1 Allele Frequencies, Genotype Frequencies, and Hardy–Weinberg Equilibrium

MOST READERS OF THIS BOOK will be familiar with the terminology of genetics. But since some terms are defined slightly differently in population genetics than in other areas of genetics and molecular biology, some definitions might be useful at the outset. A **locus** (plural: *loci*) is a position in the genome where there might be one or more alleles segregating. Some geneticists use the word *locus* as synonymous to *coding gene*. However, in population genetics, the word *locus* is generally used to represent *any position in the genome*. It could be a coding gene, such as the *MC1R* gene; it could be a microsatellite; or it could be a single nucleotide position in the genome, such as position 8,789,654 of chromosome 1 of the human genome. In general, any unit in the genome with one or more alleles is a locus. A **genotype** is the combination of alleles carried by an individual in a particular locus. For example, if an individual is homozygous TT in position 8,789,654 of chromosome 1 of the human genome, then we say that this individual has genotype TT at that locus. A diploid species, such as humans, has two copies of all its chromosomes. For a collection of N diploid individuals, there are $2N$ gene copies at each locus, and there could be one or more alleles.

A major objective of classical population genetics is to understand how allele frequencies change through time. To simplify the analyses of allele frequencies, we often use models where there are two alleles—say, allele A and allele a. We call such models **di-allelic models**. The two alleles could, for example, represent the normal and the red-hair ver-

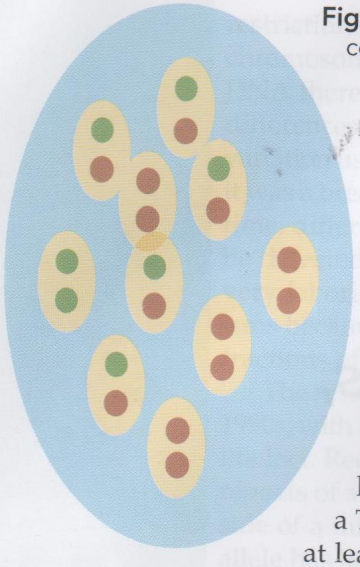

Figure 1.1 A hypothetical population with $N = 10$ individuals, 20 gene copies, and a total of 7 copies of allele *A* (green) and 13 copies of allele *a* (red), i.e., $f_A = 7/20$ and $f_a = 13/20$. The genotype frequencies are $f_{AA} = 1/10$, $f_{Aa} = 5/10$, and $f_{aa} = 4/10$.

sion of the *MC1R* gene discussed in the introduction— or two different versions of any other gene. Di-allelic models can also be used to model DNA sequences. At any position in the genome, there are four possible nucleotides, A, C, T, and G, but because mutations are rare in most organisms, you will typically tend to see at most two nucleotides in any particular position in the individuals of the population. For example, in nucleotide position 478 of the *MC1R* gene in humans, most individuals have a C, but some individuals have a T; A and G have not been observed in this position. So we can, at least as a first approximation, use a di-allelic model to describe this position in the genome.

We sometimes depict a population as in **Figure 1.1**. The blue oval represents the population, and the tan ovals within it represent individuals. The red and green balls within the individual ovals represent two alleles segregating in the population, alleles *A* and *a*. The combination of alleles within each tan oval represents the genotype of an individual; thus, an oval with a green and a red ball represents a heterozygous individual of genotype *Aa*.

Allele Frequencies

The frequency of an allele is defined as the number of copies of the allele in the population divided by the total number of gene copies in the population. In a diploid population (in which all individuals carry two copies of each chromosome) with N individuals, there are $2N$ gene copies. So the frequencies of alleles *A* and *a* are:

$$f_A = \frac{N_A}{2N} \text{ and } f_a = \frac{N_a}{2N} \quad (1.1)$$

where N_A and N_a are the numbers of *A* and *a* alleles segregating in the population, respectively. Of course, the allele frequencies must add up to 1, so $f_A + f_a = 1$. Much population genetic theory concentrates on describing the changes of f_A and f_a with time. If we can describe how we expect allele frequencies to change through time, we have learned a great deal about evolution.

Genotype Frequencies

The allele frequencies in the population can be calculated from the genotype frequencies. In a di-allelic locus, there are three possible genotypes: *AA*, *Aa*,

and *aa*. If the number of copies of genotypes *AA*, *Aa*, and *aa* are N_{AA}, N_{Aa}, and N_{aa}, respectively, then the genotype frequencies are:

$$f_{AA} = \frac{N_{AA}}{N} \qquad f_{Aa} = \frac{N_{Aa}}{N} \qquad f_{aa} = \frac{N_{aa}}{N} \tag{1.2}$$

Notice that while the denominator in Equation 1.2 is N, the denominator in Equation 1.1 is $2N$, as there are $2N$ gene copies in a diploid population of N individuals. The genotype frequencies will add up to 1: $f_{aa} + f_{Aa} + f_{AA} = 1$.

