Chapter 2

Genetic Drift

The discussion of random mating and the Hardy-Weinberg law in the previous chapter was premised on the population size being infinite. Sometimes real populations are very large (roughly 10^9 for our own species), in which case the infinite assumption might seem reasonable, at least as a first approximation. However, the population sizes of many species are not very large. Bird watchers will tell you, for example, that there are fewer than 100 Bachman's warblers in the cypress swamps of South Carolina. For these warblers, the infinite population size assumption of the Hardy-Weinberg law may be hard to accept. In finite populations, random changes in allele frequencies result from variation in the number of offspring between individuals and, if the species is diploid and sexual, from Mendel's law of segregation.

Genetic drift, the name given to these random changes, affects evolution in two important ways. One is as a dispersive force that removes genetic variation from populations. The rate of removal is inversely proportional to the population size, so genetic drift is a very weak dispersive force in most natural populations. The other is drift's effect on the probability of survival of new mutations, an effect that is important even in the largest of populations. In fact, we will see that the survival probability of beneficial mutations is (approximately) independent of the population size.

The dispersive aspect of genetic drift is countered by mutation, which puts variation back into populations. We will show how these two forces reach an equilibrium and how they can account for much of the molecular variation described in the previous chapter.

The neutral theory states that much of molecular variation is due to the interaction of drift and mutation. This theory, one of the great accomplishments of population genetics because it is the first fully developed theory to satisfy the Great Obsession, has remained controversial partly because it has been difficult to test and partly because of its seemingly outrageous claim that most of evolution is due to genetic drift rather than natural selection, as Darwin imagined. The theory will be developed in this chapter and will reappear in several later chapters as we master additional topics relevant to the theory.

2.1 A first look at genetic drift

Simple computer simulations, as shown in Figure 2.1, may be used to illustrate the consequences of genetic drift. These particular simulations model a population of $N = 20$ diploid individuals with two segregating alleles, A_1 and A_2. The frequency of A_1 at the start of each simulation is $p = 0.2$, which represents 8 A_1 alleles and 32 A_2 alleles. Each new generation is obtained from the previous generation by repeating the following three steps $2N = 40$ times.

1. Choose an allele at random from among the $2N$ alleles in the parent generation.
2. Make an exact copy of the allele.
3. Place the copy in the new generation.

After 40 cycles through the algorithm, a new population is created with an allele frequency that will, in general, be different from that of the original population. The reason for the difference is the randomness introduced in step 1.

As written, these steps may be simulated on a computer or with a bag of marbles of two colors, initially 8 of one color and 32 of another (providing that you have all of your marbles). The results of five independent simulations are illustrated in Figure 2.1. Obviously, allele frequencies do change at random. Nothing could be farther from the constancy promised by Hardy-Weinberg.

In natural populations, there are two main sources of randomness. One is Mendel's law of segregation. When a parent produces a gamete, each of its two homologous alleles is equally likely to appear in the gamete. The second is demographic stochasticity.* Different individuals have different numbers of offspring for complex reasons that collectively appear to be random. Neither of these sources gives any preference to particular alleles. Each of the $2N$ alleles in the parent generation has an equal chance of having a copy appear in the offspring generation.

Problem 2.1 *What is the probability that a particular allele has at least one copy in the next generation? The surprising answer quickly becomes independent of the population size as N increases. (Hint: Use one minus the probability that the allele has no copies in the next generation and Equation A.7)*

You may have noticed that the computer simulations do not explicitly incorporate either segregation or demographic stochasticity, even though these two sources of randomness are the causes of genetic drift. Nonetheless, they do represent genetic drift as conceived by most population geneticists. A more realistic simulation with both sources of randomness would behave almost exactly like our simple one. Why then, do we use the simpler simulation? The answer is a recurring one in population genetics: The simpler model is easier to understand, is easier to analyze mathematically, and captures the essence of the biological

Stochastic is a synonym for *random*.

