

Population Genetics — BI 515 — Exam 1, Spring 2014

Answer the following questions. Do your own work. Please type your answers in this document and submit electronically. All of the questions can be answered with one or a few sentences and/or numerical results.

1. Early doubters of Mendelian genetics pointed to the general lack of 3:1 phenotypic ratios in natural populations as evidence that Mendel's results on peas were not generally applicable. Why is this argument flawed?

A 3:1 phenotypic ratio is predicted for the offspring produced by a pair of heterozygous parents or for a population with an overall allele frequency of 0.5 (assuming in both cases that there are two alleles, one fully dominant, the other recessive), but this prediction does not apply generally to population-level data because allele frequencies other than 0.5 will generate different ratios of phenotypes at the population level. (E.g., suppose A is at frequency 0.2, we predict 0.04 AA genotypes plus 0.32 Aa genotypes, so a 64:36 ratio of the recessive phenotype to dominant phenotype.)

2. Provide a simple verbal explanation for why the probability of fixation for a new, neutral mutation is $1/(2N)$.

In a population of finite size, the process of coalescence ultimately leads to a single ancestral allele in the past. Assuming constant population size, that allele was one of the $2N$ alleles in an ancestral population. Moving forward in time, any of those $2N$ alleles might have been the one to survive and go to fixation, assuming no differences in fitness (selection).

3. Explain the subtle distinction between the terms autozygous and homozygous.

An individual is autozygous (and also homozygous) if the two alleles at a locus are identical by descent. The individual is homozygous (but not necessarily autozygous) if the two alleles at a locus are identical in state.

4. What factors influence effective population size in natural populations and what is the direction of their effects?

1) Variation in population size over time, 2) variation among individuals in offspring production (i.e., greater variation than expected under a Poisson process), 3) a difference in the effective number of breeding males and females, 4) differences in ploidy (e.g., sex chromosomes, mtDNA), and 5) populations structure. The first 3 reduce effective population size relative to census population size. Reduced ploidy also reduces effective population size (assuming equal variance in male and female reproductive success), whereas population structure increases effective population size (it's unlikely that subpopulations will all drift in the same direction, resulting in longer than expected maintenance of diversity in the overall population).

5. We have not yet considered models of natural selection, but you should be able to solve this problem using basic Mendelian and Hardy-Weinberg logic. Strong selection is one possible reason for a population deviating from Hardy-Weinberg equilibrium. Suppose that a gene has a dominant allele (**A**) and a recessive allele (**a**) and that survival during early life stages for individuals homozygous for the recessive allele is only 80% as high as for individuals with the dominant phenotype.

Population Genetics — BI 515 — Exam 1, Spring 2014

- a) If the population allele frequency is 0.7 A and 0.3 a in generation 1 adults, what are the expected proportions of the three genotypes in generation 2 zygotes?

$$AA = 0.49; Aa = 0.42; aa = 0.09$$

- b) What are the allele frequencies and expected proportions of the three genotypes in generation 2 adults? (Assume that a very large number of offspring is produced and that 20% of **aa** individuals dies immediately, before a random sample of the remaining individuals is selected to form the adult population for the next generation.)

20% mortality reduces aa from 0.09 to 0.072, so new frequencies are:

$$AA = 0.49/0.982 = 0.49898$$

$$Aa = 0.42/0.982 = 0.42770$$

$$aa = 0.072/0.982 = 0.07332$$

$$\text{freq } A = (2 \cdot 0.49898 + 0.41752)/2 = 0.71283$$

$$\text{freq } a = (2 \cdot 0.07332 + 0.41752)/2 = 0.28717$$

- c) Would a sample size of 1000 adults be enough to detect a significant deviation from Hardy-Weinberg equilibrium in the generation 2 adults?

