3.1 Random Genetic Drift and Binomial Sampling 95
3.2 The Wright–Fisher Model of Random Genetic Drift 102
3.3 The Diffusion Approximation 105
An Approach Looking Forward 106
An Approach Looking Backward 110
Absorption Time and Time to Fixation 112
3.4 Random Drift in a Subdivided Population 113
3.5 Effective Population Size 121
Fluctuation in Population Size 121
Unequal Sex Ratio, Sex Chromosomes, Organelle Genes 123
Variance in Offspring Number 126
Effective Size of a Subdivided Population 127
3.6 Gene Trees and Coalescence 128
Coalescent Effective Size 133
Coalescence with Population Growth 134
Coalescent Models with Mutation 136
Applications of Coalescent Methods 137
3.7 Theoretical Implications of Coalescence 138
Coalescent Models with Recombination 140
Linkage Disequilibrium Mapping 143

CHAPTER 3

Random Genetic Drift

In every living organism, more gametes are formed than can possibly survive. This is one of the principle tenets of Darwin's theory of natural selection. Which particular gametes survive and which perish are determined partly by chance: the luck of the draw. The element of randomness implies that chance alone can change allele frequency from generation to generation. Because the sampling process does not change the allele frequencies in any predetermined way, this process is known as **random genetic drift**. The subtlety and importance of random genetic drift are the subject of this chapter.

3.1 RANDOM GENETIC DRIFT AND BINOMIAL SAMPLING

To introduce the process of random genetic drift, we first consider a large population at Hardy-Weinberg equilibrium with alleles A and a at equal frequencies $p = q = \frac{1}{2}$. In this population, the genotype frequencies are $\frac{1}{4}\,AA$, $\frac{1}{2}\,Aa$ and $\frac{1}{4}\,aa$. Suppose the population were to "crash," and that only four randomly chosen individuals survive to perpetuate the group. It is possible, by chance alone, that the survivors will consist of 4 AA individuals: this possibility has a probability of $(\frac{1}{4})^4 = \frac{1}{256}$. Similarly, it is possible that all four will be aa. Any other possible combination of genotypes could be realized, and it is not difficult to work out the probability for each combination. If the size of the new colony remains at just four individuals in each generation, this type of random sampling occurs each generation. In any reproductive cycle, there

is an opportunity for a possibly large change in gene frequency caused purely by the sampling process. One consequence of random drift soon becomes clear: Eventually the population will have either all A alleles or all a alleles. The reason is that, once the population reaches such a "fixation" state, it is stuck. Only new mutations or migrants into the population can reintroduce variation.

In the example above, we sampled four diploid individuals each generation. If mating takes place at random, sampling four diploid individuals is completely equivalent to sampling eight haploid gametes. When eight gametes are drawn at random from a population with $p = \frac{1}{2}$, there are nine possible outcomes, having 0, 1, 2, 3, ... 8 copies of the A allele and the remaining copies being the a allele. The probability of each of the nine possibilities is given by the **binomial distribution**, corresponding to successive terms in the expansion of $(\frac{1}{2}\ A + \frac{1}{2}\ a)^8$. The probability of fixation of the A allele in the next generation corresponds to the probability of drawing eight copies of the A allele. Since each successive draw is considered independent and has a chance of $\frac{1}{2}$ of yielding an A, this implies that the probability of drawing eight consecutive A alleles is $(\frac{1}{2})^8 = \frac{1}{256}$. The result is identical to the probability of drawing four AA genotypes calculated earlier, and it illustrates the principle that, with random mating, random sampling of diploid individuals is equivalent to random sampling of twice as many haploid gametes.

The process of sampling gametes from a finite population is depicted in Figure 3.1. The assumptions are the same as those that yield Hardy-Weinberg frequencies, but in this case the allele frequencies may change from generation to generation because of chance variation due to the finite population size. In the model in Figure 3.1, the reproducing adults in each generation comprise N diploid individuals. These individuals produce an essentially infinite pool of gametes in which the allele frequencies are the same as those in the adults. From this infinite pool of gametes, $2N$ are drawn and united at random to form the zygotes of the next generation. This model of the sam-

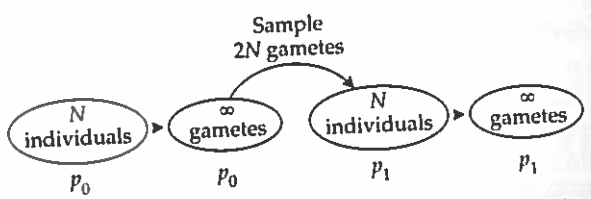

FIGURE 3.1 The gene frequencies and sampling that occur in the Wright-Fisher model. Initially there are N diploid adults with a gene whose frequency is p_0. The adults produce an infinite number of gametes having the same allele frequency. From this pool, $2N$ gametes are drawn at random to constitute the N diploid individuals for the next generation.

pling process yields a binomial distribution of all possible combinations of A and a.

To take a specific example, a population of nine diploid organisms arises from a sample of just 18 gametes, but the gametes can be thought of as being sampled from an essentially infinite pool of gametes. Because small samples are frequently not representative, an allele frequency in the sample may differ from that in the entire pool of gametes. Suppose, for example, that a pool of gametes contains the alleles A and a at frequencies p and q, respectively, with $p + q = 1$. Then if $2N$ gametes are drawn at random to produce the zygotes of the next generation, the probability that the sample contains exactly j alleles of type A is the binomial probability

$$\Pr\{j \text{ alleles of type } A\} = \binom{2N}{j} p^j q^{2N-j} = \frac{(2N)!}{j!(2N-j)!} p^j q^{2N-j} \qquad (3.1)$$

where j can take on any integer value between 0 and $2N$. The binomial coefficient (in parentheses in the middle expression) is often read as "two N choose j," because it is the number of ways that exactly j elements can be chosen from a total of $2N$. After one generation of random sampling as embodied in Equation 3.1, the new allele frequency of A in the population (call it p') is given by $j/(2N)$ because, by definition, the allele frequency of A equals the number of A alleles (in this case j) divided by the total (in this case $2N$). In the subsequent generation, the sampling process occurs anew according to Equation 3.1 with p replaced by p' and q by $1 - p'$. In this way, the allele frequency can change at random from generation to generation.

Computer-generated examples based on random sampling according to Equation 3.1 are shown in Figure 3.2. Each line in Figure 3.2A gives the number of A alleles in 20 successive generations of random genetic drift in a population of size $N = 9$ (so $2N = 18$). As you can see, individual populations behave very erratically. In seven populations, the A allele becomes **fixed** (that is, $p = 1$); in five populations, A becomes **lost** (that is, $p = 0$). In the other eight populations remain *unfixed* or **segregating** for both A and a; however, the final allele frequency among the unfixed populations is as likely to be one value as any other. Figure 3.2B shows the same kind of simulation, except now with $2N = 100$. With a larger population size, the rate at which populations go to fixation is obviously slower. The principal conclusion from Figure 3.2 is that allele frequencies behave so erratically in any one population that prediction is virtually impossible.

Although changes in allele frequency due to random genetic drift in any individual population may defy prediction, the *average* behavior of allele frequencies in a large number of populations can be predicted. Consider a large number of populations all starting at the same time with the same allele frequency and same population size N. Each of these populations is assumed to undergo drift independently of the other populations. Except for their finite

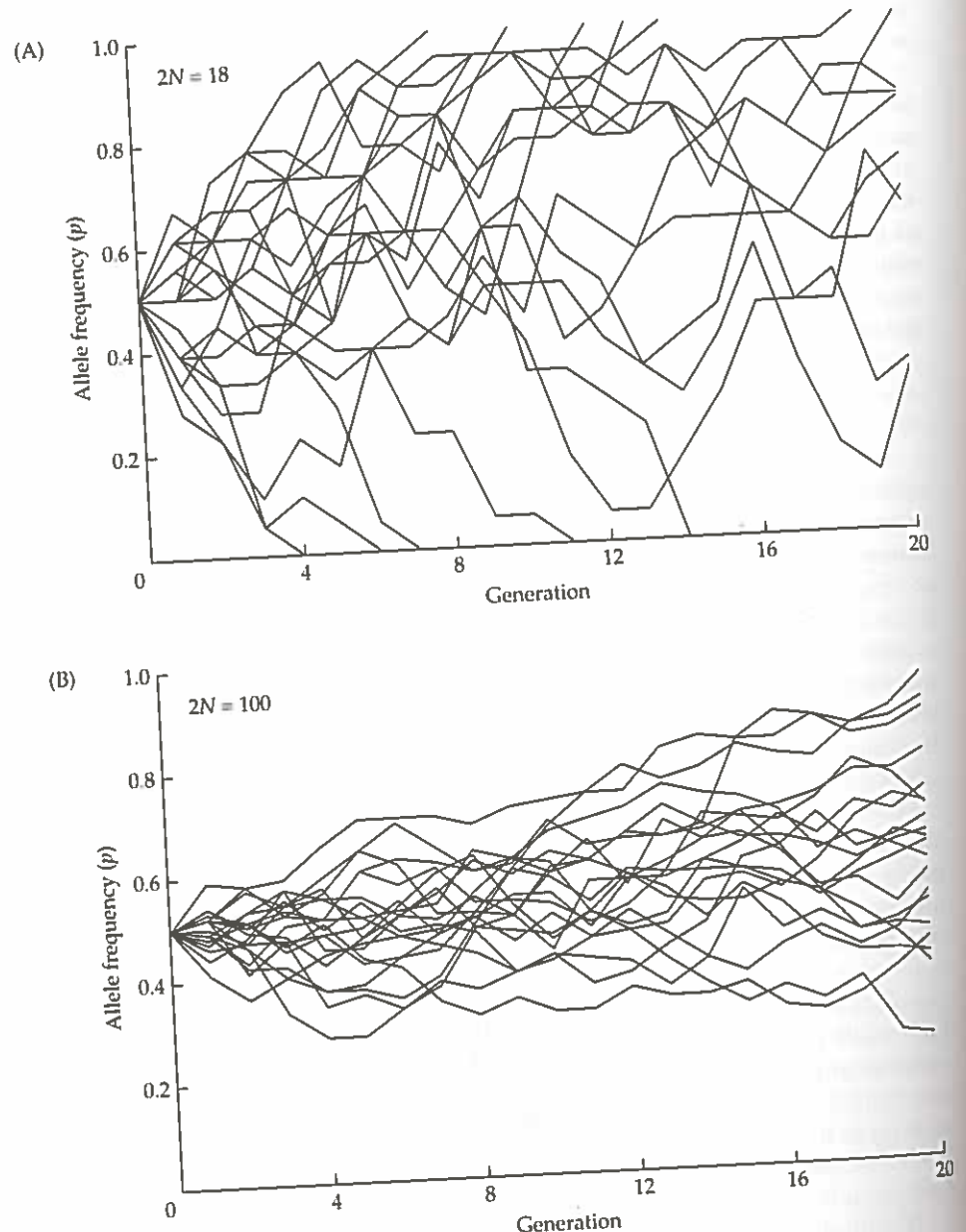

FIGURE 3.2 Computer simulations of the Wright-Fisher model of random genetic drift. Each line represents a population of size (A) $2N = 18$ or (B) $2N = 100$, simulated for 20 generations. In each generation, alleles are sampled from an infinite pool of gametes. An allele frequency of $p = 0.5$ in A implies that there are nine copies of the *A* allele, and nine copies of the *a* allele. In B, an allele frequency of 0.5 implies 50 copies of each allele. Note that the larger population size in B results in smaller fluctuations in allele frequency and a slower rate of fixation.