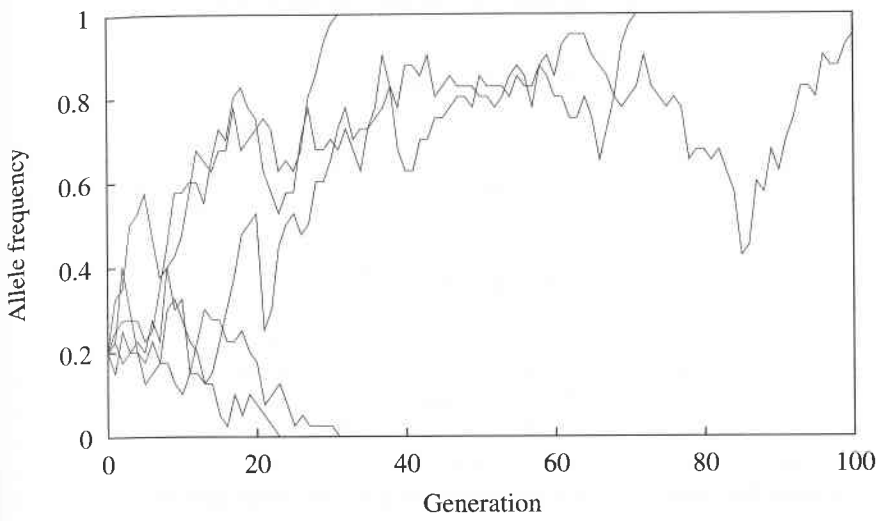

Figure 2.1: A computer simulation of genetic drift. The frequency of the A_1 allele, p, is graphed for 100 generations in five replicate populations each of size $N = 20$ and with initial allele frequency $p = 0.2$.

situation. With drift, the essence is that each allele in the parental generation is equally likely to appear in the offspring generation. In addition, the probability that a particular allele appears in the offspring generation is nearly independent of the identities of other alleles in the offspring generation. The simple algorithm in the simulation does have both of these properties and simulates what is called the Wright-Fisher model in honor of Sewall Wright and R. A. Fisher, two pioneers of population genetics who were among the first to investigate genetic drift. Most of this chapter is based on the Wright-Fisher model. Section 2.8, on the other hand, uses a more general approach that incorporates randomness in offspring numbers.

Important features of genetic drift are illustrated in Figure 2.1. One, of course, is that genetic drift causes random changes in allele frequencies. Each of the five populations behaves differently even though they all have the same initial allele frequency and the same population size. By implication, evolution can never be repeated. A second feature is that alleles are lost from the population. In two cases, the A_1 allele was lost; in two other cases, the A_2 allele was lost. In the fifth case, both alleles are still in the population after 100 generations. From this we might reasonably conclude that genetic drift removes genetic variation from populations. The third feature is more subtle: the direction of the random changes is neutral. There is no systematic tendency for the frequency of alleles to move up or down. A few simulations cannot establish this feature with certainty. That must wait for our mathematical development, beginning in the next section.

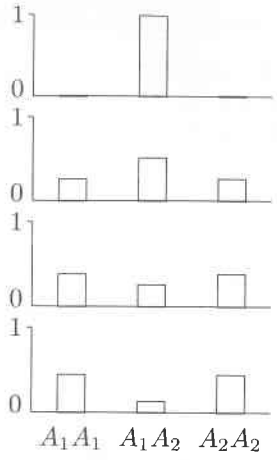

Figure 2.2: Genotype frequencies for four generations of drift with $N = 1$.

Problem 2.2 *If you know how to program a computer, write a simulation of genetic drift.*

For another view of genetic drift, consider the simplest non-trivial example: a population made up of a single hermaphroditic individual. If the individual is an A_1A_2 heterozygote, the frequency of the A_1 allele in the population is one-half. When the population reproduces by mating at random (a strange notion, but accurate) and the size of the population is kept constant at one, the heterozygote is replaced by an A_1A_1, A_1A_2, or A_2A_2 individual with probabilities 1/4, 1/2, and 1/4, respectively. These probabilities are the probabilities that the allele frequency becomes 1, 1/2, or 0 after a single round of random mating. In the first or third outcome, the population is a single homozygote individual and will remain homozygous forever. In the second outcome, the composition of the population remains unchanged.

After another round of random mating, the probability that the population is a heterozygous individual is 1/4, which is the probability that it is heterozygous in the second generation, 1/2, times the probability that it is heterozygous in the third generation given that it is heterozygous in the second generation, 1/2. The probabilities for the first four generations are illustrated in Figure 2.2. It should be clear from the figure that the probability that the population is a heterozygote after t generations of random mating is $(1/2)^t$, which approaches zero as t increases. On average, it takes only two generations for the population to become homozygous. When it does, it is as likely to be homozygous for the A_1 allele as for the A_2 allele.

