Chapter 2

Population Genetics Models

2.1 Introduction

This chapter is an introduction to elementary mathematical models of population genetics, the most mathematical of the traditional disciplines of biology. The central object of study is a population of inter-breeding individuals of the same species, considered as it evolves in time. It is a fundamental empirical fact that individuals belonging to natural populations vary in their many different phenotypic (that is, observable) traits—size and shape, external markings, fecundity, disease resistance, blood type, etc.; everyday experience with the diversity of individuals in human populations will provide you with a ready example. The inherited component of an individual's traits depends on its *genotype*, which, as defined in chapter 1, is the specification of the alleles it carries at gene loci. The phenotypic variation in a population reflects variation in its genotypes. Population genetics is concerned mainly with genotypic variation. It applies what is known about the mechanism of heredity—Mendel's laws, the molecular basis of heredity—to understanding how genetic variation varies in time, as one generation succeeds another, under the influence of selection, mutation, and the randomness inherent in mating.

A population genetics study will typically focus on only a small number of genetically based traits, for example, eye color in humans, or color and texture of seeds, as in Mendel's pea experiment. Thus, if you were to participate in a genetic study of eye color, all other distinctive features that characterize you—your good looks, superior intelligence, and athletic prowess—would be

ignored and you would be entered into the data set only as an eye color (and, probably, a sex, as inheritance of many traits is sex-linked). The underlying model would classify individuals only by their genotype at the alleles influencing eye color. Thus in population genetics models, individuals are not considered in their concrete entirety, but rather, are identified with the list of their alleles at the small number of loci under consideration.

The basic quantities of interest in population genetics are genotype and allele frequencies. The frequency of a genotype in a population is simply the total number of individuals with the genotype divided by the population size. Because we can recover the number of each genotype from its frequency and the population size, the collection of genotype frequencies characterizes the genotype population. The degree of genotypic variability is revealed in the distribution of genotype frequencies. If the frequency of one genotype is close to 1 and the frequencies of all others are small, there is not much variation. By contrast, in highly variable populations the different genotypes tend to have about equal frequencies.

At the quantitative level, population genetics seeks to explain the genotype frequencies observed in nature and to understand the mechansims by which they arise and are maintained. This study requires model building and analysis, and it is here that mathematics, and, especially, probability, enter. There is a random component to the matings and to the survival of the consequent offspring that produce the genotypes of successive generations. The modeller makes assumptions about how genes are passed down over the generations and translates them into precise mathematical rules. Then the implications of these rules are analyzed and compared to data.

In this chapter, we introduce and analyze very basic models of population genetics. To build models, a number of questions need to be addresed. What type of random mixing of genotypes do we postulate in a mating population. How is the influence of selection and mutation to be modelled? How do generational and age structures affect evolution of inheritance? Section 2 of this chapter discusses these issues as a preliminary to model formulation, which is done in section 3. Once a model is built, we need to analyze it. What does it predict about the nature and evolution of genotype frequencies? How does it help understand actual populations? These questions are taken up in the last sections of the chapter.

As you read, you will of course be trying to master the models and the probabilistic reasoning behind their formulations. But, it is just as important to pay attention to the modelling process itself. Elementary population

genetics provides an excellent example of the art of mathematical modelling. First one begins with extremely oversimplified models; nevertheless, they are easy to analyze and lead already to interesting conclusions, in particular, the Hardy-Weinberg equilibrium concept. Then, one explores what happens as the simplifications are removed, one at time. Do the conclusions of the simplest model continue to hold, at least in modified form? Or do very different behaviors arise? This process—building simplified models, testing their behavior and sensitivity to assumptions, modifying them—is a paradigm for mathematical modelling. There is a tension in modelling between precision in the assumptions, so that they account for possible subtle effects, and abstraction and simplification. The precise model will, it is hoped, be more accurate, but it will also be more complicated, so much so that it cannot be easily analyzed. In a simplified model, on the other hand, it will be easier to understand the consequences of each assumption. If the simplified model accounts for the behavior of the natural system at least broadly, that is evidence that it captures something of the essential mechanisms at work, even if it is inaccurate in details. The simplified model can then serve as the basis for refined models. We will do enough in this chapter so that you can see how probabilistic principles are used to create models and how the models are refined and analyzed. You will be given an opportunity to model and develop refinements in many of the exercises, which you should do an integral part of reading this chapter.