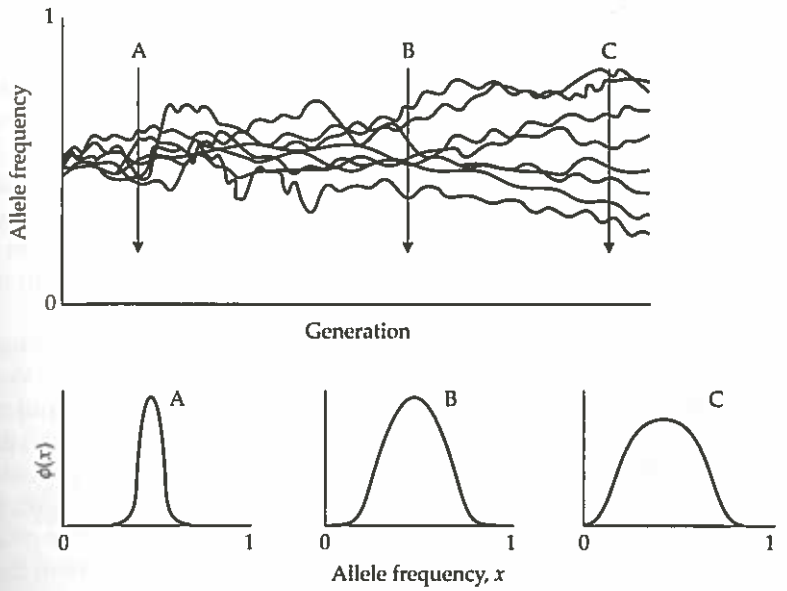

FIGURE 3.3 The implications of random genetic drift can be appreciated by imagining a large collection of subpopulations undergoing the process of repeated sampling. As the top part of the figure indicates, the allele frequency in each subpopulation changes erratically, and the allele frequencies in different subpopulations tend to drift apart. At time intervals, a snapshot of the subpopulations would produce a distribution of allele frequencies, the variance of which increases over time.

size, the subpopulations are assumed to satisfy all the assumptions of the Hardy-Weinberg model, with the additional stipulations that (1) the number of males and females is equal, and (2) each individual has an equal chance of contributing successful gametes to the next generation. The key point—illustrated in Figure 3.3—is that we can describe how these populations change in allele frequency by considering time slices through the graph and tallying a histogram of the counts of populations having each specified allele frequency. Initially, the populations will all be close to the starting allele frequency. As time passes, the populations "drift" apart, and eventually they become spread over all possible allele frequencies. Finally, as we will see, each population must go to fixation for one allele or the other.

To appreciate why ultimate fixation or loss is inevitable, consider an infinitely long bowling alley with minor imperfections that displace an imaginary weightless bowling ball one way or the other. Since the ball has no mass, it has no momentum, and hence at every instant is subject willy-nilly to the bumps and shallows of the alley. This means that, like allele frequencies, the future of the bowling ball depends only on its current position, not

on how it got there. The gutters represent the fixation states of $p = 0$ and $p = 1$. Once the ball goes into the gutter, it cannot get out again. The imperfections keep the ball from rolling in a straight line, and eventually it rolls into one or the other gutter. In this analogy, the size of the population corresponds to the width of the bowling alley; a larger population implies a wider alley. The imperfections still deflect the ball but, in proportion to the width of the alley, the ball's zigs and zags are of a smaller magnitude. Consequently, the ball remains out of the gutter for a longer time, analogous to the longer time to fixation for a larger population, but eventually the ball will end up in the gutter.

For a full understanding of random genetic drift, we must learn how to deduce the distributions of allele frequencies plotted in Figure 3.3. We just described what would happen after one generation—the set of populations would have a range of allele frequencies as described by the binomial distribution in Equation 3.1. The binomial distribution gives us the probability that a population has allele frequency p' after one generation of drift. If we consider 1000 populations all starting at p, the binomial distribution gives us the fraction of those populations with allele frequency p'. What about the following generation? For each population, one can imagine the whole sampling process as starting over again. Because no population remembers where it was the previous generation, the binomial sampling occurs anew in each generation. But because the allele frequency changes, the new allele frequency must be used in Equation 3.1. For 1000 populations, Equation 3.1 would have to be applied to each one individually, and then summed across these distributions to obtain the overall probability of each possible outcome of random drift. Fortunately, there is an easier approach that is described after we examine the following experiment.

An actual experiment designed along the lines of Figure 3.3 yielded the results shown in Figure 3.4. The graph shows the history of 19 generations of random genetic drift in 107 subpopulations of *Drosophila melanogaster*. Each subpopulation was initiated with 16 heterozygous bw^{75}/bw flies (bw = brown eyes) and maintained at a constant size of 16 individuals by randomly choosing eight males and eight females to produce the next generation. Each histogram in Figure 3.4 gives the number of subpopulations containing 0, 1, 2, ... 32 bw^{75} alleles. The pattern of change in allele frequency in Figure 3.4 may at first appear to be complicated, but in reality a simple thing is happening. The initially humped distribution of allele frequency gradually becomes flat as populations fixed for bw^{75} or bw begin to pile up at the boundaries. The piling up occurs because, once an allele has been fixed or lost, it remains fixed or lost since mutation is negligible over such a small number of generations in small populations. After 19 generations, most of the subpopulations are fixed for one allele or the other, and among the unfixed populations, the distribution of allele frequencies is essentially flat.

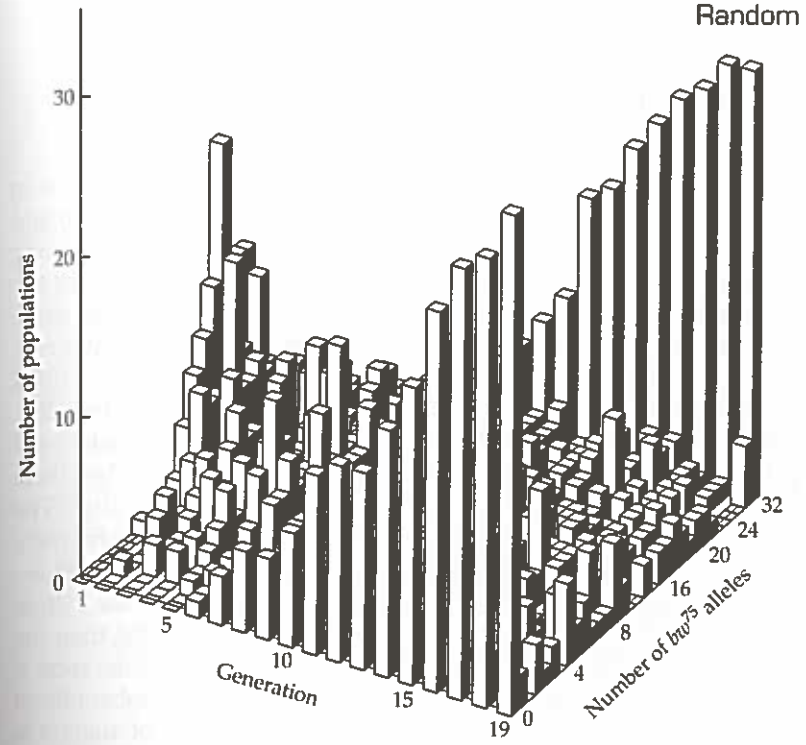

FIGURE 3.4 Random genetic drift in 107 actual populations of *Drosophila melanogaster*. Each of the initial 107 populations consisted of 16 bw^{75}/bw heterozygotes ($N = 16$; bw = brown eyes). From among the progeny in each generation, eight males and eight females were chosen at random to be the parents of the next generation. The horizontal axis of each curve gives the number of bw^{75} alleles in the population, and the vertical axis gives the corresponding number of populations. (Data from Buri 1956.)

PROBLEM 3.1 Consider a self-pollinating plant population consisting of a single heterozygous (*Aa*) individual on a small barren island. Suppose the plant reproduces and dies, so that the generations are discrete and the population can only consist of a single plant. What is the probability that the population is homozygous at this genetic locus by the second generation?

ANSWER The chance that the first generation offspring is *AA* is $\frac{1}{4}$ and the chance that it is *aa* is also $\frac{1}{4}$, so the chance of fixation in one generation is $\frac{1}{2}$. If the first generation offspring is *Aa*, then the probability of fixation in the second generation (given that the population is not fixed in the first generation) is again $\frac{1}{2}$. The probability of not fixing in generation 1 and then fixing in generation 2 is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Add to this the chance of fixing in one generation ($\frac{1}{2}$), and we get $\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$ as the probability of fixation by two generations. Note that the probability of *not* going to fixation each generation is $\frac{1}{2}$, and so the chance of not fixing for two generations is $\frac{1}{2} \times \frac{1}{2}$, which equals $1 - \frac{3}{4}$.

3.2 THE WRIGHT-FISHER MODEL OF RANDOM GENETIC DRIFT

The model of random genetic drift with binomial sampling described in Equation 3.1 is known as the **Wright-Fisher model** because Fisher (1930) and Wright (1931) derived the expected distribution of allele frequencies among subpopulations. Although neither author formulated the problem in the manner used here, our approach makes the problem much simpler and gives the same results. If a population contains $2N$ alleles among which two alleles A and a may be present, then the state of the population can be described by the number of A alleles in the population. The possible states are then 0, 1, 2,... $2N$. The states 0 and $2N$ are special in that these are fixation states, and once the population get into either of these states, it cannot leave unless there is a new mutation (and for the moment we exclude this possibility). The states 0 and $2N$ are called *absorbing states*. From any nonfixed allele frequency, it is possible for the population to drift to any other allele frequency. However, the population is more likely to remain close to its present state than to take a large jump. To use an example from Figure 3.4, if $2N = 32$, then the chance of drifting from 30 copies of gene A to 29 copies in one generation is 0.186, whereas the chance of drifting to 27 copies is 0.033. The probability of the population drifting from a state having i copies to j copies of allele A is known as the *transition probability*. The transition probability for the Wright-Fisher model is obtained directly from the binomial distribution (see Equation 3.1). In particular, if a population has i copies of allele A and $2N - i$ copies of allele a, then the transition probability, T_{ij}, of going from i copies of A to j copies of A in one generation of random genetic drift is given by

$$T_{ij} = \binom{2N}{j}\left(\frac{i}{2N}\right)^{j}\left(\frac{2N-i}{2N}\right)^{2N-j} = \frac{(2N)!}{j!(2N-j)!}p^{j}q^{2N-j} \qquad (3.2)$$

where $p = i/2N$ is the initial allele frequency of A and $q = (2N - i)/2N$ is the initial allele frequency of a.