If we let $\mathcal{H}_t$ be the probability that two alleles drawn at random from our population are different by state, we have $\mathcal{H}_t = (1/2)^t$. This is a special case of

an important formula to be derived in the next section:

$$\mathcal{H}_t = \mathcal{H}_0 \left(1 - \frac{1}{2N}\right)^t,$$

where $\mathcal{H}_0$ is the initial probability of being a heterozygote (one in our example) and N is the population size (also one in our example).

Problem 2.3 *Convince yourself that the average time for the population to become homozygous is, in fact, two generations.*

When discussing the computer simulations, our attention focused on the random changes of allele frequencies in a single population as illustrated by each of the five trajectories in Figure 2.1. A mathematical description of one of these trajectories is formidable as it must include all of the random twists and turns seen in the figure. Familiar functions like e^x or x^n are not up to the task. By contrast, the histograms in Figure 2.2 are easy to derive by simply multiplying probabilities. The histograms give the probability of each possible state of the population in a particular generation. The probabilities are just numbers, they are not random objects like trajectories. As such, they are much easier to describe using standard mathematical arguments. For this reason, our treatment of genetic drift will be much more akin to the histograms than to the trajectories. Motoo Kimura, an important contributor to the mathematics of genetic drift, once commented that R. A. Fisher thought of genetic drift in terms of trajectories while Sewall Wright concentrated on histograms.

2.2 The decay of heterozygosity

The mathematical description of genetic drift can be quite complicated for populations with more than one individual. Fortunately, there is a simple and elegant way to study one of the most important aspects of genetic drift: the rate of decay of heterozygosity. As usual, we will be studying an autosomal locus in a randomly mating population made up of N diploid hermaphroditic individuals. The state of the population will be described by the variable $\mathcal{G}$, defined to be the probability that two alleles different by origin (equivalently, drawn at random from the population without replacement) are identical by state. These alleles are assumed to be completely equivalent in function and, thus, equally fit in the eyes of natural selection. Such alleles are called neutral alleles. $\mathcal{G}$ is a measure of the genetic variation in the population, which is almost the same as the homozygosity of the population as defined in Equation 1.3. When there is no genetic variation, $\mathcal{G} = 1$. When every allele is different by state from every other allele, $\mathcal{G} = 0$.

The value of $\mathcal{G}$ after one round of random mating, $\mathcal{G}'$, as a function of its current value, is

$$\mathcal{G}' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\mathcal{G}. \tag{2.1}$$

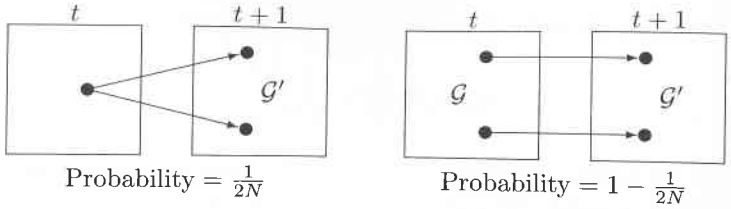

Figure 2.3: The derivation of $\mathcal{G}'$. The circles represent alleles; the arrows indicate the ancestry of the alleles.

The derivation, illustrated in Figure 2.3, goes as follows. $\mathcal{G}'$ is the probability that two alleles that are different by origin in the next generation, called generation $t+1$ in Figure 2.3, are identical by state. Identity by state could happen in two different ways. One way is when the two alleles are copies of the same allele in the previous generation, as illustrated by the left-hand side of Figure 2.3. The probability that the two alleles do share an ancestor allele in the previous generation is $1/(2N)$. (Pick one allele, and the probability that the allele picked next has the same parent allele as the first is $1/(2N)$, as all alleles are equally likely to be chosen.) The second way for the alleles to be identical by state is when the two alleles do not have the same ancestor allele in the previous generation, as illustrated by the right-hand side of Figure 2.3, but their two ancestor alleles are themselves identical by state. This ancestry occurs with probability $1 - 1/(2N)$, and the probability that the two ancestor alleles are identical by state is $\mathcal{G}$ (by definition). As these two events are independent, the probability of the second way is $[1 - 1/(2N)]\mathcal{G}$. Finally, as the two ways to be identical by state are mutually exclusive, the full probability of identity by state in the next generation is obtained by summation, as seen in the right-hand side of Equation 2.1.