2.2 Modelling Issues

The principle considerations going into modelling are population size, how one generation succeeds another, whether a population is divided into sexes or not, the nature of randomness in mating, and the number of loci and alleles under consideration.

Sex. (And you never thought you'd see this word in a math course!) Organisms in which sexual reproduction occurs are either **monecious** or **dioecious**. Monecious (literally, *in one house*) means that each mature individual houses both male and female sex organs; plants with flowers that both contain an ovum and produce pollen provide a common example. In dioecious (*in two houses*) organisms, such as humans, individuals are either male or female, but not both. In a population of monecious organisms, any individual can mate with any other or even with itself (selfing, or self

fertilization), whereas with dioecious organisms, mating occurs only between males and females.

Genetic mechanisms of sex determination in dioecious species vary widely, and in general complicate analysis of inheritance. In most mammals, sex is determined by X and Y chromosomes. A pair of X chromosomes and no Y chromosome produces a female. A single X and a single Y determine a male. Although the X and Y act as a pair in meiosis in males, the X and Y are not copies of each other. Thus the male will only contain a single copy of the allele of any gene on the X or Y chromosome, and its genotype for such genes with therefore consist only of a single allele. The remaining chromosomes, those not involved in sex determination, are called autosomal chromosomes and are all paired in the standard manner in both males and females. Other animals have similar sex-linked chromosomes, but the details differ. Consequently, it is important to differentiate between dioecious and monecious in setting up the basic models, and, for the dioecious case, between loci on autosomal and on sex chromosomes. We will see that for autosomal chromosomes, the ultimate consequences of random mating are much the same and so the distinction does not ultimately matter. But analysis of sex-linked genes has special features.

Genotypes, genotype frequencies, and allele frequencies. We will classify models by the number of loci being studied and the number of alleles at each locus. The simplest case is that of one locus with two alleles. In this situation, the two alleles shall be denoted by A and a, and the genotypes are AA, Aa, and aa. For the purposes of population genetics models, it will be useful to think of the population as a collection of boxes, one box per individual, in each of which is placed the letters denoting the individual's genotype. Thus, in the one locus/two allele case, each box contains one of the pairs of letters, AA, Aa, or aa. (If you want an even more concrete visualization, imagine representing each A allele as a red marble and each a allele as a blue marble, and think of the population as a collection of boxes, each containing two marbles.) Next in order of simplicity is a one locus/m allele model, where $m > 2$. In this case one might label the alleles as $A_1, A_2, \ldots, A_m$; the population could then be modelled as a set of boxes, each containing a pair of letters of the form A_iA_j, where $1 \leq i, j \leq m$. The most general case considers multiple loci with multiple alleles. Rather than state general notation at this point, let us illustrate the two locus/two allele case. Denote the alleles at locus 1 by A and a and those at locus 2 by B and b.

Then the possible genotypes are the nine four letter strings, $AABB$, $AAbb$, $AABb$, $aaBB$, $aabb$, $aaBb$, $AaBB$, $Aabb$, and $AaBb$; these are the labels contained by the boxes representing the population. There is a complication in these genotype listings if the loci are on the same chromosome, but for the moment we defer further discussion.