The transition probabilities can be put in a square matrix **T**, with elements T_{ij} giving the transition probability from state i to state j for $i, j = 0,1,2,\ldots 2N$. The matrix **T** contains everything that is needed to predict the expected distribution of populations like those in Figure 3.4 over a series of generations. This type of model, expressed in terms of discrete states with fixed probabilities of going from one state to another, is known as a **Markov chain**, and it has some very elegant mathematical properties. Iterations of the Wright-Fisher model give the expected outcome of a pure drift process (Figure 3.5). In a few minutes, we will use the Wright-Fisher model to show an important result regarding fixation probabilities.

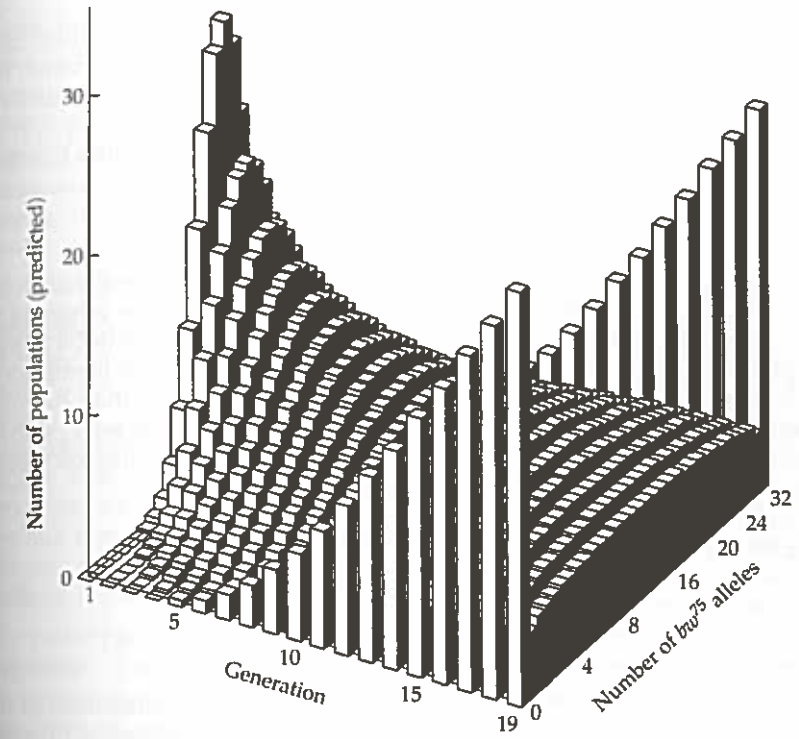

FIGURE 3.5 Prediction of the Wright-Fisher model for the distribution of allele frequencies $\phi(p, x; t)$ in subpopulations of size $N = 16$, where x represents the allele frequency in generation t. Time runs for 19 generations, and all subpopulations start with an initial allele frequency of $p = 0.5$. The values of $\phi(p, x; t)$ were generated by successive multiplication of the Markov transition probability matrix, whose entries are given by the binomial distribution in Equation 3.2. The model with $2N = 32$ predicts that fewer populations have fixed by generation 19 than actually did go to fixation in the experiment in Figure 3.4. This is because the variance in offspring number is about 70% greater than that assumed in the Wright-Fisher model.

PROBLEM 3.2 Consider a population of four diploid individuals. Calculate the probability that a population with four copies of allele A (allele frequency $p = \frac{1}{2}$) drifts in one generation to having three copies. What is the probability that the population will have four copies of A? Five copies? Now consider a population of the same size, but initially with two copies of A. What is its probability of drifting to one, two, or three copies?

ANSWER Applying Equation 3.2, we get $T_{43} = [8!/(5!3!)](\frac{1}{2})^8 = 7/32 = 0.219$. $T_{44} = [8!/(4!4!)](\frac{1}{2})^8 = 70/256 = 0.273$. $T_{45} = 0.219$ ($= T_{4,3}$). (Note that the binomial distribution is symmetric when $p = \frac{1}{2}$ so there is equal probability for samples that are symmetrically divergent from $p = \frac{1}{2}$.) In the case when the initial frequency is $\frac{2}{8}$ we get $T_{21} = [8!/1!7!)](\frac{1}{4})(\frac{3}{4})^7 = 0.267$, $T_{22} = [8!/(2!6!)](\frac{1}{4})^2(\frac{3}{4})^6 = 0.311$, and $T_{23} = [8!/(3!5!)](\frac{1}{4})^3(\frac{3}{4})^5 = 0.208$.

PROBLEM 3.3 An alternative formulation of random genetic drift is due to Moran (1958). This model has considerable intuitive appeal, and it also permits explicit expressions to be derived for various quantities of evolutionary interest (Ewens 2004). The Moran model strictly applies only to haploid populations, but to make it comparable to the Wright-Fisher model we will suppose a population of $2N$ haploid individuals. In each generation, the drift process begins by sampling two individuals at random. The sampling is carried out "with replacement," so the same individual could be chosen twice. If the two sampled individuals are different, pick one of them at random to produce a single offspring, and return the parent and the offspring to the population; discard the other individual. If the two sampled individuals are the same, then return only the offspring to the population. In the Moran model, if a population of $2N$ haploid individuals contains i of type A and $2N - i$ of type a, then the only nonzero transition probabilities are T_{ij} with $j = i - 1$, $j = i$, or $j = i + 1$. These transition probabilities are given by

$$T_{ii} = \frac{i^2 + (2N - i)^2}{(2N)^2} = p^2 + q^2 \qquad (3.3)$$

$$T_{ij} = \frac{i(2N - i)}{(2N)^2} = pq \text{ for } j = i+1 \text{ or } j = i-1$$

Calculate the transition probabilities in the Moran model for the examples in Problem 3.2.

ANSWER Applying Equation 3.3 to the case $p = \frac{4}{8}$, we obtain $T_{43} = 0.25$, $T_{44} = 0.50$, $T_{45} = 0.25$, and for the case $p = \frac{2}{8}$ we obtain $T_{43} = 0.1875$, $T_{44} = 0.6250$, $T_{45} = 0.1875$. Unlike the Wright-Fisher model, the transition probabilities, for either keeping the same number of alleles or else for increasing or decreasing by exactly 1, sum to 1.

Both the Wright-Fisher model and the Moran model incorporate an important feature of random genetic drift. It is that the magnitude of random change in allele frequency is greater when the allele frequency is $\frac{1}{2}$ than when the allele frequency is more skewed. The changes are greater because the variance in the sampling distribution is greatest when $p = \frac{1}{2}$. In the Wright-Fisher model, the variance in allele frequency from one generation of random genetic drift is given by $pq/(2N)$, corresponding to the variance of the proportion in a binomial distribution. The variance drops to zero at $p = 0$ and $p = 1$. In the Moran model. the variance resulting from a single birth/death event is $2pq/(2N)^2$. This looks very different from the variance in the Wright-Fisher model, however multiplying by a factor of $2N$ is needed to convert the individual births and deaths into replacement of the entire population. There is still a factor of 2 in the numerator, which reflects the subtle fact that the variance in offspring number per individual is exactly twice as large in the Moran model as in the Wright-Fisher model (Ewens 2004). In either formulation of random drift, the variance formula makes it clear that a large population will change allele frequency more slowly than a smaller population, because the sampling variance varies as the reciprocal of population size.

PROBLEM 3.4 Simulating random drift can be a very time-consuming proposition. If one wants to simulate a population of 1000 individuals for 1000 generations, one has to draw 10^6 random numbers and for each decide whether to accept or reject each genotype. Kimura (1980b) came up with a shortcut that relates very closely to how the diffusion approximation works (see the next section). The trick is to use the recursion: $p' = p + (2U - 1)\sqrt{(3pq/2N)}$, where U is a random number uniformly distributed in the range between 0 and 1. In each generation, you pick a random number U, and then calculate the realization of the next generation's allele frequency from the above recursion. Why does this approach work? (*Hint:* The variance in a uniform distribution is the square of the range divided by 12.)

ANSWER The expression $2U - 1$, where U is a number between 0 and 1, yields a value from −1 to +1, or a range of 2. The range of $(2U - 1)\sqrt{(3pq/2N)}$ is therefore $2\sqrt{(3pq/2N)}$. Squaring this expression and dividing by 12, the variance of this uniform random variable equals $pq/2N$, which is exactly that from a binomial sampling distribution. Each generation the allele frequency has an equal chance of increasing or decreasing, and the variance in the allele frequency change is $pq/2N$. Even though the distribution of change in allele frequency is uniform in the pseudosampling simulation instead of binomial (as it is in the Wright-Fisher model), this process can reproduce most of the results of the complete brute-force simulation at a tiny fraction of the computer time. The trade-off is that one must be a little careful when near the fixation states, because the algorithm as described can yield allele frequencies less than 0 or greater than 1.

3.3 THE DIFFUSION APPROXIMATION

The pattern of change in allele frequency shown in Figure 3.4 is very nearly that expected theoretically for an ideal population, as shown in Figure 3.5. This distribution was obtained by successive multiplication of a matrix whose elements are given by the transition probabilities in Equation 3.2. Although the full-blown theory of random genetic drift requires mathematics beyond the scope of this book (see Kimura 1955, 1964, 1976; Wright 1969; Crow and Kimura 1970; Kimura and Ohta 1971; Ewens 2004), in the next section we provide an introductory tidbit to impart the flavor. If you are a student with no background in calculus, the discussion may seem quite

mysterious, but please do not be discouraged because a detailed understanding is not necessary to understand the rest of this chapter or anything later in the book.

An Approach Looking Forward

An elegant alternative to successive matrix multiplication is based on a diffusion approximation (Fisher 1922; Wright 1945; Kimura 1957, 1964). The diffusion approximation assumes that random drift disperses allele frequencies among subpopulations in a manner analogous to heat diffusing through a metal rod or tiny particles diffusing under Brownian motion (Kolmogorov 1931). The idea is to assume that the subpopulations are large enough that the allele frequencies change smoothly through time, not in large jumps. Then the statistical distribution of allele frequencies at any time is a continuous function that we may denote as $\phi(p, x; t)$, where x represents the allele frequency at time t among a large number of segregating populations ($0 < x < 1$), and p is the initial frequency among these populations. The theoretical problem is to formulate an equation that describes how $\phi(p, x; t)$ changes under random genetic drift, and to solve the equation. At any time t, the function $\phi(p, x; t)$ is a smooth, continuous function approximating the histogram of allele frequencies among the subpopulations in Figure 3.5, except that $\phi(p, x; t)$ pertains only to the unfixed subpopulations still segregating for A and a.

There are actually two approaches for obtaining a diffusion equation, each of which has advantages and limitations. One approach is to ask how the distribution $\phi(p, x; t)$ changes as we go forward in time. To explain the meaning of the equation, we will allow x and t to change only in small, discrete increments of Δx and Δt. There are two reasons why the state x could change in the time Δt. One is random genetic drift, the other is a systematic force that might include mutation or selection. We will assume that A is the favored allele, and define $M(x)$ as the probability that x increases by the amount Δx because of the systematic force. The force of random drift is measured by the probability $V(x)$ that x changes because of drift, either decreasing by the amount Δx with probability $V(x)/2$ or increasing by the amount Δx with probability $V(x)/2$. In any time interval Δt, therefore, the probability that x remains at x equals $1 - M(x) - V(x)$.