Individuals of genotype *AA* carry two copies of allele *A* and individuals of genotype *Aa* carry one copy of allele *A*. The allele frequency of allele *A* can, therefore, be calculated as:

$$f_A = \frac{2N_{AA} + N_{Aa}}{2N} = f_{AA} + f_{Aa}/2 \tag{1.3}$$

Similarly, $f_a = f_{aa} + f_{Aa}/2$. The proportion of individuals that are heterozygous in the population (f_{Aa}) is called the **heterozygosity** of the population. The proportion that is homozygous ($1 - f_{Aa} = f_{AA} + f_{aa}$), is the **homozygosity** of the population.

K-allelic Loci

A locus in which there are k different alleles, where k could be any positive natural number, is usually referred to as a ***k*-allelic locus**. Microsatellite loci often have more than two alleles. We can find expressions for allele and genotype frequencies for a general k-allelic locus similar to the ones we have already found for a di-allelic locus. For an allele, $i \in \{1, 2, \ldots, k\}$, with N_i copies in the population, the allele frequency is $f_i = N_i/2N$, and for a genotype ij ($= ji$), the genotype frequency is $f_{ij} = N_{ij}/N$. The allele frequency can then be calculated from the genotype frequencies as:

$$f_i = f_{ii} + \sum_{j:j \neq i} f_{ij}/2 \tag{1.4}$$

The concepts of homozygosity and heterozygosity can also be extended to k-allelic loci, with $\sum_i f_{ii}$ being the homozygosity and $\sum_{(i,j):i<j} f_{ij}$ being the heterozygosity. In this book we will mostly concentrate on di-allelic loci, because the mathematical notation is simpler for such loci. However, much of the theory discussed easily extends to loci with more than two alleles.

Example: The MC1R Gene

Let us again consider position 478 of the *MC1R* gene. Suppose we obtain a random sample of 30 individuals from the United States and find 25 individuals of genotype CC, 5 individuals of genotype CT, and 0 individuals of genotype TT. The genotype frequencies can then be estimated as $f_{CC} = 25/30 = 0.833$; $f_{CT} = 5/30 = 0.167$; and $f_{TT} = 0/30 = 0$. The allele frequencies can be estimated as $f_C = 0.833 + 0.167/2 = 0.917$ and $f_T = 1 - 0.917 = 0.083$.

Notice here that we used the word *estimated*. We cannot know the true genotype or allele frequencies in the entire population without examining all the individuals in the population, but we can hope that this sample of 30 individuals is representative. Had we taken another sample of 30 different individuals, we might have obtained a slightly different answer.

Hardy–Weinberg Equilibrium

We have seen how allele frequencies can be calculated from genotype frequencies. But can we also predict genotype frequencies from allele frequencies? For example, knowing that the frequency of T in position 478 of the *MC1R* locus is approximately 0.08, what proportion of the population would we expect to have genotype TT?

We can answer this question, but only if we make some assumptions. One particularly useful simplifying assumption is that mating is random, i.e., that individuals mate with each other without regard to genotype. Imagine a pool of parental males and a pool of parental females that mate randomly, i.e, the next generation is produced by randomly choosing the father and the mother from these pools of potential parents independently of each other for each individual in the offspring generation. For now, assume that the allele frequency among males is the same as among females, and that there are only two alleles, A and a, for the locus under consideration. Given these assumptions, the chance that an individual offspring is of genotype AA is given by the probability of receiving an A allele from the father and an A allele from the mother. The probability that an A allele is transmitted to the next generation is simply the frequency of the allele, f_A, because all gene copies have the same probability of transmission under Mendel's First Law. The assumption of random mating ensures that we can multiply the probabilities from the father (f_A) and the mother (f_A), so the probability that an individual in the population is of type of AA is simply f_A^2.

Likewise, an individual offspring can be heterozygous by getting an A allele from the father and an a allele from the mother—or by getting an a allele from the father and an A allele from the mother. The probability than an individual is of genotype Aa is then $f_A f_a + f_a f_A = 2f_A f_a$. Finally, using the same logic, we find that the probability that an individual is homozygous, aa, is f_a^2. The expected proportion of individuals of a particular genotype

TABLE 1.1 Genotype frequencies under Hardy–Weinberg Equilibrium

Genotype	AA	Aa	aa
Frequency	f_A^2	$2f_A f_a$	f_a^2

in the population is simply the genotype probabilities we have calculated, and we have arrived at the famous Hardy–Weinberg equilibrium theory: The **expected homozygosity** in the population is then $f_a^2 + f_A^2$ and the **expected heterozygosity** is $2f_A f_a$.

