Dannie Durand
durand@cs.princeton.edu

Simulating Population Genetics

Some introductory genetics: Higher organisms, including mice and men (also women), are *diploid*; that is, they have two copies of every chromosome (except the sex chromosomes, X and Y). A gene is a locus on the chromosome, a DNA sequence at a particular position that encodes a protein. Since diploid organisms have two copies of very chromosome, they also have two, possibly different, copies of every gene. The variants of a gene are called its *alleles*.

Consider a diploid organism, say a mouse, that has two possible alleles, a and A at some gene locus. When diploid organisms reproduce they make sex cells containing a single copy of each chromosome in a fifty-fifty ratio. Thus fifty percent of the sperm from an Aa mouse will contain the a allele and the other fifty percent will contain the A allele. Each parent donates one sex cell to its offspring, resulting in an embryo with two copies of each chromosome, one from each parent. So, for example, two Aa mice would have a 25% chance of producing an AA mouse, a 25% chance of producing an aa mouse, and a 50% chance of producing an Aa mouse. (Aa and aA are indistinguishable.) The combination of alleles found in a particular mouse is called its *genotype*.

If we have a population with allele frequencies $P(A) = p$ and $P(a) = (1 - p) = q$, we would like to know the probability of finding a mouse in the population with a particular genotype (AA, Aa or aa), how these probabilities vary over time and what the steady state probabilities will be. G. H. Hardy and W. Weinberg independently solved this problem in 1908 under a set of idealized assumptions:

- The population is infinite.
- All male, female pairs are equally likely to mate.
- Alleles do not spontaneously appear or disappear from the population (i.e. no migration or mutation.)
- All alleles are equally fit.

Under these conditions, the following steady-state genotype frequencies are established:

$$\begin{array}{ccc} P(AA) & P(Aa) & P(aa) \\ p^2 & 2pq & q^2 \end{array}$$

Your assignment: The first two conditions never hold; the remaining two conditions hold some of the time, at best. Your assignment is to use simulation to study the limitations of the first assumption: infinite population size.

Simulate a finite population of mice with two alleles at a single locus. Allow mice to mate randomly and assume that generations are synchronized: all the mice in generation i are offspring of mice in generation $i - 1$. You may also assume that the population size remains fixed for the duration of a single experiment. Track how the frequencies of these alleles change over time. For example, you might use histograms, means and standard deviations to study allele frequency in your population.

How do the allele frequencies vary over time? Is an equilibrium condition reached? How do you recognize when (if?) equilibrium is reached and how quickly do you reach this condition, if it occurs? Does it differ from the steady state allele frequencies predicted by Hardy and Weinberg?

How does population size affect the answers to these questions? How about the allele frequencies in the initial population? How many times should you run each simulation (i.e. each set of conditions) to have confidence in your results?

Extra credit:

- Relax the random mating condition. Partition your population into subpopulations and only allow the mice to mate within subpopulations. What are the allele frequencies in the subpopulations? In the total population? How do they differ from the Hardy-Weinberg frequencies?
- Relax the assumption of synchronized generations. What happens if some mice are allowed to survive in each generation?