The reasoning is outlined in Table 3.1. Because changes in state are limited to $+\Delta x$ or $-\Delta x$, a subpopulation can be in state x at time $t + \Delta t$ only if it was in state $x + \Delta x$, x, or $x - \Delta x$ at time t, and these have probabilities proportional to $\phi(p, x + \Delta; t)$, $\phi(p, x; t)$, and $\phi(p, x - \Delta x; t)$, respectively. A subpopulation in state $x - \Delta x$ can change to state x with probability $M(x - \Delta x) + V(x - \Delta x)/2$ according to whether it was pushed by a systematic force (for example, mutation or selection), or else changed randomly because of random drift. A subpopulation in state $x + \Delta x$ can change to state x with probability $V(x - \Delta x)/2$ due to random drift. Finally, a subpopulation in state x can remain in state x with probability $1 - M(x) - V(x)$. The required function for $\phi(p, x; t)$

TABLE 3.1 Random Genetic Drift Looking One Generation Forward in Time

Possibilities for frequency after t generations	Probability of specified frequency after t generations	Possibilities to change to x in next interval Δt	Probability of specified change in next interval Δt
$x - \Delta x$	$\phi(p, x - \Delta x; t)$	$x - \Delta x \rightarrow x$ by systematic force	$M(x)$
	$\phi(p, x - \Delta x; t)$	$x - \Delta x \rightarrow x$ by random drift	$V(x)/2$
$x + \Delta x$	$\phi(p, x + \Delta x; t)$	$x + \Delta x \rightarrow x$ by random drift	$V(x)/2$
x	$\phi(p, x; t)$	x remains at x	$1 - M(x) - V(x)$

obtained by summing the products of columns 2 and 4 in Table 3.1, which after some simplification yields the difference equation

$$\begin{aligned}\phi(p,x;t+\Delta t)-\phi(p,x;t) = \\ -\left[M(x)\phi(p,x;t)-M(x-\Delta x)\phi(p,x-\Delta x;t)\right] \\ +\frac{1}{2}\left\{\left[V(x+\Delta x)\phi(p,x+\Delta x;t)-V(x)\phi(p,x;t)\right]\right. \\ \left.-\left[V(x)\phi(p,x;t)-V(x-\Delta x)\phi(p,x-\Delta x;t)\right]\right\}\end{aligned}$$

On the left-hand side of the equal sign is the change in ϕ ($\Delta\phi$) for a given change in t (Δt). On the right-hand side, the first term is the change in $M\phi$ ($\Delta M\phi$) for a given change in x (Δx), and the second term is the change in the change in $V\phi$ ($\Delta\Delta V\phi$) for a two-step change in x ($\Delta\Delta x$). In symbols, the difference equation can be written as

$$\frac{\Delta\phi(p,x;t)}{\Delta t} = -\frac{\Delta\left[M(x)\phi(p,x;t)\right]}{\Delta x}+\frac{1}{2}\frac{\Delta\left\{\Delta\left[V(x)\phi(p,x;t)\right]\right\}}{\Delta(\Delta x)}$$

At this point we can take the limit as $\Delta t \rightarrow 0$ and $\Delta x \rightarrow 0$ (as we also overlook a number of technical details) to obtain what us called the **Kolmogorov forward equation**:

$$\frac{\partial\phi(p,x;t)}{\partial t} = -\frac{\partial\left[M(x)\phi(p,x;t)\right]}{\partial x}+\frac{1}{2}\frac{\partial^2\left[V(x)\phi(p,x;t)\right]}{\partial x^2} \qquad [3.4]$$

This is a partial differential equation, and given some initial function $\phi(p, x; 0)$, it can be solved (though not easily) for $\phi(p, x; t)$. We have not yet specified

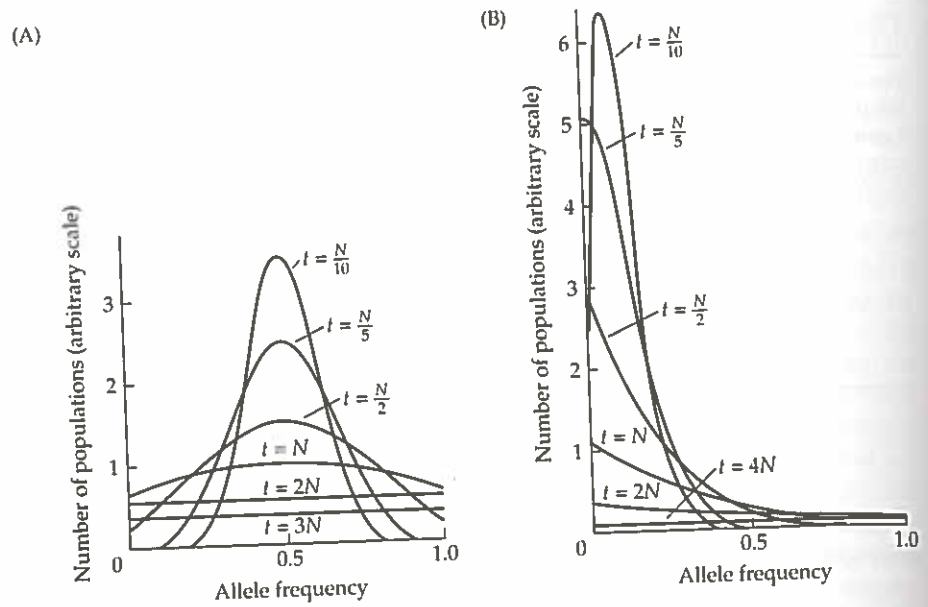

FIGURE 3.6 Theoretical results of random genetic drift. (A) Initial allele frequency = 0.5. (B) Initial allele frequency = 0.1. The curves have been scaled so that the area under each curve is equal to the proportion of populations in which fixation or loss has not yet occurred. The curves are therefore the distributions of allele frequencies among segregating populations. (From Kimura 1955.)

$M(x)$ or $V(x)$ in terms that have any relation to population genetics. The function $M(x)$ is a symbol for the change in allele frequency that occurs in one generation due to any systematic force such as mutation, migration, or selection. The function $V(x)$ also has a straightforward biological interpretation; $V(x)$ is the variance in allele frequency after one generation of binomial sampling of $2N$ alleles according to Equation 3.1; hence $V(x) = x(1-x)/(2N)$.

Many aspects of Equation 3.4 were explored by Wright (1931), and the formal solution to this equation, found by Kimura (1955), required some heavy mathematics. For our purposes, some graphs will illustrate the important properties of the forward diffusion equation. The solutions for $M(x) = 0$ are the curves plotted in Figure 3.6, which show the theoretical distributions of allele frequency among unfixed populations after various times (t) measured in units of N generations. In Figure 3.6A, all populations have an initial allele frequency of $\frac{1}{2}$, as in the actual populations in Figure 3.4; after about $t = 2N$ generations, the distribution of allele frequency is essentially flat, and by this time about half the populations are still unfixed. The distributions in Figure 3.6 refer only to those populations that are unfixed; as time goes on, more and more of the populations become fixed, and the distributions progressively pile up at 0 and 1, as in the histograms in Figure 3.4. Indeed, in Figure 3.6, the area under each curve is equal to the proportion of unfixed populations, which becomes progressively smaller. In particular, the rate at which the height of the distribution decreases once it becomes flat is about $1/(2N)$ per generation. To illustrate that the diffusion approximation and the Wright-Fisher model give very similar results, Figure 3.7 shows the diffusion approximation for the data in Figure 3.4, with $2N = 32$, $p_0 = \frac{1}{2}$, and t running from generation 1 through generation 19.

Figure 3.6B shows what happens when the initial allele frequency is 0.1; here the distributions are highly asymmetrical, and the distribution of allele frequency does not become flat until about $t = 4N$ generations, by which time only about 10% of the populations remain unfixed. Once a flat distribution of allele frequency is reached, the distribution remains flat, but random drift continues until fixation or loss has occurred in all populations.

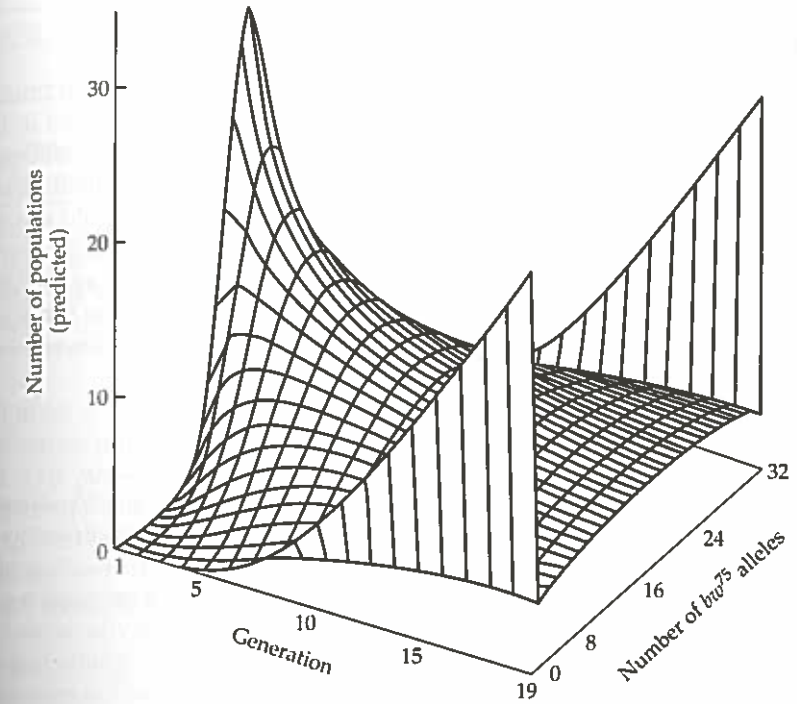

FIGURE 3.7 Kimura's (1955) solution to the diffusion equation for the particular case of $N = 16$. This is the three-dimensional view of Figure 3.6, and represents the diffusion approximation to the exact solution obtained from the Wright-Fisher model in Figure 3.5.

An Approach Looking Backward

To find another equation for $\phi(p, x; t)$, we may also look backward in time to the beginning of the process and consider what may have happened in the very first increment of time Δt. Since the subpopulations initially all begin with an allele frequency of p, in the first time increment Δt a particular subpopulation could change its state to a frequency of $p + \Delta p$, or it could change its state to $p - \Delta p$, or it could remain at p. These possibilities have relative probabilities $M(p) + V(p)/2$, $V(p)/2$, and $1 - M(p) - V(p)$, where again $M(p)$ measures the strength of any systematic force tending to increase the allele frequency and $V(p)$ measures the variance in allele frequency due to random genetic drift.