The reader may previously have encountered Hardy–Weinberg Equilibrium (HWE) theory using the notation p^2, $2pq$, and q^2 for the three genotype probabilities, respectively. Notice that this result is exactly the same as that stated in **Table 1.1**, with f_A replaced by p and f_a replaced by q. We use our notation because it generalizes more easily. As required, the genotype frequencies under HWE will add up to 1:

$$f_A^2 + 2f_A f_a + f_a^2 = (f_A + f_a)^2 = 1 \tag{1.5}$$

The concept of *probability* used here to derive HWE is discussed in **Box 1.1**. Box 1.1 also discusses the concept of *independence*. The reader may notice that the assumption of random mating implies that we draw alleles independently from male and female parents, allowing us to multiply the allele frequencies together in the offspring population. In terms of the notation from Box 1.1, we could write:

$$\begin{aligned} &\Pr(\textit{offspring genotype} = AA) \\ &= \Pr(\textit{paternal allele} = A) \times \Pr(\textit{maternal allele} = A) \\ &= f_A f_A = f_A^2 \end{aligned} \tag{1.6}$$

While the basic ideas in Box 1.1 are not prerequisite to an understanding of HWE, they will be used throughout this book, and should be reviewed at this point if they are not already familiar.

An alternative derivation of HWE, based on enumerating all possible matings, is shown in **Box 1.2**. We obtain the same result using that approach, demonstrating that random mating is, in fact, equivalent to independent sampling of paternal and maternal alleles.

Finally, notice that random mating in itself does not change the allele frequencies. The frequency of allele A in the next generation (f_A')

$$f_A' = f_A^2 + 2f_A f_a/2 = f_A^2 + 2f_A(1 - f_A)/2 = f_A \tag{1.7}$$

will be the same as in the previous generation.

The MC1R *Gene Revisited*

Now let's revisit the question regarding prediction of genotype frequencies in position 478 of the *MC1R* locus. With an allele frequency of 0.08 of allele T in the US population, how many TT homozygotes might we expect? Using HWE theory we will expect the proportions of individuals with genotypes CC, CT, and TT to be $0.92^2 = 0.8464$, $2 \times 0.92 \times 0.08 = 0.1472$, and $0.08^2 = 0.0064$, respectively. Part of the interest in this gene is caused by the fact that individuals with the TT genotype will likely have red hair (Introductory Figure). However, a much larger proportion of the population has red hair

BOX 1.1 Probability and Independence

Although there are different schools of thought regarding definitions of **probability**, we will here think of probability as expressing belief in future events, or outcomes of an experiment. For example, if we toss a coin and make the statement, "The probability of observing a head is ½ and the probability of observing a tail is ½," then we believe heads and tails are equally likely to occur in the next toss. Let X be a variable that indicates the outcome of the coin toss. The variable X can take two different values: H for heads and T for tails. We can then write

$$\Pr(X = H) = ½$$

where $X = H$ denotes the event that the coin toss results in a H. A variable such as X, that can take on different values with different probabilities, is called a **random variable**. In words, we can read the equation above as: "The probability that the random variable X takes on the value H equals one-half," a mathematical way of saying that we think heads and tails are equally likely outcomes of the coin toss.

The **sample space** of a random variable is the set of possible values that the random variable can take on. In the coin-toss case, the sample space is $\{H, T\}$.

Two random variables are **independent** if the outcome of one variable does not affect (our belief in) the outcome of the other variable. For example, if we toss a coin twice, it is reasonable to assume that the result from the first coin toss does not affect the second coin toss, so the two coin tosses are independent of each other. Just because the first coin toss resulted in an H does not mean that we think the next coin toss also will result in an H—as long as the coin is not biased.

If two random variables are independent, we can multiply their probabilities. In the coin toss example, if we let X be the result of the first coin toss, and Y be the result of the second coin toss, we find that the **joint probability** is:

$$\Pr(X = H \text{ and } Y = H) = \Pr(X = H) \times \Pr(Y = H) = 0.5 \times 0.5 = 0.25$$

However, imagine a bag full of fake coins that are biased, half of which give H with probability 0.9 and half of which give T with probability 0.9. If we randomly pick a coin from this bag and toss it twice, these two coin tosses are correlated (not independent). The chance that the first coin toss gives H is still 0.5, because half of the coins in the bag are biased toward H and half are biased toward T. However, if the first coin toss gives H, it is likely that we have picked an H-biased coin, and our belief that the second coin toss will also result in H has increased. The two coin tosses are not, mathematically speaking, independent, and the joint probability of observing an H on both the first and the second coin toss is no longer 0.25. We can no longer obtain the joint probability from the two coin tosses by multiplying the probabilities from each coin toss. These concepts are expanded upon in Appendix A.