The time course for $\mathcal{G}$ is most easily studied by using

$$\mathcal{H} = 1 - \mathcal{G},$$

the probability that two randomly drawn alleles are different by state. ($\mathcal{H}$ is similar to the heterozygosity of the population.) From Equation 2.1 and a few algebraic manipulations, we have

$$\mathcal{H}' = 1 - \mathcal{G}' = \left(1 - \frac{1}{2N}\right)\mathcal{H},$$

and finally,

$$\boxed{\Delta_N \mathcal{H} = -\frac{1}{2N}\mathcal{H}} \tag{2.2}$$

The Δ operator is used to indicate the change in a state variable that occurs in a single generation, $\Delta_N \mathcal{H} = \mathcal{H}' - \mathcal{H}$. The subscript N in $\Delta_N \mathcal{H}$ is simply a reminder that the change is due to genetic drift.

Equation 2.2 shows that the probability that two alleles are different by state decreases at a rate $1/(2N)$ each generation. For very large populations, this decrease will be very slow. Nonetheless, the eventual result is that all of the variation is driven from the population by genetic drift.

The full time course for $\mathcal{H}$ is

$$\mathcal{H}_t = \mathcal{H}_0 \left(1 - \frac{1}{2N}\right)^t, \tag{2.3}$$

where $\mathcal{H}_t$ is $\mathcal{H}$ in the tth generation. The easiest way to show this is to examine the first few generations,

$$\begin{aligned} \mathcal{H}_1 &= \mathcal{H}_0 \left(1 - \frac{1}{2N}\right) \\ \mathcal{H}_2 &= \mathcal{H}_1 \left(1 - \frac{1}{2N}\right) \\ &= \mathcal{H}_0 \left(1 - \frac{1}{2N}\right)^2, \end{aligned}$$

and then make a modest inductive leap to the final result.

Equation 2.3 shows that the decay of $\mathcal{H}$ is geometric. The probability that two alleles are different by state goes steadily down but does not hit zero in a finite number of generations. Nonetheless, the probability eventually becomes so small that most populations will, in fact, be homozygous. Every allele will be a descendent of a single allele in the founding population. All but one of the possibly thousands or millions of alleles in any particular population will fail to leave any descendents.

Problem 2.4 *Graph $\mathcal{H}_t$ and $\mathcal{G}_t$ for 100 generations with $\mathcal{H}_0 = 1$ and population sizes of 1, 10, 100, and 1,000,000.*

For large populations, genetic drift is a very weak evolutionary force, as may be shown by the number of generations required to reduce $\mathcal{H}$ by one-half. This number is the value of t that satisfies the equation $\mathcal{H}_t = \mathcal{H}_0/2$,

$$\frac{\mathcal{H}_0}{2} = \mathcal{H}_0 \left(1 - \frac{1}{2N}\right)^t.$$

Cancel $\mathcal{H}_0$ from both sides, take the natural logarithm of both sides, and solve for t to obtain

$$t_{1/2} = \frac{-\ln(2)}{\ln(1 - 1/2N)}. \tag{2.4}$$

The approximation of the log given in Equation A.3,

$$\ln(1 + x) \approx x,$$

allows us to write

$$t_{1/2} \approx 2N \ln(2). \tag{2.5}$$

In words, the time required for genetic drift to reduce $\mathcal{H}$ by one-half is proportional to the population size.

When studying population genetics, placing results in a more general context is often enlightening. For example, a population of one million individuals requires about 1.38×10^6 generations to reduce $\mathcal{H}$ by one-half. If the generation time of the species were 20 years, it would take about 28 million years to halve the genetic variation. In geologic terms, 28 million years ago Earth was in the Oligocene epoch, the Alps and the Himalayas were rising from the collision of India and Eurasia and large browsing mammals first appeared, along with the first monkey-like primates. During the succeeding 28 million years, whales, apes, large carnivores, and hominoids all appeared, while genetic drift was poking along removing one-half of the genetic variation.