The bookkeeping is shown in Table 3.2. If p changed state to $p + \Delta p$ in the first time increment, then the probability of the subpopulation achieving state x in the subsequent $t - \Delta t$ time units is proportional to $\phi(p + \Delta p, x; t - \Delta t)$. Similarly, going from state $p - \Delta p$ to state x in $t - \Delta t$ time units has a probability proportional to $\phi(p - \Delta p, x; t - \Delta t)$. Finally, going from state p at time Δt to state x at time t has a probability proportional to $\phi(p, x; t - \Delta t)$. The relevant equation for $\phi(p, x; t)$ is obtained by summing the products of columns 2 and 3 in Table 3.2. After some rearrangement we obtain

$$\begin{aligned}&\phi(p,x;t)-\phi(p,x;t-\Delta t)=\\&M(p)\left[\phi(p+\Delta p,x;t-\Delta t)-\phi(p,x;t-\Delta t)\right]\\&+\frac{V(p)}{2}\left\{\left[\phi(p+\Delta p,x;t-\Delta t)-\phi(p,x;t-\Delta t)\right]\right.\\&\left.-\left[\phi(p,x;t-\Delta t)-\phi(p-\Delta p,x;t-\Delta t)\right]\right\}\end{aligned}$$

As before, the left hand side is equal to the change in ϕ ($\Delta\phi$) for a given change in t (Δt). On the right-hand side, the first term is $M(p)$ times the change in ϕ ($M\Delta\phi$) for a given change in p (Δp), and the second term is $V(p)$ times the change in the change in ϕ ($V\Delta\Delta\phi$) for a two-step change in p ($\Delta\Delta p$). In these terms, the difference equation can be written as

$$\frac{\Delta\phi(p,x;t)}{\Delta t}=M(p)\frac{\Delta\phi(p,x;t)}{\Delta p}+\frac{V(p)}{2}\frac{\Delta\left(\Delta\phi(p,x;t)\right)}{\Delta(\Delta p)}$$

Once again we will ignore some technical requirements and simply assert that, in the limit as $\Delta t \to 0$ and $\Delta p \to 0$, the difference equation converges to a partial differential equation called the **Kolmogorov backward equation:**

$$\frac{\partial\phi(p,x;t)}{\partial t}=M(p)\frac{\partial\phi(p,x;t)}{\partial p}+\frac{V(p)}{2}\frac{\partial^2\phi(p,x;t)}{\partial p^2} \tag{3.5}$$

For answering questions of population genetic interest in random drift, the Kolmogorov backward equation (see Equation 3.5) is often more useful

TABLE 3.2 Random Genetic Drift Looking Backward at the First Generation

Possibilities for change in first generation	Probability of specified change in first-generation	Probability of changing to x in remaining $t - \Delta t$ generations
$p \to p + \Delta p$ by systematic force	$M(p)$	$\phi(p + \Delta p, x; t - \Delta t)$
$p \to p + \Delta p$ by random drift	$V(p)/2$	$\phi(p + \Delta p, x; t - \Delta t)$
$p \to p - \Delta p$ by random drift	$V(p)/2$	$\phi(p - \Delta p, x; t - \Delta t)$
$p \to$ remains at p	$1 - M(p) - V(p)$	$\phi(p, x; t - \Delta t)$

than the forward equation (see Equation 3.4). The quantities of interest include the probability of ultimate fixation of an allele, the average time to fixation of alleles that are eventually fixed, and so forth. To give a sense of how the backward equation is used for these purposes, imagine the form of Equation 3.5 at a time so advanced that the distribution of allele frequencies $\phi(p, x; t)$ is no longer changing. Since random drift will continue to change the allele frequencies as long as any subpopulations are still polymorphic, the statement that $\phi(p, x; t)$ is no longer changing means that all subpopulations have become fixed for one allele or the other, which furthermore implies that the left-hand side of Equation 3.5 equals 0 and that the right-hand side no longer depends on either x (because no populations are still segregating) or t. To emphasize that we are now dealing with a function of a single variable, population geneticists often rewrite this form of Equation 3.5 as

$$0=M(p)\frac{du(p)}{dp}+\frac{V(p)}{2}\frac{d^2u(p)}{dp^2} \tag{3.6}$$

In this equation, the symbol d is used instead of ∂ to emphasize that $u(p)$ is a function of a single variable. In words, $u(p)$ is the probability of ultimate fixation of the allele A, given an initial frequency of p. Alternatively, $u(p)$ may be interpreted as the proportion of all subpopulations in which A eventually becomes fixed. In the case of pure random drift with no systematic force, $M(p) = 0$. Equation 3.6 then becomes

$$0=\frac{V(p)}{2}\frac{d^2u(p)}{dp^2} \tag{3.7}$$

This equation defines a family of curves, but the one of interest in population genetics has the property $u(0) = 0$, which says that an allele that does not exist cannot be fixed, and the property $u(1) = 1$, which says that an allele that is already fixed is eventually fixed.

PROBLEM 3.5 For an initial frequency of the A allele of p ($0 < p < 1$), show that $u(p) = p$ is a solution of the differential equation (see Equation 3.7).

ANSWER What needs to be shown is that Equation 3.7 is satisfied when $u(p) = p$. Although $V(p) = p(1 - p)/2N$, this is not relevant to the solution. The solution follows from the fact that, when $u(p) = p$, then $du(p)/dp = 1$ and $d^2u(p)/dp^2 = 0$. Hence $u(p) = p$ is a solution of Equation 3.7 so long as $V(p) \neq 0$. The biological meaning of $u(p) = p$ is that, because of random genetic drift, an allele present in a population at frequency p will ultimately be fixed with probability p and lost with probability $1 - p$, provided it has no effects on the organisms's ability to survive and reproduce (such alleles are often called *selectively neutral alleles*).

Absorption Time and Time to Fixation

For a selectively neutral allele, as indicated in Problem 3.5, the probability of ultimate fixation is equal to its initial allele frequency. Many other important results also follow from an analysis of the Kolmogorov backward equation (see Equation 3.5). These include the expected time for a neutral allele to go to fixation (given that it is eventually fixed) or to loss (given that it is eventually lost). Assuming an initial allele frequency p, Kimura and Ohta (1969) showed that the mean time [$\bar{t}_1(p)$, in generations] until the allele is fixed (given that it is eventually fixed) is

$$\bar{t}_1(p) = -4N\left(\frac{1-p}{p}\right)\ln(1-p) \qquad [3.8]$$

Similarly, they showed that the mean time to loss $\bar{t}_0(p)$ (given that the allele is eventually lost) is

$$\bar{t}_0(p) = -4N\left(\frac{p}{1-p}\right)\ln(p) \qquad [3.9]$$

Combining Equations 3.8 and 3.9, the mean persistence time of an allele [$\bar{t}(p)$, the average length of time that a population is segregating for A and a] is given by $\bar{t}(p) = p\bar{t}_1(p) + (1 - p)\,\bar{t}_0(p)$, which equals

$$\bar{t}(p) = -4N\left[(1-p)\ln(1-p) + p\ln(p)\right] \qquad [3.10]$$

Figure 3.8 shows the average times to fixation, loss, and persistence of a neutral allele. An allele is expected to remain in a population for the longest time when its initial frequency is $\frac{1}{2}$. When $p = \frac{1}{2}$, the average time that a population remains unfixed is about $2.77N$ generations.

Equations 3.8 and 3.9 are of particular interest when $p = 1/(2N)$, that is, when a new neutral mutation has just occurred and there is only one copy in the population. In this case, the probability of eventual fixation is $1/(2N)$, and, given that the allele is eventually fixed, the average time to fixation is approximately $4N$ generations. On the other hand, the probability that a new neutral mutation is eventually lost is $1 - 1/(2N)$, and, given that the allele is eventually lost, the average time to loss is approximately $2\ln(2N)$ generations. In other words, new neutral alleles that are eventually fixed usually take a long time to be fixed, whereas those that are lost are lost very quickly. For the specific example of $N = 500$, the average new neutral mutation that is eventually fixed requires 2000 generations to be fixed, whereas the average new neutral mutation that is destined to be lost requires fewer than 14 generations to be lost.

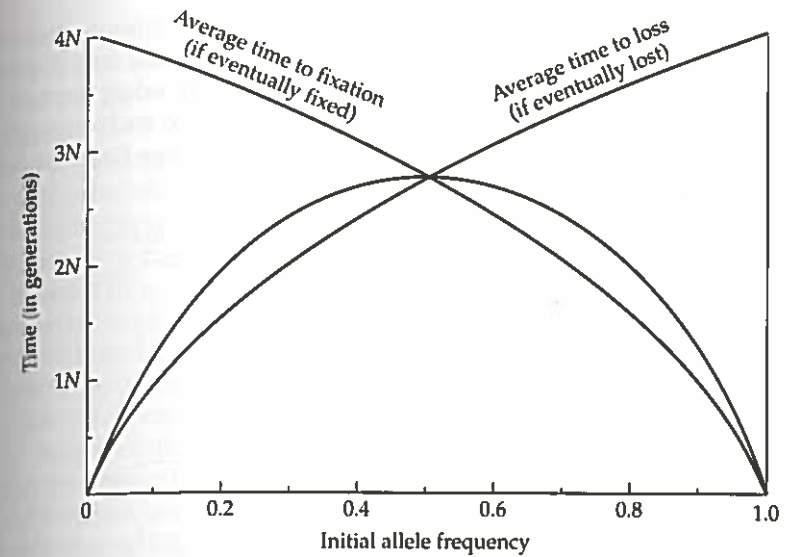

FIGURE 3.8 Average persistence of a neutral allele in an ideal diploid population of size N, plotted against initial allele frequency.

3.4 RANDOM DRIFT IN A SUBDIVIDED POPULATION

Most real populations are subdivided into smaller units, for example, humans are concentrated in cities, towns, and villages; animals form herds, flocks, or schools; and plants are aggregated into stands. This kind of subdivision is reminiscent of the population structure in Figure 3.5, except that, in nature, the subpopulations are not genetically isolated from one another owing to some *migration*, or movement, of individuals among the subpopulations, which results in **gene flow**, or exchange of genes, between them.

Nevertheless, random genetic drift will tend to cause subpopulations to undergo differentiation in their allele frequencies, even in the face of some gene flow. To see this point, consider the four subpopulations diagrammed in Figure 3.9. Each begins with an allele frequency of $p = \frac{1}{2}$, and each undergoes random drift independently following binomial sampling (see Equation 3.2).