Short answer: "No"

Suppose 499 AA, 428 Aa and 73 aa individuals. Observed allele frequencies are:

$$A = (2 \cdot 499 + 428)/2000 = 0.713$$

$$a = (2 \cdot 73 + 428)/2000 = 0.287$$

Expected genotype frequencies:

$$AA = 0.713^2 \times 1000 = 508.369$$

$$Aa = 2 \cdot 0.713 \cdot 0.287 \times 1000 = 409.262$$

$$aa = 0.287^2 \times 1000 = 82.369$$

(observed – expected)²/expected:

$$-9.369^2/508.369 = 0.1727$$

$$18.738^2/409.262 = 0.8579$$

$$-9.369^2/82.369 = 1.0657$$

chi-squared = 2.0963, **not significant at $p < 0.05$**

hints: critical value for the χ^2 distribution with 1 *df* and *p*-value = 0.05 is 3.84

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

6. The table below shows the genotypes for one individual at nine microsatellite loci as well as the population level allele frequencies for the allele(s) present in this individual. Using the data in this table, calculate the probability that a randomly selected individual from the same population would match the genotype show in the table at right:

a) Probability: 2.82147E-13

- b) This genotype was derived from a blood smear at a crime scene. If a suspect, who happens to be Asian, had a matching genotype, how would you argue the case if you were the prosecutor?

Using one the most basic principles of population genetics and statistics, we can calculate that the probability of an individual exactly matching the genotype at all 10 of these microsatellite loci is only

Population Genetics — BI 515 — Exam 1, Spring 2014

2.82147E-13. Put another way, there is only a 1 in 3.544 trillion chance that a randomly selected person would match this genotype by chance! We can be certain he's guilty.

c) Likewise, if you were the defense attorney?

1) We are provided no information about the population from which these allele frequencies were derived. For all we know the table provides data for a Caucasian population and therefore the calculation of any statistical probability of a match is completely invalid. For all we know, the alleles in this genotype are common Asian alleles and there may be less diversity in these loci in the Asian population, making a random match far more likely. 2) Mistakes in the lab (including sample mix-ups, cross contamination, genotyping errors). 3) Just because the defendant's blood was at the crime scene does not mean he's guilty of the crime. 4) How do we know the defendant does not have an identical twin (or other close relative, for that matter, who would have a higher probability of matching).

Locus	Allele 1	Allele 2	Allele 1: Population Frequency	Allele 2: Population Frequency
D3S1358	15	15	0.24	—
vWA	17	19	0.49	0.11
D21S11	30	32	0.02	0.23
D18S51	13	16	0.29	0.09
D13S317	11	11	0.10	—
FGA	20	24	0.39	0.05
D8S1179	13	13	0.30	—
D5S818	11	12	0.39	0.24
D7S820	9	12	0.06	0.12

7. The equation for Wright's fixation index for a random-breeding population of size N is as follows:

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_{t-1}$$

a) Explain in words the basic logic of this equation.

The equation includes two parts, the first being the probability that two randomly selected alleles are identical copies from the prior generation ($1/2N$) and the second being the probability that two alleles are identical copies from an earlier generation, which is expressed as the probability that the two alleles are **not** identical from the previous generation ($1-1/2N$) times the fixation index for the previous generation.

b) Assuming F_0 (F at time zero) = 0, what is the expected value of F after 100 generations for a population of size 250 ($2N = 500$)?

$$F_{100} = 0.181433195$$

c) The equation above assumes no mutation. Write the equation for F_t when mutation is possible.

Population Genetics — BI 515 — Exam 1, Spring 2014

$$F_t = \left(\frac{1}{2N}\right)(1-\mu)^2 + \left(1 - \frac{1}{2N}\right)(1-\mu)^2 F_{t-1}$$

d) What is the equilibrium value of F if the per locus mutation rate is 0.0001?

$$\hat{F} = \frac{1}{1+4N\mu} = 0.909090909$$

e) Explain in words why the population reaches an equilibrium value of F and stays there.

An equilibrium is reached when the rate at which diversity is lost through genetic drift is equal to the introduction of new diversity due to mutation.