BOX 1.2 Derivation of HWE Genotype Frequencies

In the text, we derived the Hardy–Weinberg genotype frequencies by assuming that gametes inherited from the mother and father assorted independently. We derive the frequencies here by considering all possible matings, show in the table below. The frequency of each type of mating is the product of the genotype frequencies. That is what is meant by **random mating**. The genotypes of the offspring then follow from Mendel's First Law (random gamete assortment).

Mother	Father	Frequency	Offspring		
			AA	*Aa*	*aa*
AA	*AA*	f_{AA}^2	1	0	0
AA	*Aa*	$f_{AA}f_{Aa}$	½	½	0
AA	*aa*	$f_{AA}f_{aa}$	0	1	0
Aa	*AA*	$f_{Aa}f_{AA}$	½	½	0
Aa	*Aa*	f_{Aa}^2	¼	½	¼
Aa	*aa*	$f_{Aa}f_{aa}$	0	½	½
aa	*AA*	$f_{aa}f_{AA}$	0	1	0
aa	*Aa*	$f_{aa}f_{Aa}$	0	½	½
aa	*aa*	f_{aa}^2	0	0	1

The genotype frequencies in the offspring are found by adding over all the families:

$$f'_{AA} = (1)f_{AA}^2 + (½)f_{AA}f_{Aa} + (½)f_{Aa}f_{AA} + (¼)f_{Aa}^2 = (f_{AA} + (1 - f_{Aa}/2)^2 = f_A^2$$

$$f'_{Aa} = (½)f_{AA}f_{Aa} + (1)f_{AA}f_{aa} + (½)f_{Aa}f_{AA} + (½)f_{Aa}^2 + (½)f_{Aa}f_{aa} + (1)f_{aa}f_{AA} + (½)f_{aa}f_{Aa}$$
$$= 2(f_{AA} + f_{Aa}/2)(f_{aa} + f_{Aa}/2) = 2f_Af_a$$

$$f'_{aa} = (¼)f_{Aa}^2 + (½)f_{Aa}f_{aa} + (½)f_{aa}f_{Aa} + (1)f_{aa}^2 = (f_{aa} + f_{Aa}/2)^2 = f_a^2$$

where the prime (′) indicates the frequencies among the offspring. No matter what the genotype frequencies are, one generation of random mating will establish the HWE genotype frequencies.

than the expected 0.64% from this calculation, telling us that other factors are important for the development of red hair than being homozygous TT at position 478 of the *MC1R* locus.

Tay–Sachs Disease

HWE has many applications, including analysis of allele frequencies that impact health in humans: the frequency of individuals affected by diseases caused by recessive deleterious mutations can be predicted from the allele frequencies. An example is Tay–Sachs disease, which causes deterioration of mental and physical abilities and usually ends in death by the age of

four. Individuals homozygous for certain mutations in the *HEXA* gene will be affected by this disease. A four-base-pair insertion in the gene, causing a change in reading frame that essentially destroys the function of the gene, is common among Ashkenazi Jews. In fact, the allele frequency of this mutation among Ashkenazi Jews is as high as 2%. What is the proportion of offspring of Ashkenazi Jewish couples that will be affected by Tay–Sachs disease because they are homozygous for the disease mutation? Using HWE, we find the answer to be $0.02^2 = 0.0004$ or 0.04%. This disease risk is sufficiently high that Ashkenazi Jewish couples in the United States and Israel are often genetically screened for Tay–Sachs Disease.

Extensions and Generalizations of HWE

HWE shows that if the allele frequencies are identical in males and females, after one round of random mating, the genotype frequencies can be obtained simply by multiplying together the appropriate allele frequencies. If the allele frequencies are different in males and females, it takes two generations before HWE is established. After one generation of random mating, the allele frequencies in males and females will become the same. The next generation of random mating then establishes HWE. (The demonstration of this principle is left as an exercise at the end of the chapter.) In real populations, there is no real reason to expect that allele frequencies are initially different in males and females, and any observed deviations from HWE are unlikely to be caused by this very transient effect.

HWE can also be generalized to loci with more than two alleles. Imagine a k-allelic locus with allele frequencies $f_1, f_2, \ldots, f_k$, assumed to be equal among males and females. After one generation of random mating, the genotype frequencies can be obtained by multiplying the appropriate allele frequencies together. So the expected genotype frequency of homozygous individuals with genotype ii is f_i^2 for any allele i, and the genotype frequencies of heterozygous individuals with genotype ij is $2f_if_j$, for any pair of (different) alleles i and j.

