n_{22})!}{n_{11}!n_{12}!n_{22}!} P_{11}^{n_{11}} P_{12}^{n_{12}} P_{22}^{n_{22}} \\ &= \frac{(n_{11}+n_{12}+n_{22})!}{n_{11}!n_{12}!n_{22}!} p^{2n_{11}} [2p(1-p)]^{n_{12}} (1-p)^{2n_{22}} \end{aligned}$$

$$P(n_1, n_2) = \frac{(2n)!}{n_1! n_2!} p^{n_1} (1-p)^{n_2}$$

b. Suppose we observe $n_{11} = 10, n_{12} = 1, n_{22} = 2$. Determine the p-value for rejecting HWE based on the table of exact test probabilities shown below.

n_{11}	n_{12}	n_{22}	$P(n_{11}, n_{12}, n_{22} \mid n_1, n_2)$	**Cumul. Prob.**
10	1	2	9.1×10^{-5}	9.1×10^{-5}
9	3	1	0.35	0.35
8	5	0	0.63	0.97

Solution:

The p-value is 9.1×10^{-5} because only the observed dataset is equally or more extreme (lower probability) than the one observed.

c. Why don't we include the probability of $n_{11} = 11, n_{12} = 0, n_{22} = 2$ in the calculation of the more extreme datasets?

Solution:

The exact test is based on the conditional probability $P(n_{11}, n_{12}, n_{22} \mid n_1, n_2)$ that conditions on the observed allele counts. The configuration $n_{11} = 11, n_{12} = 0, n_{22} = 2$ changes the marginal allele counts n_1 and n_2.

3 Linkage Disequilibrium (20 pts)

a. Compute the linkage disequilibrium for the following haplotype data $n_{AB} = 36, n_{Ab} = 38, n_{aB} = 4, n_{ab} = 22$.

Solution:

$$\begin{aligned} 2n &= n_{AB} + n_{Ab} + n_{aB} + n_{ab} = 100 \\ \hat{D}_{AB} &= \hat{p}_{AB} - \hat{p}_A \hat{p}_B \\ \hat{p}_{AB} &= \frac{36}{100} = 0.36 \\ \hat{p}_A &= \frac{36+38}{100} = 0.74 \\ \hat{p}_B &= \frac{36+4}{100} = 0.40 \\ \hat{D}_{AB} &= 0.36 - 0.74 \times 0.40 = 0.064 \end{aligned}$$

b. How many generations would it take to decrease the estimated disequilibrium by half if the recombination proportion between the two loci is $r = 0.009$?

Solution:

The disequilibrium t generations after the current D_{AB} is given by

$$D_{AB}(t) = (1-r)^t D_{AB}$$

where we will substitute in $\hat{D}_{AB}$ for D_{AB}. Rearrange this equation to estimate t when $D_{AB}(t) = \frac{\hat{D}_{AB}}{2}$.

$$\begin{aligned} \frac{1}{2} &= (1-r)^t \\ -\ln(2) &= t\ln(1-r) \\ t &= \frac{-\ln(2)}{\ln(1-r)} = 76.67 \end{aligned}$$

Therefore, it will take at least 77 generations to reduce the linkage by half.

4 Nonrandom Mating (30 pts)

Here you will consider the problem of modeling and estimating the degree of assortative mating by studying a biallelic locus, assuming HWE. Let F measure the correlation of identical genotypes in mating pairs.

a. Formulate a model of assortative mating using the parameters p_A and F, i.e. write down formulas for the probabilities of the different mating types, e.g. $AA \times AA, AA \times Aa$, etc. [Hint: motivate your thinking with the inbreeding model of genotype frequencies, namely $P_{AA} = p_A^2 + f p_A p_a$ and $P_{Aa} = 2p_A p_a(1 - f)$].

Solution:

Use the exact same model except instead of alleles merging into genotypes, consider genotypes merging into mating pairs, e.g.

$$\begin{aligned} P(AA \times AA) &= P_{AA}^2 + F P_{AA}(1 - P_{AA}) \\ P(AA \times Aa) &= 2P_{AA}P_{Aa}(1 - F) \\ &\vdots \end{aligned}$$

Applying HWE to rewrite genotype frequencies in terms of allele frequencies, we get

$$\begin{array}{cc} P(AA \times AA) = p_A^4 + F p_A^2(1 - p_A^2) & P(AA \times Aa) = 4p_A^3 p_a(1 - F) \\ P(aa \times aa) = p_a^4 + F p_a^2(1 - p_a^2) & P(AA \times aa) = 2p_A^2 p_a^2(1 - F) \\ P(Aa \times Aa) = 4p_A^2 p_a^2 + 2F p_A p_a(1 - 2p_A p_a) & P(Aa \times aa) = 4p_A p_a^3(1 - F) \end{array}$$

You can verify these sum to 1.