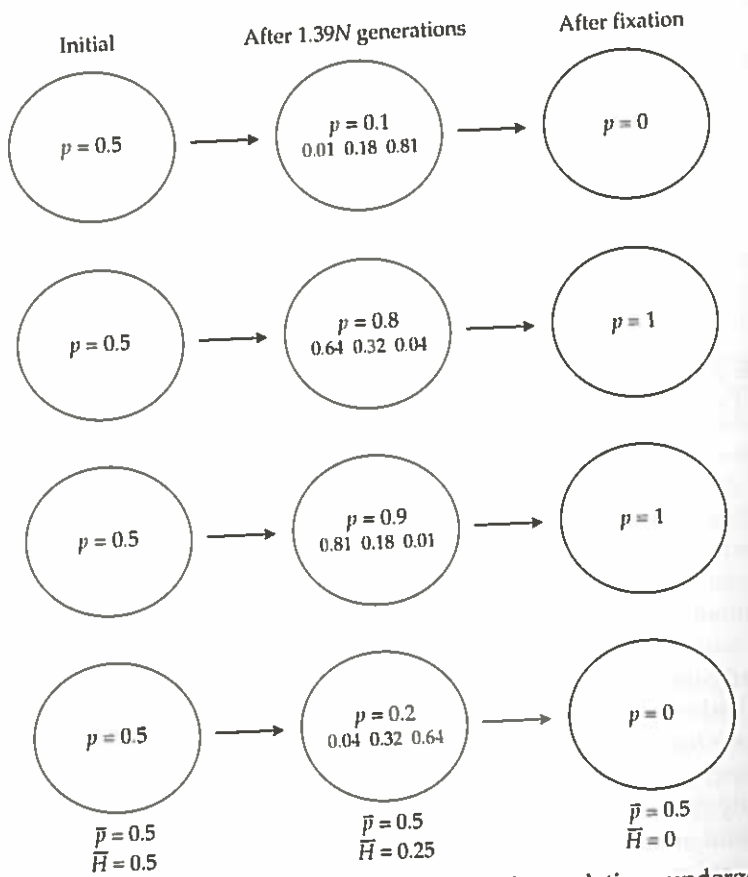

FIGURE 3.9 Schematic showing a set of four subpopulations undergoing the process of random genetic drift. Initially the allele frequency is 0.5 in all four subpopulations, and the average heterozygosity is also 0.5. As the subpopulations drift in allele frequency, the average allele frequency is expected to remain the same (indicated by $\bar{p}$ remaining at the value 0.5), but the average heterozygosity decreases. For the intermediate generation when $t = 1.39N$ generations, the allele and genotype frequencies in each subpopulation are given, as well as the average allele frequency and heterozygosity across subpopulations. By this time the average heterozygosity is reduced to 50% of the value expected without population subdivision. Ultimately, all subpopulations go to fixation, half fix one allele and half fix the other, so the average allele frequency is still 0.5, whereas the heterozygosity is zero.

We assume that random mating takes place within any particular subpopulation (call it subpopulation number i). Therefore, if the allele frequencies of A and a in the ith subpopulation are denoted p_i and q_i, then the genotype frequencies of AA, Aa, and aa are given by the familiar Hardy-Weinberg principle as p_i^2, $2p_iq_i$ and q_i^2. Furthermore, picture the situation in Figure 3.9 at a time so advanced that all subpopulations are fixed for one allele or the other. Within the ith subpopulation, therefore, either p_i equals 0 or else p_i equals 1. The genotype frequencies of AA, Aa, and aa in that subpopulation are either 0, 0, and 1 (if $p_i = 0$), or 1, 0, and 0 (if $p_i = 1$). These genotype frequencies, though extreme, still satisfy the Hardy-Weinberg principle. Thus, within any one subpopulation in Figure 3.9, the frequency of heterozygotes is that expected with random mating.

The total population in Figure 3.9 is composed of the aggregate of the four subpopulations, and in the total population there is a deficiency of heterozygous genotypes. Suppose that we were unaware of the subpopulation structure and sampled from the total population as if it were a single randomly mating population. If we were to sample randomly from the far right of Figure 3.9, when no populations are still segregating, we would obtain an allele frequency of $p = \frac{1}{2}$. Assuming Hardy-Weinberg equilibrium, we would naively expect a fraction $2pq = \frac{1}{2}$ of the genotypes to be heterozygous. In fact, we would have sampled no heterozygous genotypes at all! This rather paradoxical result—that there is a deficiency of heterozygotes in the total population even though random mating occurs within each subpopulation—is a consequence of the random genetic drift of allele frequencies among subpopulations due to their finite size. The extreme case when each subpopulation is fixed is easy to understand: A population with allele frequency $\frac{1}{2}$ could only be made up of two subpopulations fixed for A and two subpopulations fixed for a; the average allele frequency is $\frac{1}{2}$, but the total population has no heterozygotes.

We are now in a position to quantify the manner in which subpopulations diverge in allele frequency under random genetic drift. To do this efficiently, we need to introduce a concept known as allele identity by descent. Two alleles are **identical by descent** if they are replicas (by DNA replication) of a gene present in some previous generation. This definition does not speak for itself, because if one goes far enough backward in time, every pair of alleles must be identical by descent, and so the concept may seem vacuous. The way out of this trap is to choose some arbitrary time in the past, which may be recent or remote according to the application, and declare that at this time every allele is distinct from every other allele. In this fashion, any earlier identity by descent is erased, and therefore the identity by descent spoken of in the definition is common ancestry through DNA replication since that arbitrary time in the past when every allele was declared distinct.

The concept of identity by descent is useful because it allows us to distinguish two types of homozygous genotypes. In particular, the A alleles in a

homozygous AA genotype could be alleles that are not identical by descent (which means that these alleles both existed in the population at the time when every allele was declared distinct), or they could be identical by descent (which means that they originated by DNA replication of a single A allele since that time). In some cases alleles may be indistinguishable by means of an experimental procedure (for example, protein electrophoresis), but their status in regard to identity by descent is unknown. Such alleles are said to be *identical by kind* or *identical by state*.

The probability that the alleles in an individual are identical by descent is often denoted F, following Wright (1922) who called it the **fixation index**. In the context of population subdivision, F as used in this chapter is the same quantity that in Chapter 6 will be denoted F_{ST}. In this chapter we will drop the subscript because we will want to track changes in F_{ST} through time, and in this case the probability of allele identity by descent in generation t is conveniently represented as F_t.

Now we can be more specific about what we mean by saying that one can choose some arbitrary time in the past and declare that at this time every allele is distinct. In the context of population subdivision as illustrated in Figure 3.9, the time in the past when the alleles are declared as distinct is in the initial populations, when the population subdivision first takes place, and all subpopulations have the same allele frequencies. In symbols, we declare that, at time $t = 0$ when the subpopulations are first established, $F_t = 0$. As time goes on, and each subpopulation undergoes random drift, the genotype frequencies in each subpopulation will satisfy the Hardy-Weinberg principle because mating within subpopulations is random. However, the allele frequencies among the subpopulations will change because of random genetic drift, and moreover the value of F_t will gradually increase as more and more alleles within any subpopulation become identical by descent owing to common ancestry.

The rate of increase in F_t can be calculated with the aid of the diagram in Figure 3.10. This figure shows the $2N$ alleles in a breeding population of generation $t - 1$. In sampling alleles for generation t, the first allele chosen may be any of those present in generation $t - 1$ with equal probability. Having chosen the first allele, the probability that the second allele chosen is of the same type as the first is $1/(2N)$ (in which case $F = 1$), because this is the frequency of each allelic type in the pool of gamates; the probability that the second chosen allele is of a different type from the first is accordingly $1 - 1/(2N)$ (in which case $F = F_{t-1}$). Putting these two possibilities together, the relationship between F_t and F_{t-1} is seen to be

$$F_t = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)F_{t-1} \tag{3.11}$$

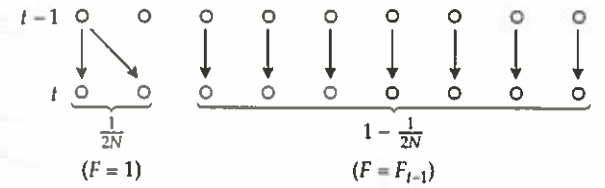

FIGURE 3.10 Diagram illustrating the reasoning behind the recursion for F in a finite population. When the gametes are drawn to make up the population at generation t, there is a chance $1/(2N)$ that any pair of alleles will have been identical in generation $t - 1$. If this happens, the probability of identity is 1. For the allele pairs drawn in generation t from two distinct alleles at generation $t - 1$ [the probability of this happening is $1 - 1/(2N)$], the probability of identity is F_{t-1}. Adding the probabilities of these two events, we get $F_t = 1/(2N) + [1 - 1/(2N)]F_{t-1}$.

Multiplying both sides by –1 and then adding 1 to each side leads to

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

and so

$$1 - F_t = \left(1 - \frac{1}{2N}\right)^t \left(1 - F_0\right) \tag{3.12}$$

or, when $F_0 = 0$,

$$F_t = 1 - \left(1 - \frac{1}{2N}\right)^t \tag{3.13}$$

Figure 3.11 shows the rapid increase of F_t in small populations. Even though the genotype frequencies in each individual subpopulation are in Hardy-Weinberg proportions, the frequency of homozygous genotypes in the overall population steadily increases. Conversely, as the frequency of homozygous genotypes increases, the frequency of heterozygous genotypes decreases until, when $F_t = 1$, there are no heterozygous genotypes left and all subpopulations are fixed for either A or a. At any time, the average frequency of heterozygous genotypes among the subpopulations, H_t, relative to what it would be without population subdivision, H_0, decreases linearly in F_t, hence we have $H_t/H_0 = 1 - F_t$, or $H_t = (1 - F_t)H_0$. Solving Equation 3.13 for $1 - F_t$ and substituting, we obtain

$$H_t = \left(1 - \frac{1}{2N}\right)^t H_0 \approx H_0 e^{-t/2N} \tag{3.14}$$

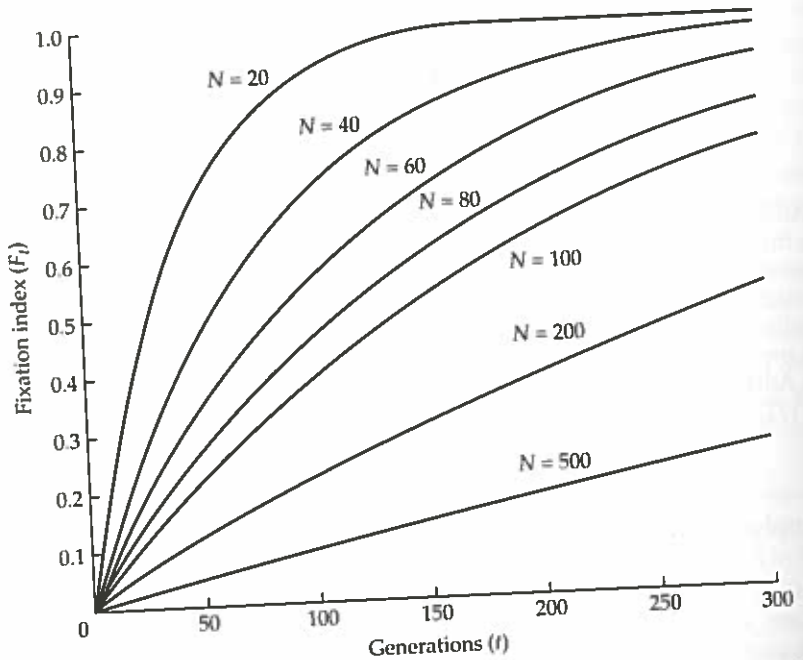

FIGURE 3.11 Increase of F_t in ideal populations as a function of time and effective population size N.

We emphasize again that each individual subpopulation undergoes random drift and remains in approximate Hardy-Weinberg proportions, and that the symbol H_t represents a sort of "virtual heterozygosity" in which the frequency of heterozygous genotypes is averaged across many subpopulations. Equation 3.14 shows that pure random drift should result in the heterozygosity decreasing at a geometric rate, since H_t is multiplied by the constant $[1 - 1/(2N)]$ each generation. Experimental tests of this prediction are shown in Figure 3.12. Figure 3.12A shows how the heterozygosity averaged across the subpopulations in Figure 3.4 declines over generations, but the theoretical curve when $N = 16$ does not fit the data very well. In fact, the rate of decline of heterozygosity is greater than the theoretical expectation, as though the population size were smaller than $N = 16$. In other words, the populations in Figure 3.4 decrease in heterozygosity as if each had a population size of $N = 9$ rather than its actual size of $N = 16$. We call $N = 9$ the *effective size* of the subpopulations, as distinct from the actual size (see Section 3.5). The theory also predicts that the allele frequency, averaged across populations, is not expected to change, and the data agree with this aspect of the theory quite well (Figure 3.12B).