Deviations from HWE 1: Assortative Mating

There are many factors that can cause deviations from HWE equilibrium. First, mating may not be random with respect to genotype. For example, individuals may be more likely to mate with other individuals of the same, or similar, genotype. This is called **assortative mating**. Clearly, if *AA* individuals prefer to mate with other *AA* individuals, *aa* individuals prefer to mate with other *aa* individuals, and *AA* and *aa* individuals rarely mate, there will be fewer heterozygous individuals in the next generation than predicted by HWE. For example, consider a population initially in HWE with an allele frequency of $f_A = 0.5$ and genotype frequencies $f_{AA} = 0.25$, $f_{Aa} = 0.5$, and $f_{aa} = 0.25$. If the population then undergoes one generation of strong assortative mating in which individuals only mate with other indi-

viduals of the same genotype, the genotype frequency of the *AA* genotype will become $f_{AA} = 0.25 + 0.25 \times 0.5 = 0.375$. All offspring of $AA \times AA$ matings (25% of all matings) will be of type *AA* and a quarter of all offspring of $Aa \times Aa$ matings (50% of all matings) will be of type *AA*. Using similar arguments we can also find the frequency of *aa* offspring to be $f_{Aa} = 0.375$, and the frequency of heterozygous offspring will then be $f_{Aa} = 1 - f_{AA} - f_{aa} = 0.25$. The allele frequency is still $f_A = 0.5$ in this example, but there are now only half as many heterozygous individuals as under HWE. If this processes continues for many generations, the population will eventually become entirely depleted of heterozygous individuals.

The opposite situation, where individuals prefer not to mate with individuals of their own genotype, is called *negative assortative* mating or **dis-assortative mating**. Dis-assortative mating can result in numbers of heterozygous individuals in excess of those expected under HWE.

Deviations from HWE 2: Inbreeding

Another mating pattern that can cause deviations from HWE is **inbreeding**. Inbreeding occurs as a result of matings between individuals that are related because they have one or more ancestors in common. The effect of such matings is very much the same as for assortative mating. If these matings are more common than expected under random mating, the proportion of heterozygous individuals will be smaller than under HWE. An extreme type of inbreeding occurs when organisms reproduce by self-fertilization, as many plants do. This type of inbreeding will quickly cause strong deviations from HWE. Assortative mating and inbreeding have similar effects on genotype frequencies: they both increase the proportion of homozygous individuals. The difference is that inbreeding affects the whole genome, while assortative mating affects only those loci that determine the trait or traits that affect mating preference. Assortative mating does not affect genotype frequencies at other loci.

In the early population genetic literature, deviations from HWE were often thought to be a consequence of inbreeding in one way or another. For this reason, we measure deviations from HWE in terms of an inbreeding coefficient (F). We will discuss the inbreeding coefficient in more detail a little later in this chapter.

Deviations from HWE 3: Population Structure

When deriving the HWE theory, we assumed that parents were sampled at random from a population. But what if the population were structured so that it really contained two or more subpopulations? Imagine, for example, a species of lizards inhabiting different islands in the Caribbean.

If we obtained a sample from multiple islands, ignoring this structure of the population, it clearly could not be true that the individuals in the sample had been produced by random mating: individuals from different islands are not likely to mate with each other. Consider the extreme case

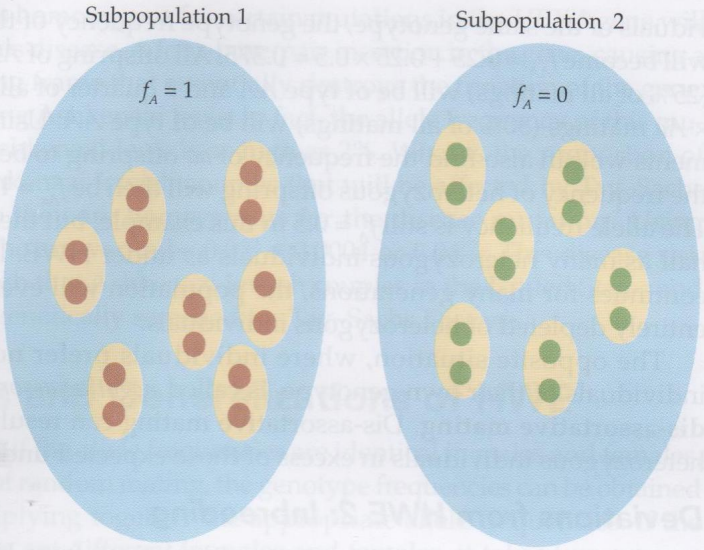

Figure 1.2 Two subpopulations with allele frequencies $f_A = 0$ and 1, respectively. In the combined population, obtained by pooling individuals from subpopulation 1 and subpopulation 2, all individuals are homozygous and there is an apparent deficit of heterozygous individuals compared to the HWE expectation.