Problem 2.5 *Graph simultaneously both Formula 2.4 and Formula 2.5 as a function of N for N from 1 to 100. Is the approximation to your liking?*

Another property of genetic drift that is easy to derive is the probability that the A_1 allele will be the sole surviving allele in the population. This probability is called the fixation probability. In Figure 2.1, the A_1 allele was fixed in two of the four replicate populations in which a fixation occurred. We could use the simulation to guess that the fixation probability of the A_1 allele is about one-half. In fact, the fixation probability is 0.2, as will emerge from a few simple observations.

As all variation is ultimately lost, we know that eventually one allele will be the ancestor of all of the alleles in the population. As there are $2N$ alleles in the population, the chance that any particular one of them is the ancestor of all (once $H = 0$) is just $1/(2N)$. If there were, say, i copies of the A_1 allele, then the chance that one of the i copies is the ancestor is $i/(2N)$. Equivalently, if the frequency of the A_1 allele is p, then the probability that all alleles are ultimately A_1 is p. In this case, we say that the A_1 allele is fixed in the population. Thus, the probability of ultimate fixation of a neutral allele is its current frequency,

$$\pi(p) = p, \tag{2.6}$$

to introduce a notation that will be used later in the book. This is as trivial as it is because all alleles are equivalent; there is no natural selection.

Problem 2.6 *$\mathcal{G}$ is almost the same as the homozygosity of the population, G. Suppose we were to define the homozygosity of a population as the probability that two alleles chosen at random from the population with replacement are identical by state. Show that this is equivalent to the definition given in Equation 1.3. Next, show that*

$$G = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)\mathcal{G}.$$

Use this to justify the claim that G and $\mathcal{G}$ are "almost the same." It should be clear that we could have used the term heterozygosity *everywhere that we used $\mathcal{H}$ without being seriously misled.*

Genetic drift appears to call into question the validity, or at least the utility, of the Hardy-Weinberg law. However, this is not the case except in the smallest of populations. The attainment of Hardy-Weinberg frequencies takes only a generation or two. Viewed as an evolutionary force, random mating has a time scale of one or two generations. Drift has a time scale of $2N$ generations, vastly larger than one or two for natural populations. When two forces have such different time scales, they rarely interact in an interesting way. This is certainly true for the interaction of drift and random mating. In any particular generation, the population will appear to be in Hardy-Weinberg equilibrium. The deviation of the frequency of a genotype from the Hardy-Weinberg expectation will be no more than about $1/(2N)$, certainly not a measurable deviation. Moreover, the allele frequency will not change by a measurable amount in a single generation. Thus, there is nothing that an experimenter could do to tell, based on allele and genotype frequencies, that the population does not adhere faithfully to the Hardy-Weinberg predictions.

2.3 Mutation and drift

If genetic drift removes variation from natural populations, why aren't all populations devoid of genetic variation? The answer, of course, is that mutation restores the genetic variation that genetic drift eliminates. The interaction between drift and mutation is particularly important for molecular population genetics and its neutral theory. The neutral theory claims that most of the DNA sequence differences between alleles within a population or between species are due to neutral mutations. The mathematical aspects of the theory will be developed in this section. The following section will bring the mathematics and the data together.

Mutation introduces variation into the population at a rate $2Nu$, where u is the mutation rate to neutral alleles. Genetic drift gets rid of variation at a rate $1/(2N)$. At equilibrium, the probability that two alleles different by origin are identical by state is given by the classic formula

$$\boxed{\hat{\mathcal{G}} = \frac{1}{1 + 4Nu}} \tag{2.7}$$

There are many ways to obtain Formula 2.7. We will use a traditional approach that follows directly from Equation 2.1 with the addition of mutation.

The probability that a mutation appears in a gamete at the locus under study is u, which is called the mutation rate even though mutation probability would be a more accurate term. When a mutation does occur, it is assumed to be to a unique allele, one that differs by state from all alleles that have ever existed in the population. A population that has been around for a long time will have seen a very large number of different alleles. Consequently, our model of mutation is often called the infinite-allele model. The infinite-allele model is meant to approximate the large, though finite, number of alleles that are possible at the molecular level.