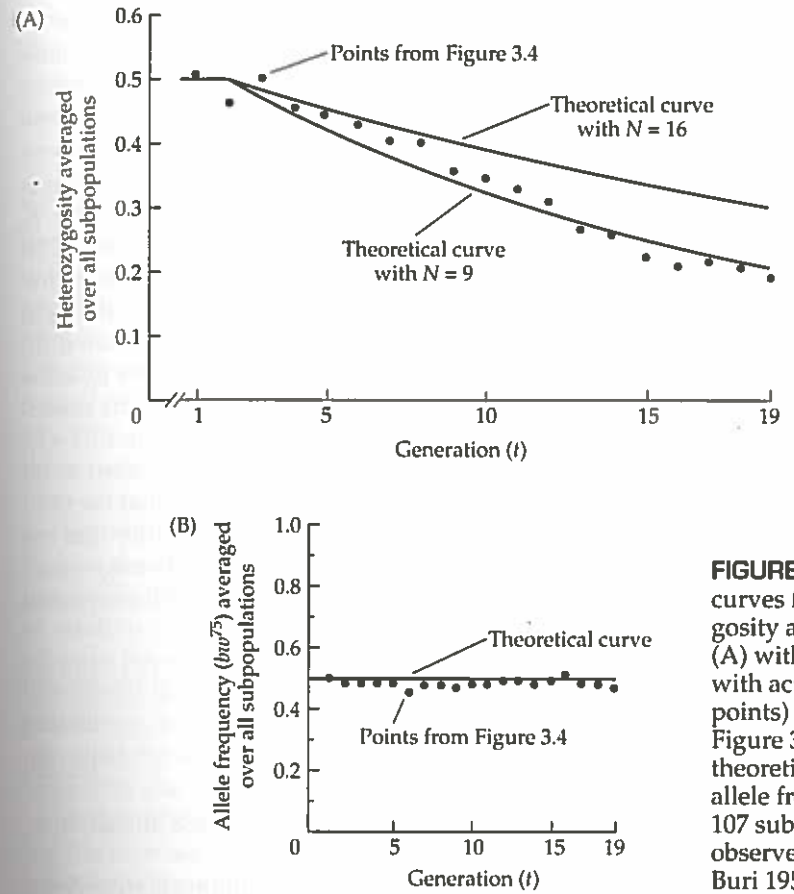

FIGURE 3.12 Theoretical curves for average heterozygosity among subpopulations (A) with $N = 9$ or $N = 16$, along with actual values (plotted as points) from the experiment in Figure 3.4. Part (B) shows the theoretically expected average allele frequency among the 107 subpopulations and the observed average. (Data from Buri 1956.)

PROBLEM 3.6 Use Equation 3.14 to determine the average length of time it would take for a finite population of size N to reduce its initial heterozygosity by a factor of two.

ANSWER Set $H_t = \frac{1}{2} H_0 = H_0 e^{-(t/2N)}$. Now divide both sides by H_0 and take the natural logarithm (base e) to obtain $\ln(\frac{1}{2}) = -t/(2N)$, or $t = -2N \ln(\frac{1}{2}) = 1.39N$ generations. In words, this result says that it requires an average of $1.39N$ generations to halve the heterozygosity, whatever its initial value. Fisher (1918) showed that it also takes $1.39N$ generations to halve what he called the *genic variance* in the population. Since the variance of a binomial sample is $pq/2N$, and the average heterozygosity in a population decreases in proportion to the variance in allele frequency among subpopulations, it follows that the average heterozygosity decreases at the same rate as the variance in allele frequency among subpopulations increases.

Several important consequences of the population structure in Figure 3.9 can now be summarized. First, although each subpopulation is finite in size, we can imagine so many of them that the size of the total population is effectively infinite. For an infinite population, the allele frequencies must remain constant. That is, even though the allele frequency in any individual subpopulation may change willy-nilly due to random genetic drift, the overall average allele frequency of A among subpopulations remains p_0, where p_0 represents the allele frequency of A in the initial populations. Figure 3.12B shows an experimental demonstration of the constancy of average allele frequency. Since F_t is the probability of identity by descent of the two alleles in an individual in generation t, the probability that the two alleles in an individual in generation t are not identical by descent is $1 - F_t$. Because p_0 is the overall allele frequency of A, averaged across all subpopulations, the probability that a randomly chosen individual will be genotypically AA is $p_0^2(1 - F_t)$ [for the case of nonidentity by descent] + p_0F_t [for the case of identity by descent], which equals $p_0^2(1 - F_t) + p_0F_t$. Similarly, the probability that the individual will be Aa equals $2p_0q_0(1 - F_t)$, and likewise the probability that the individual will be aa equals $q_0^2(1 - F_t) + q_0F_t$. To summarize, the average genotype frequencies among subpopulations at any time t have the expected values:

$$AA: \; p_0^2(1 - F_t) + p_0F_t = p_0^2 + p_0q_0F_t \tag{3.15a}$$

$$Aa: \; 2p_0q_0(1 - F_t) = 2p_0q_0 - 2p_0q_0F_t \tag{3.15b}$$

$$aa: \; q_0^2(1 - F_t) + q_0F_t = q_0^2 + p_0q_0F_t \tag{3.15c}$$

where $q_0 = 1 - p_0$ is the average frequency of a, averaged across all subpopulations.

Note that, while each individual subpopulation maintains Hardy-Weinberg frequencies, the average genotypic frequencies in the total population are different because there is an excess of homozygotes and a deficiency of heterozygotes. Equation 3.13 implies that the average heterozygosity among subpopulations at time t equals $2p_0q_0(1 - F_t) = 2p_0q_0[1-1/(2N)]^t$, and this is the theoretical curve plotted in Figure 3.12A (with $p_0 = q_0 = \frac{1}{2}$). Additionally, the comment about the variance in the answer to Problem 3.6 can be stated in symbols by saying that, at any time t, the expected variance in allele frequencies among the subpopulations equals $2p_0q_0F_t$.

Since F_t eventually goes to 1, all subpopulations eventually become fixed for one allele or the other (see Equations 3.15). Because the average allele frequency of A remains p_0 even when all subpopulations have become fixed, the proportion of subpopulations that eventually become fixed for A must be p_0, and the proportion that eventually become fixed for a must be q_0. Stated another way, the probability of ultimate fixation of an allele in any ideal subpopulation is equal to the frequency of that allele in the initial population.

This point follows from the diffusion approximation (see Problem 3.5) and is illustrated in the experiment in Figure 3.4, where $p_0 = \frac{1}{2}$; in this case, by generation 19, a total of 58 populations have become fixed, among which 30 are fixed for the *bw* allele and 28 fixed for the bw^{75} allele.

3.5 EFFECTIVE POPULATION SIZE

As we saw in the *Drosophila* experiments in Figure 3.12, populations generally fluctuate in allele frequency by an amount greater than $pq/(2N)$. No real population can be expected to satisfy the assumptions of a theoretically ideal population in all respects. Hence, in any actual case, there must be corrections for such complications as fluctuations in population size, unequal numbers of males and females, skewed distributions in family size, population structure, and so forth (Crow and Kimura 1970; Ewens 2004). The effects of these complicating circumstances on the change in allele frequencies and rates of allele fixation can be approximated by calculating the *effective size* of the population and using this value in the theory for an ideal population. That is, the **effective population size** of an actual population is the number of individuals in a theoretically ideal population having the same magnitude of random genetic drift as the actual population. There are three kinds of effective population size based on how we choose to measure "magnitude," namely: (1) the change in probability of identity by descent (F), (2) the change in variance in allele frequency, or (3) the rate of loss of heterozygosity. These are called the *inbreeding effective size*, the *variance effective size*, and the *eigenvalue effective size*, respectively.

Wright (1931) first worked out the effective population size by considering the increase in identity by descent in various situations. As noted, the effective population size can also be calculated by determining the rate of change in variance in allele frequency among subpopulations, and Kimura and Crow (1963) first applied this approach to the problem of overlapping generations. Usually, the inbreeding effective size and the variance effective size are the same, but exceptions do occur. Similarly, the variance effective size and the eigenvalue effective size can be distinct (Ewens 1982, 2004). Some of the various factors that require calculation of an effective population size will now be illustrated. We will focus on the inbreeding effective size because this concept is the most widely used.

Fluctuation in Population Size

Correction for fluctuating population size is important because natural populations actually do change in size, sometimes by a factor of 10 or more in a single generation. For the sake of simplicity, assume that the population is ideal in all respects except that its size is not constant. We will consider the situation over just two generations. Suppose that the population sizes in two

successive generations are N_0 and N_1. The arguments laid out in Figure 3.10 imply that

$$1 - F_2 = \left(1 - \frac{1}{2N_1}\right)(1 - F_1) \tag{3.16}$$

and

$$1 - F_1 = \left(1 - \frac{1}{2N_0}\right)(1 - F_0) \tag{3.17}$$

Substituting from the second equation into the first leads to

$$1 - F_2 = \left(1 - \frac{1}{2N_1}\right)\left(1 - \frac{1}{2N_0}\right)(1 - F_0) \tag{3.18}$$

By analogy with the constant N case, it is appropriate to try to express this equation in the general form

$$1 - F_t = \left(1 - \frac{1}{2N}\right)^t (1 - F_0) \tag{3.19}$$

where N is now the *effective* population size, usually symbolized as N_e. In our example $t = 2$, so

$$1 - F_2 = \left(1 - \frac{1}{2N}\right)^2 (1 - F_0) \tag{3.20}$$

Setting the two expressions for $1 - F_2$ equal to each other, we obtain

$$\left(1 - \frac{1}{2N}\right)^2 = \left(1 - \frac{1}{2N_0}\right)\left(1 - \frac{1}{2N_1}\right) \tag{3.21}$$

from which $1/N = \frac{1}{2}(1/N_0 + 1/N_1)$ turns out to be an excellent approximation. In general,

$$\frac{1}{N_e} = \frac{1}{t}\left(\frac{1}{N_0} + \frac{1}{N_1} + \cdots + \frac{1}{N_{t-1}}\right) \tag{3.22}$$

and so the effective size N_e is the **harmonic mean** of the actual numbers—the reciprocal of the average of the reciprocals. As illustrated in the problem below, the harmonic mean tends to be dominated by the smallest terms. In biological reality, this means that a single period of small population size, called a **bottleneck**, can result in a serious loss in heterozygosity. Population bottlenecks are thought to account for the very low levels of polymorphism found in extant populations of the elephant seal (Bonnell and Selander 1974) and the cheetah (O'Brien et al. 1985, 1987). A severe population bottleneck often occurs in nature when a small group of emigrants from an established subpopulation founds a new subpopulation; the accompanying random genetic drift is then known as a **founder effect** (see Holgate 1966; Nei et al. 1975; Chakraborty and Nei 1977; Neel and Thompson 1978). Founder effects in human populations have implications in medical genetics, because human populations derived from small numbers of founders may have an elevated incidence of an otherwise rare genetic disorder. Examples include Tay-Sachs diseases in Ashkenazi Jews, diastrophic dystrophy in Finns, familial hyperchylomicronemia in Quebecois, and congenital total color blindness in Pingelap Islanders (reviewed in Scriver 2001). In addition to reducing the effective population size, and thereby increasing F, population bottlenecks and founder effects may affect many other aspects of the genetic variation, including causing a reduced number of alleles, a distorted distribution of allele frequencies, and an increase in linkage disequilibrium.