where there are two subpopulations, subpopulation 1 and 2, and the frequency of allele *A* in subpopulation 1 is 100%, while in subpopulation 2, it is 0% (**Figure 1.2**). Even if there is random mating within subpopulation 1 and within subpopulation 2, all individuals will be of either genotype *AA* (subpopulation 1) or *aa* (subpopulation 2). The combined population will very much be out of HWE because it contains only homozygous individuals. Clearly, if there are more than one subpopulation within a larger population (**population structure**), there may be deviations from HWE. This is also true in less extreme cases where allele frequencies differ only marginally between subpopulations. Deviations from HWE will also arise when there are no discrete subpopulations but a continuous spatial distribution of individuals, or in cases when only one subpopulation has been sampled but this subpopulation occasionally receives migrants from another subpopulation. The effect is quite general and is not specific to any particular model of population structure. In real populations, population structure and inbreeding are likely the most important reasons for observations of deviations from HWE. Even relatively small differences in allele frequencies in different subpopulations can cause deviations from HWE. The effects of population structure on deviations from HWE will be discussed in more detail in Chapter 4.

Deviations from HWE 4: Selection

Natural selection occurs when there is differential survival or reproduction among individuals due to their genotypes. It is of such importance in population genetics that we devote three chapters to it. For now, suffice it to say that natural selection also can cause deviations from HWE. Take,

for example, the genotype frequencies in the HEXA gene among adults. As individuals homozygous for this disease-causing mutation die before they reach adulthood, the adult population must be slightly out of HWE with a modest excess of heterozygotes. At most, 0.04% of the population is affected by disease, so you would need to examine many thousands of individuals to actually detect this deviation from HWE. Most of the time, we do expect natural selection to be strong enough in humans to cause very severe deviations from HWE. Also worth noting is that deviations from HWE due to selection only can be detected if the population is sampled after selection has been acting. In the case of Tay–Sachs, we do not expect natural selection to cause deviations from HWE among infants.

Some geneticists also include effects of small population sizes and mutations among forces that can cause deviations from HWE. However, as the effect of these factors are extremely small and cause only small random deviations from HWE that do not accumulate over time, we do list them among forces that can cause deviations from HWE.

The Inbreeding Coefficient

Although factors other than inbreeding (such as selection) can cause deviations from HWE, the most common statistic we use to measure deviations from HWE is called the *inbreeding coefficient* (F). To further confuse students, population geneticists have a bad habit of using F to describe the degree to which heterozygosity is reduced both in individuals and in populations as a result of inbreeding. In this book we will use F solely to denote the decrease in heterozygosity in a population beyond that expected under HWE. For a di-allelic locus, we define F as:

$$F = \frac{(2f_A f_a - f_{Aa})}{2f_A f_a} \tag{1.8}$$

Notice that the first term in the numerator, $2f_A f_a$, is the proportion of individuals expected to be heterozygous under HWE. So F measures the difference between the expected and the observed heterozygosity, standardized by the expected heterozygosity. If $F = 0$, the population is in HWE, and if $F = 1$, there are no heterozygotes in the population. Also notice that if there are more heterozygotes than expected under HWE, F is negative.

By rearranging Equation 1.4, we find:

$$f_{Aa} = 2f_A f_a(1-F) \tag{1.9}$$

which shows that, with this definition, the proportion of heterozygotes in the population is reduced by a factor F from that expected under HWE. If we know the value of F, and the allele frequencies, we can predict the proportion of heterozygote individuals in the population without assuming HWE.

Many plant species are predominantly self-fertilizing and in those species, genotype frequencies are typically far from HWE. For example, in a

Figure 1.3 The flower of wild oats (*Avena fatua*) has both male and female reproductive organs (stamens and pistils) and is capable of self-fertilization, which leads to high levels of inbreeding. Many plants are capable of self-fertilization, but many are not, because they are dioecious (having male and female flowers on separate plants) or because they have evolved other mechanisms to avoid self-fertilization— for example, by separating the flowering times of male and female flowers on the same plant or by evolving genetic self-incompatibility.

population of wild oats, *Avena fatua* (**Figure 1.3**), the genotype frequencies at one locus were found by Marshall and Allard to be $f_{AA} = 0.58$, $f_{Aa} = 0.07$, and $f_{aa} = 0.35$, which obviously deviates from HWE. This species is self-fertile and extensive self-fertilization accounts for the lower frequency of heterozygotes. We can calculate F for this species using the formulas given above. We first find the allele frequencies as $f_A = 0.58 + 0.07/2 = 0.615$, $f_a = 1 - 0.615 = 0.385$. We then find $F = (2 \times 0.385 \times 0.615 - 0.07)/(2 \times 0.385 \times 0.615) = 0.852$.