8. Consider the following sample of 6 gene sequences, with the alignment below showing only the variable positions along a sequence of 1000 bases.

a) Calculate the number of segregating sites (S) and nucleotide diversity (Π = the average number of pairwise mismatches)?

$$S = 10$$

$$\Pi = (3 \times 2 \times 4 + 2 \times 5 \times 1 + 5 \times 3 \times 3)/15 = 5.267$$

b) Give two estimates of θ that can be derived from these data?

$$\theta = \Pi = 5.267$$

$$\theta = S/(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}) = 4.380$$

c) Why might the two estimates of θ not match?

The theory showing that both of these quantities estimate $\theta (= 4N\mu)$ assumes constant population size. Changes in population size affect the shape of coalescent trees and thus the patterns of nucleotide diversity observed in DNA sequences. In a growing population, for example, we expect to see relatively more segregating sites, leading to a higher value for θ_s than θ_π .

Sample1	A	A	G	C	C	T	G	T	G	T
Sample2	A	A	G	C	C	T	G	T	A	T
Sample3	A	A	G	C	T	T	G	T	A	T
Sample4	A	G	A	T	T	T	A	C	A	C
Sample5	T	A	A	T	T	C	A	C	A	C
Sample6	T	A	A	T	T	C	A	C	A	C

9. Briefly describe the Wright-Fisher model of random genetic drift.

In the Wright-Fisher model, a population of N individuals with $2N$ alleles produces a very large number of gametes (effectively infinite, equivalent to sampling with replacement), from which $2N$ alleles are randomly drawn to form the next generation of N adults. (There is no accounting of different sexes such that it is possible (with probability $1/2N$) that the two randomly selected gametes selected to form a

Population Genetics — BI 515 — Exam 1, Spring 2014

new diploid individuals are identical copies from the previous generation – this doesn't happen in organisms with separate sexes and sexual reproduction, but a proper diploid sexual model yields essentially the same results.)

- a) An classic experimental study using *Drosophila* randomly selected 8 males and 8 females to produce each successive generation, thus maintaining a constant population size of $2N = 32$. We looked at the expected and observed results for this experiment in lecture (see PowerPoints). There is, overall, a reasonably good fit between observation and theory, but drift appears to have proceeded more quickly than expected in the experimental populations. What is a likely explanation for the discrepancy?

Drift likely proceeded more quickly because the variance in reproduction among males and/or females in the experimental populations was higher than expected under a random Poisson process, thus reducing the effective population size.

10. (*Fill in the blank!*) In generations, the average time to coalescence for two randomly sampled alleles is the reciprocal of the probability of coalescence in a single generation (= $2N$ generations). Now consider the case where k alleles have been sampled. Explain in as simple terms as possible why it makes sense that the average time until 2 of the k alleles coalesce is $\frac{4N}{k(k-1)}$? Hint: It may be helpful to focus on how the probability of coalescence changes with increasing k . You may want to use both words and simple equations in your answer.

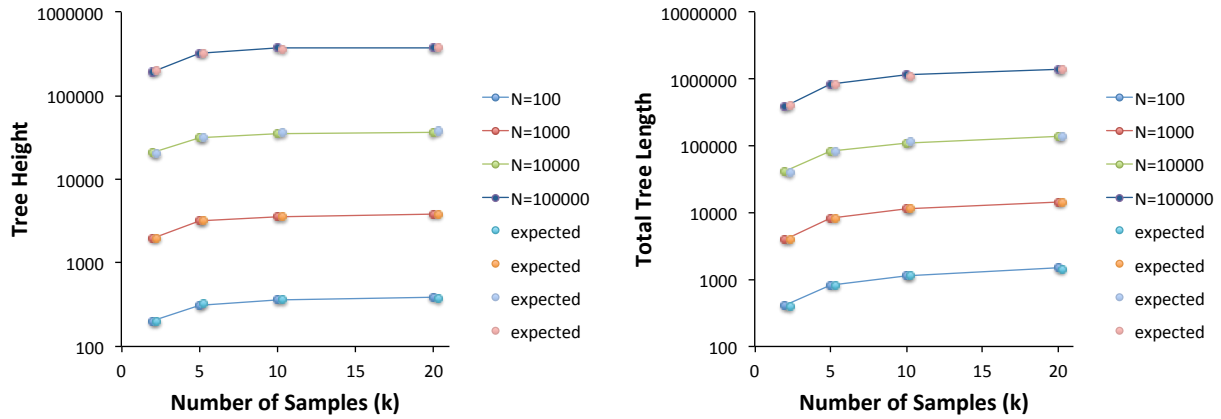
With k alleles, there are “ k choose 2” pairs of alleles that might coalesce in each preceding generation. “ k choose 2” equals $k(k-1)/2$ – given more chances for coalescence every generation, the time until the