b. If you are given $p_A = 0.8$, $F = 0.1$, work out the joint probability

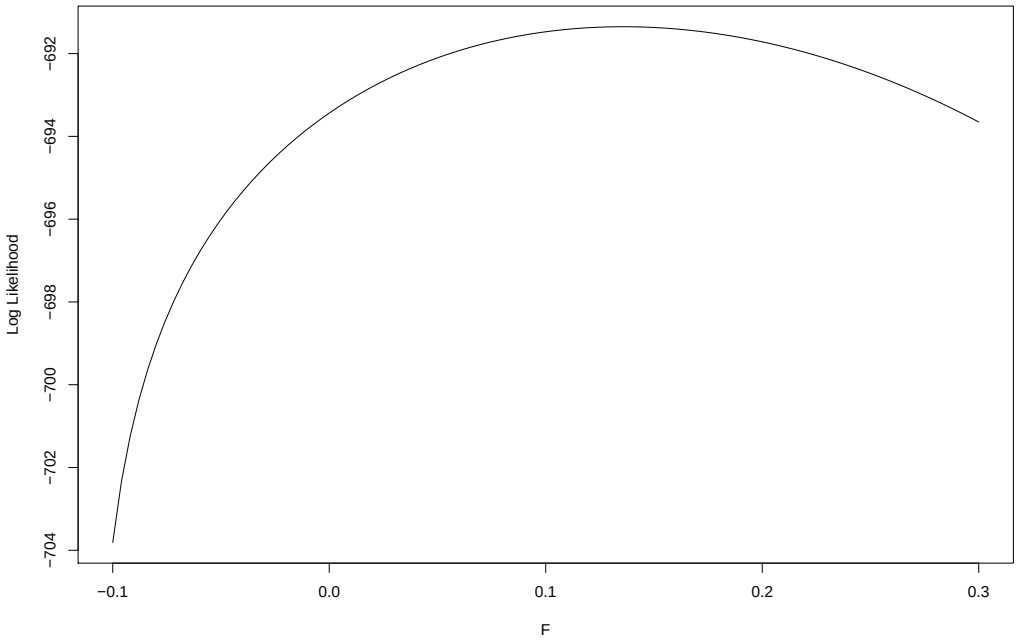
$$P(n_{AA} = 0, n_{Aa} = 1, n_{aa} = 9)$$

where (n_{AA}, n_{Aa}, n_{aa}) is the number of each offspring type for a particular mating pair.

Solution:

Let M represent the mating type. Then,

$$\begin{aligned} P(n_{AA} = 0, n_{Aa} = 1, n_{aa} = 9) &= \sum_M P(0, 1, 9 \mid M) P(M) \\ &= \frac{10!}{1!9!} \left[\left(\frac{1}{2}\right) \left(\frac{1}{4}\right)^9 P(Aa \times Aa) + \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^9 P(Aa \times aa) \right] \\ &= \frac{10!}{1!9!} \left(\frac{1}{2}\right) \left[\left(\frac{1}{4}\right)^9 \left(4(0.8)^2(0.2)^2 + 2(0.1)(0.8)(0.2)[1 - 2(0.8)(0.2)] \right) \right. \\ &\quad \left. + \left(\frac{1}{2}\right)^9 4(0.8)(0.2)^3(1 - 0.1) \right] \\ &= 0.0002273682 \end{aligned}$$



c. Now, suppose you observe many such families, which you treat as iid. You believe a the MLE of allele frequency is very close to

$$\hat{p}_A = \frac{2\sum_i n_{i,AA} + \sum_i n_{i,Aa}}{2\sum_i (n_{i,AA} + n_{i,Aa} + n_{i,aa})}$$

where $n_{i,jk}$ is the number of offspring in family i of the genotype jk. You set $p_A = \hat{p}_A$ in the likelihood and then evaluate the likelihood at many different values of F to produce the above figure. What is a good estimate of the F measure of assortative mating (by eye)? Can you reject the null hypothesis of random mating? [Note: the chi-square critical value for $\alpha = 0.05$ and 1 d.f. is 3.84 or state what critical value you need to answer the question.]

Solution:

The log likelihood is maximized around $\hat{F} = 0.14$. At this point the value of the log likelihood is $\ln L_A \approx -691.5$. To test the hypothesis of random mating, i.e. $H_0 : F = 0$, we need $\ln L_0 \approx -693.5$. The likelihood ratio statistic

$$\Lambda = -2(\ln L_0 - \ln L_A) \approx 4$$

Since H_0 is nested in our full alternative model (with F), $\Lambda \sim \chi^2_1$ asymptotically if H_0 is true. In fact, if $F = 0$, then Λ should not be very large. The one-sided hypothesis test for $H_0 : \Lambda = 0$ vs. $H_A : \Lambda > 0$ has critical value `qchisq(p=0.95,df=1)`$= 3.84$. It seems we have just enough power to reject the null hypothesis of random mating.