Testing for Deviations from HWE

If we take a sample from a population, we may randomly tend to get a few more homozygotes or heterozygotes than expected under HWE, even though the population actually is in HWE. To determine if the population is out of HWE, we need a formal statistical test. In such a test, we wish to test the null hypothesis that genotype frequencies follow those predicted by HWE (e.g., Table 1.1 in the di-allelic case). One way of doing this is to use a chi-square test (**Box 1.3**). To perform a chi-square test, we need to obtain the observed and expected values, and to find the degrees of freedom. The genotype counts in the data are the observed values. The expected values are given by the HWE theory and can be calculated by the allele frequencies. There is just one degree of freedom, because there are three categories and two constraints. The first constraint is the same as in the coin toss example in Box 1.1: the genotype counts must add to the total number of observations. The second constraint comes from the fact that the allele frequencies under the expected genotypes should equal the observed allele frequencies.

As an example, consider a locus with the following genotypic counts for forty individuals: $N_{AA} = 20$, $N_{Aa} = 10$, $N_{aa} = 10$. The genotype frequencies are $f_{AA} = ½$, $f_{Aa} = ¼$, and $f_{aa} = ¼$ and the allele frequencies are then $f_A = ½ + (¼)/2 = ⅝$ and $f_a = ¼ + (¼)/2 = ⅜$. We next need to find the expected

BOX 1.3 The Chi-Square Test

A chi-square test, in the definition used in this book, is used to test the goodness-of-fit of a model using **categorical data**—data that can be presented as the counts of different types of observations, such as the number of different alleles or the number of different genotypes. It also assumes we have a null-hypothesis model that predicts the expected frequencies of each count. It is this model we wish to test. If the observed counts are so different from the expected counts that they cannot be attributed to chance, then the null hypothesis can be rejected (we no longer believe that model to be true). Assume there are k categories of observations, and let the observed counts be $O_1, O_2, \ldots O_k$, and the expected counts under the model be $E_1, E_2, \ldots E_k$. The chi-square test statistic is then calculated as

$$\chi^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$$

If χ^2 is very large, it means that we can reject the null model because the observed and expected counts are more different from each other than expected by chance. But how do we figure out if χ^2 is sufficiently large to reject the null model? It turns out that standard statistical theory shows that, for large amounts of data (under suitable assumptions), χ^2 follows a chi-square distribution with degrees of freedom equal to $k - p$, where p is the reduction in the degree of freedom due to constraints imposed by the model when calculating the expected values. A chi-square test is performed by calculating χ^2, calculating p, and then comparing the value of χ^2 to a chi-square distribution with $k - p$ degrees of freedom. Chi-square distributions with different degrees of freedom are given in Appendix D.

As an example, imagine that we are interested in testing the null hypothesis that a coin is fair, i.e., that it produces H and T each with probability 0.5 (see Box 1.1). To test this, we toss a coin 50 times and get 29 H and 21 T. Does this show that the coin is biased (not fair)? The expected numbers under the null model of a fair coin are clearly $E_1 = 25$ and $E_2 = 25$, so we get

$$\chi^2 = \frac{(25-29)^2}{25} + \frac{(25-21)^2}{25} = 1.28$$

In this case, the number of categories is $k = 2$, and the only constraint we have on the counts of H and T is that they should sum to 50, implying that $p = 1$, so there is one degree of freedom. Consulting the table in Appendix A we find that the probability of observing a value of $\chi^2 = 1.28$ or larger is close to 0.25. To reject the null model, this probability would need to be much smaller, say less than 0.05, or less than 0.01, so in this case we cannot reject the null hypothesis that the coin is fair. The cut-off value we choose for the probability is called the **significance level**. The choice of significance level is somewhat arbitrary, but most studies choose 0.05 or 0.01.