first pair of alleles coalescence is proportionally shorter: $\frac{2N}{k(k-1)/2} = \frac{4N}{k(k-1)}$

11. Download a new version of the coalescent simulation codes from the course web page (CoalSim3.py) and test whether the code is producing results that are consistent with theoretical expectations (note that this code adds the simulation of mutations on the coalescent tree, necessary for the next question below). Specifically, how does tree height and tree length vary with k and N ? A perfect answer will be two graphs: 1) one graph showing tree height as a function of N with multiple lines showing the result for different values of k . Plot both the simulated results with standard errors and the theoretical expectations. You can calculate the mean and standard error in Excel using, for example, =AVERAGE(A1:A100) and =STDEV(A1:A100)/SQRT(100); and 2) the same for total tree length.

I should have instructed you to plot tree height and total tree length against k with different lines for N , easier to see the results that way – see next page.

Population Genetics — BI 515 — Exam 1, Spring 2014



12. Use the same python code (CoalSim3.py) to answer the following question: **If you're going to collect twice as much data to get a better estimate of θ for a given population (assuming that the population has been at constant size and affected only by drift and mutation), is it better to collect data for twice as many samples, twice as many loci, or twice the length of DNA sequence per locus?** Start the simulation with $k = 10$, $n_{loci} = 10$, $seq_len = 500$, $N = 10000$, and $\mu = 0.0000001$. Then double each parameter (k , n_{loci} , seq_len) one at a time to generate estimates of θ based on the number of segregating sites (S) and nucleotide diversity (Π). A better estimate is one that is subject to less variation, and a good way to compare variability is with the coefficient of variation. You can calculate this in Excel using, for example, $=STDEV(A1:A100)/AVERAGE(A1:A100)$. Does doubling the sample size, doubling the number of loci, or doubling the length of each locus results in the greatest improvement in the coefficient of variation of the θ estimates? Explain why this is the case. A table would work well to summarize your results.

Model	θ_S	CV	θ_π	CV
k, n_loci, seq_length				
10, 10, 500	2.0094	0.1937	2.0141	0.2242
20, 10, 500	1.9931	0.1655	1.9961	0.2139
10, 20, 500	2.0056	0.1324	2.0017	0.1520
10, 10, 1000	4.0045	0.1765	4.0028	0.2008

Any increase in the size of the data set (more samples, more loci, longer sequences) improves both estimates of theta (i.e., reduces the coefficient of variation), but increasing the number of loci has the largest effect. In other words, if you have the opportunity to double the size of your data set, your time and money is best invested in doubling the number of loci.

Adding samples only marginally increases the historical information at a given locus, because most of the coalescent history at a given locus is captured by a small number of samples; adding more sequence improves the genealogical estimate for a given locus but does not increase the number of coalescent events represented by the data; increasing the number of loci increases the number of **independent** estimates of the history, as each locus has its own coalescent history. Thus, adding more loci adds the greatest additional information about history.

Population Genetics — BI 515 — Exam 1, Spring 2014

** Note: the results you need to answer the two questions above will be saved to the file “[CoalSimResults.out](#)” – you can change the name of this file at line 12 of the code if that makes it easier to keep track of the results from different runs. You can open the output file(s) in Excel to summarize the results across runs. If you want, you can increase the number of replicate runs (line 15 of the code) to get a better estimate of the average outcomes. If you know **what** you need to do but are not sure about **how** to do it, I’m happy to give you some help with Python, etc.