PROBLEM 3.7 Suppose a population went through a bottleneck as follows: $N_0 = 1000$, $N_1 = 10$, and $N_2 = 1000$. Calculate the effective size of this population across all three generations.

ANSWER Using Equation 3.22, we get $1/N_e = (\frac{1}{3})(\frac{1}{1000} + \frac{1}{10} + \frac{1}{1000}) = 0.034$, or $N_e = 1/0.034 = 29.4$. The average effective number over the three-generation period is only 29.4, whereas the arithmetic average number of individuals is $(\frac{1}{3})(1000 + 10 + 1000) = 670$.

Unequal Sex Ratio, Sex Chromosomes, Organelle Genes

A second important case in which the effective size of a nonideal population can readily be calculated concerns sexual populations in which the number of males and females is unequal. This inequality creates a peculiar sort of bottleneck; because half of the alleles in any generation must come from each sex, any departure of the sex ratio from equality will enhance the opportunity for random genetic drift. This situation is important in wildlife management, where, for many game animals (pheasants and deer come immediately to mind), the legal bag limit for males is much larger than for females. Although some management goals are served by such hunting regulations (for example, the species involved are usually polygamous, so one male can fertilize many females and overall actual population size can be maintained), it must be remembered that the resultant inequality in sex ratio reduces the effective population size. Specifically, if a sexual population consists of N_m males and N_f females, the actual size is

$$N_a = N_m + N_f \tag{3.23}$$

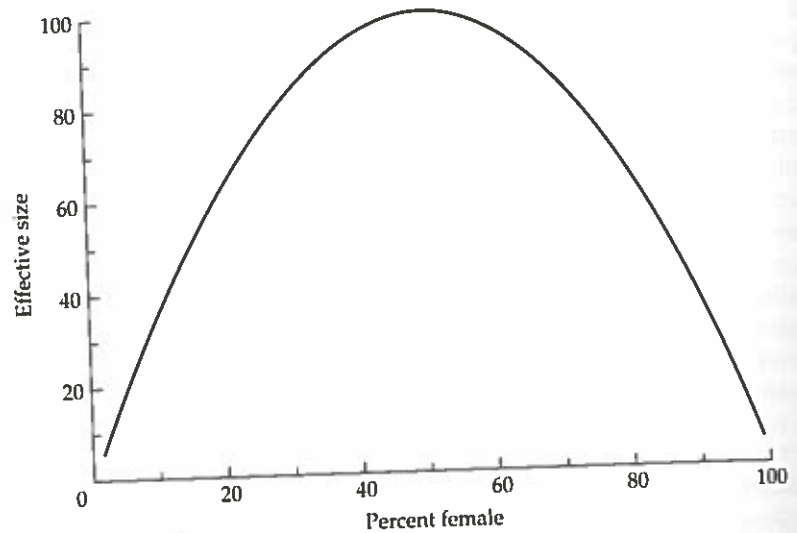

FIGURE 3.13 Effective size falls off rapidly in populations with a skewed sex ratio.

However, the effective population size is

$$N_e = \frac{4N_m N_f}{N_m + N_f} \tag{3.24}$$

Figure 3.13 shows the relationship between sex ratio and the reduction in effective population size. To take a realistic example, if hunting is permitted to a level at which the number of surviving males is one-tenth the number of females, then the effective population size is a mere one-third of the actual number of individuals in the population.

A related problem is the effective population size for an X-linked gene, in which case $\frac{2}{3}$ of the X chromosomes in any generation come from females in the previous generation and $\frac{1}{3}$ come from males. The variance effective population size for an X-linked gene is

$$N_e = \frac{9N_m N_f}{4N_m + 2N_f} \tag{3.25}$$

Equation 3.25 can be justified by noting that the sampling variance for the X chromosomes from males is $p_m q_m / N_m$, whereas the sampling variance for X chromosomes from females is $p_f q_f / 2N_f$, in which p_m and p_f are the frequencies of allele A in males and females, respectively. The frequency of an A-bearing X chromosome in the population is

$$p = \frac{1}{3}p_m + \frac{2}{3}p_f \tag{3.26}$$

Now we use the fact that, if a and b are constants and X and Y are independent random variables, then $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$. In this case $a = \frac{1}{3}$, $b = \frac{2}{3}$, and the variances of p_m and p_f are the binomial variances, and so

$$Var(p) = \frac{1}{9}\left(\frac{p_m q_m}{N_m}\right) + \frac{4}{9}\left(\frac{p_f q_f}{2N_f}\right) \tag{3.27}$$

At steady state, $p_m = p_f = p$ and $q_m = q_f = q$. Making these substitutions and factoring pq results in

$$Var(p) = pq\left(\frac{1}{9}\frac{1}{N_m} + \frac{4}{9}\frac{1}{2N_f}\right) = \frac{pq}{2\left[\dfrac{9N_m N_f}{4N_m + 2N_f}\right]} \tag{3.28}$$

The term in the square brackets corresponds to the N_e in Equation 3.25. It shows why this is a variance effective size: The binomial sampling variance in an ideal population is $pq/(2N_e)$.

PROBLEM 3.8 What is the effective population size for mitochondrial DNA? (Assume transmission is exclusively from mothers to all offspring.) What is the effective population size for a gene on the Y chromosome, given that the population consists of N diploid individuals and is in all respects a theoretically ideal population? (Assume XX individuals are female and XY individuals are male.)

ANSWER Mitochondrial DNA is transmitted essentially exclusively by females, and therefore the chance of drawing two mtDNAs that are identical by descent is $1/N_f$, where N_f is the number of females in the population. However, the probability that two randomly chosen autosomal genes are identical by descent is $1/(2N_e)$. Equating $1/(2N_e) = 1/N_f$ yields $N_e = N_f/2$ as the effective size for the population of mitochondrial DNA molecules. Since $N_f = N/2$ in an ideal population, the effective size for mitochondrial DNA, relative to an autosomal gene in an ideal population, is $N/4$. Similarly, the effective population size for the Y chromosome is $N_m/2$, where N_m is the number of males in the population. As with mitochondrial DNA, the effective size for Y chromosomal DNA, relative to an autosomal gene in an ideal population, is $N/4$. Note that, when $N_f = N_m$, even though mtDNA is present in all individuals, whereas the Y chromosome is present only in males, the effective size of mtDNA is not larger than that of the Y chromosome. The effective size depends on the sampling properties of a gene, which depends not only on how many individuals carry the gene but also on the gene's mode of transmission.

Variance in Offspring Number

An ideal population is one in which each breeding individual has an equal chance of contributing offspring to the next generation. Technically, this means that the statistical distribution of the number of offspring per individual is a binomial distribution with mean 1 and variance $1 - 1/N$. The distribution is binomial because its range is the fixed interval $[0, N]$, owing to the fact that no individual can have more than N progeny. If N is reasonably large, this binomial distribution is virtually identical to a Poisson distribution with mean and variance equal to 1. Nevertheless, the assumption that each individual has the same distribution of number of progeny is usually unrealistic because, in real organisms, breeding individuals can manifest large differences in their number of progeny. A more realistic model is one in which there are N individuals in the population and in which the ith individual ($i = 1, 2, \ldots, N$) produces n_i offspring. In this situation, the effective size of the population is defined as the reciprocal of the probability P that two randomly chosen gametes in the next generation come from the same parent in the previous generation (Crow and Kimura 1970). We will denote the mean and variance of the distribution of offspring number as ξ (Greek xi) and σ^2, respectively. With these definitions,

$$\xi = \frac{\Sigma n_i}{N} \text{ and } \sigma^2 = \frac{\Sigma n_i}{N} - \left(\frac{\Sigma n_i}{N}\right)^2 \tag{3.29}$$

The probability P that two randomly chosen gametes come from the same parent is given by

$$P = \frac{\Sigma\binom{n_i}{2}}{\binom{N\xi}{2}} = \frac{\Sigma n_i(n_i-1)}{N\xi(N\xi-1)} = \frac{\Sigma n_i^2 - \Sigma n_i}{N\xi(N\xi-1)} \tag{3.30}$$

The rationale for Equation 3.30 is that the numerator is the number of ways that two randomly chosen alleles can be present in offspring from the same parent, and the denominator is the number of ways that two randomly chosen alleles can have any parents. Substitution of Equation 3.29 into Equation 3.30 and a little rearrangement yields

$$P = \frac{(\sigma^2/\xi) + (\xi - 1)}{N - 1}$$

But since $N_e = 1/P$ by definition, we can write

$$N_e = \frac{N-1}{(\sigma^2/\xi) + (\xi - 1)} \tag{3.31}$$

and so, when $\xi = 1$, N_e is approximately equal to N/σ^2. Therefore, a large variance in offspring number reduces the effective population size by a factor of $1/\sigma^2$, thereby speeding up the process of random genetic drift. The flip side of this principle suggests a management strategy for endangered species: Loss of genetic variation can be reduced when the variance in offspring number is minimized, because if σ^2 is smaller than 1, the effective population size can be larger than the actual population size.

Variance in offspring number can have a large effect on random genetic drift, as can be seen in particularly important cases in which genes are transmitted by different mechanisms in males and females (for example, in the X and Y chromosomes, or in mitochondrial and chloroplast DNA). Generally, even for nuclear genes, the variance in offspring number of males is far greater than that of females, and one particular consequence is that the effective size for the Y chromosome is much smaller than the theoretical value of $N_m/2$ implied by Problem 3.8.

Effective Size of a Subdivided Population

Finally we will consider a model in which a population is subdivided into D subpopulations (**demes**), each consisting of N diploid individuals, with migration among demes measured by a quantity m equal to the probability that a randomly chosen allele in any deme originates from one of the remaining $D - 1$ demes. The population subdivision creates a situation in which two levels of random drift take place simultaneously. There is a drift process within each deme, which takes place relatively rapidly, and another drift process in the population as a whole, which takes place more slowly. Since the mathematics is somewhat rough going (Wakeley 1999, 2000), we shall present only the main result, which is that, when D is reasonably large, the effective population size of the entire population is given by

$$N_e = ND\left(1 + \frac{1}{4Nm}\right) \tag{3.32}$$

In this equation, the factor ND comes from the within-deme phase of the random genetic drift, and the factor $1 + 1/(4Nm)$ comes from the among-deme phase. An interesting and important feature of the model is that, unless $4Nm$ is very large, the effective population size (N_e) is larger than the actual population size (ND). This seeming paradox results from the population subdivision. When there are many demes connected by low rates of migration, then even if one knew which allele in some deme is destined ultimately to become the common ancestor of all alleles in the population at some future time, the process by which this lucky allele spreads among the