Examples of chi-square tests are given throughout this book; the first is in the section on testing HWE.

genotype counts under HWE, given the allele frequencies: $E_{AA} = 40 \times (5/8)^2 = 15.625$; $E_{Aa} = 40 \times 2 \times 3/8 \times 5/8 = 18.75$; and $E_{aa} = 40 \times (3/8)^2 = 5.625$. We then calculate the chi-square statistic (as in Box 1.3) as

$$\chi^2 = \frac{(15.625-20)^2}{15.625} + \frac{(18.75-10)^2}{18.75} + \frac{(5.625-10)^2}{5.625} = 8.711 \quad (1.10)$$

Comparing our observed value of 8.711 to the critical values for a chi-square distribution with one degree of freedom in Appendix 4, we see that the probability of observing a value this high or higher is between 0.01 and 0.001. Using a traditional significance level of 0.05 (critical value = 3.841), we find $p < 0.05$ and reject the null hypothesis of HWE. The genotype frequencies are statistically significantly different from those expected under HWE.

The chi-square test can also be extended to k-allelic loci. The hardest part is to calculate the degrees of freedom. For k alleles there are $k(k + 1)/2$ possible genotypes, i.e., categories in a chi-square test. But there are k constraints, because the allele frequencies in the expected categories have to match the observed allele frequencies. So the degrees of freedom are calculated as $k(k + 1)/2 - k = k(k - 1)/2$.

Using Allele Frequencies to Identify Individuals

The DNA from an individual can be used to identify the individual. This principle has been used extensively in many connections, most importantly in forensics where DNA is used to determine paternity and to identify someone who was at a crime scene. In the context of forensics, the use of DNA to identify individuals is called **DNA fingerprinting** or **DNA profiling**. In the United States, thirteen microsatellite loci are usually used in forensics. An individual matches a DNA profile if the genotype is identical to the profile at all thirteen loci. But with only thirteen loci, there is some chance than an individual will match a profile by chance alone. To assess the probability (Box 1.1) of a random match, forensic scientists compare the profile to a database of allele frequencies. If the individual carries two alleles for a locus, say allele 1 and allele 2, then the **match probability** is simply $2f_1f_2$ for a heterozygous individual, and f_1^2 or f_2^2 for a homozygous individual, assuming HW equilibrium. The probabilities calculated for all loci are then multiplied together to provide one final match probability.

There are several problems that arise in the interpretation of match probabilities based on databases. First, the database may not be representative for the population to which the individual belongs. For example, a database of Caucasian individuals may not be appropriate as a reference for an individual from a non-Caucasian background. For this reason, the United States and many other countries have devoted significant efforts

to developing large representative databases. Second, the individual may have siblings or other close relatives who also have a high probability of matching the profile. Third, assumptions regarding HW equilibrium and simple multiplication of probabilities among loci may not always be valid. Considerable statistical research has been devoted to these concerns.

References

*Chen J., 2010. The Hardy–Weinberg principle and its applications in modern population genetics. *Frontiers in Biology* 5: 348–353.

*Evett I. W. and Weir B. S., 1998. *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.* Sinauer, Sunderland, MA.

Marshall D. R. and Allard R. W., 1970. Maintenance of isozyme polymorphism in natural populations of *Avena barbata*. *Genetics* 66: 393–399.

*Valverde P., Healy E., Jackson I., et al., 1995. Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature Genetics* 11: 328–30.

*Recommended reading

EXERCISES

1.1 A researcher examines a locus in which there is a particular C/T polymorphism. She obtains the following genotypic counts: CC: 42, CT: 16, TT: 32. Calculate the genotype frequencies and the allele frequencies in the sample.

1.2 For the data from Exercise 1.1, find the expected homozygosity and the expected heterozygosity, given the observed allele frequencies, and calculate the inbreeding coefficient (F).

1.3 For the data in Exercise 1.1, test if the population is in HWE using a chi-square test at the 5% significance level.

1.4 The proportion of a population suffering from a specific rare genetic disease is 0.02%. Assume that the disease is caused by a single recessive allele and assume that the population is in HWE. How many individuals carry the disease allele in the heterozygous state?

1.5 In another locus there are three alleles—A, C, T—and the genotypic counts in the sample are AA: 10, AC: 10, AT: 5, CC: 20, CT: 5, and TT: 20. Calculate the genotype frequencies and the allele frequencies in the sample.

1.6 For the data from Exercise 1.5, find the expected homozygosity and the expected heterozygosity, given the observed allele frequencies.

1.7 For the data in Exercise 1.5, test if the population is in HWE, using a chi-square test at the 5% significance level.

1.8 An individual has genotype CT for the locus discussed in Exercise 1.1, and genotype AC in the locus discussed in Exercise 1.5. At a crime scene, forensic evidence is found with the exact same (TT, CC) genotype. What is the chance of such a match by random, assuming HWE and the allele frequencies calculated in Exercises 1.1 and 1.5? What is the match probability if the calculation is done using observed genotype frequencies instead?

1.9 Show mathematically that it takes two generations to achieve HWE when the allele frequencies differ between males and females (assume a di-allelic locus).