There are two kinds of quantities we want to track in our model, genotype frequencies and allele frequencies. Consider a genotype G in a population of size N; (G here stands for a string of letters denoting all the alleles occuring at the loci under study.) Then the frequency of G is

$$f_G \triangleq \frac{\text{number of times } G \text{ occurs in the population}}{N}. \tag{2.1}$$

Recall from chapter 1 that a random sample of a population is a random draw of one individual done in such a way that each individual is equally likely to be chosen. We have the following simple, but important, equivalence:

$$f_g = \text{probability to draw } G \text{ in a random sample from the population} \tag{2.2}$$

Let A be an allele of some gene. The allele frequency is defined to be

$$p_A \triangleq \frac{\text{number of times } A \text{ occurs in all genotypes of population}}{\text{total number of alleles in population}}.$$

To clarify, think of the population in terms of boxes containing genotypes. Count the number of letters in all the boxes. This is the total number of alleles. Now count all occurences of A and divide by the total number of letters, The ratio is p_A.

Example 1. Consider the one locus/two allele case. Suppose there are 12 individuals, of whom 3 are AA, 5 are aa, and 4 are Aa. The frequency of genotype AA is $f_{AA} = 1/12 = 0.25$. Since each individual has two alleles, the total number of alleles in the population is 24. Of these 10 are A's. Thus, $p_A = 10/24 = 5/12$.

There are two probability interpretations of p_A for an allele A. First, suppose all the boxes of genotypes are emptied into one container; this just contains all the alleles without regard to whom they belong, and so we call this the allele pool. Since p_A is just the frequency of A in the container, it is also the probability of of drawing A in a random sample from the allele pool.

Consider next a different experiment. Go back to the population of genotypes. First take a random sample from the population (choose an individual

box at random). Next choose one of the alleles in the individual's genotype at random (randomly sample the contents of the chosen box). Call this experiment the random choice of an allele from a randomly chosen individual.

Theorem 1 *The allele frequency p_A for allele A, that is the probability of drawing A from the allele pool, is equal to the probability of drawing A in making a random choice of allele from a randomly chosen individual.*

This result is independent of the number of loci or alleles in the model.

We will prove this theorem in a moment. First it is useful to introduce some notation, which we will use a lot in the one locus/two allele case, and a useful relation. Let

$$K \triangleq f_{AA}, \qquad L \triangleq \frac{1}{2} f_{Aa}, \qquad M \triangleq f_{aa} \tag{2.3}$$

When we are looking at a population evolving in t, we shall use the notation $K(t)$, $L(t)$, and $M(t)$ to describe the corresponding frequencies at time t. Notice that $K + 2L + M = 1$, since the frequencies of all possible genotypes must add to 1. The following identities hold:

$$p_A = K + L \qquad \text{and} \qquad p_a = L + M. \tag{2.4}$$

To see this imagine a population of N individuals. Since each individual has 2 alleles, the size of the allele pool is $2N$. How many A alleles are there? There are NK individuals with genotype AA contributing a total of $2NK$ alleles A, and there are $N(2L)$ individuals with genotype Aa, contributing $2NL$ more A's. Hence

$$p_A = \frac{2NK + 2NL}{2N} = K + L.$$

The argument for p_a is similar, or one can observe $p_a = 1 - p_A = K + 2L + M - (K + L) = L + M$.

To prove the theorem, we shall use conditional probabilities and the rule of total probabilities—see I.4 in the Appendix. First, we give the proof in the one locus/two allele case, as the notation is simpler. Let U denote the event of drawing allele A in the experiment of first drawing and individual at random and then choosing one of its alleles at random. Let V_1 be the

event that, in the first step, an AA is drawn, let V_2 be the event that an Aa is drawn, and finally let V_3 be the event that an aa is drawn. Clearly, $\mathbb{P}(U/V_1) = 1$, because if the individual AA is drawn, allele A must be drawn. Similarly $\mathbb{P}(U/V_3) = 0$. However, if Aa is drawn, then a random sample of A, a yields A with probability 0.5, so $\mathbb{P}(U/V_2) = 0.5$. Now, by (2.2), $\mathbb{P}(V_1) = K$, $\mathbb{P}(V_2) = 2L$ and $\mathbb{P}(V_3) = M$. Putting this all together, using the rule of total probabilities and the relation in (2.4),

$$\begin{aligned}\mathbb{P}(U) &= \mathbb{P}(U/V_1)\mathbb{P}(V_1) + \mathbb{P}(U/V_2)\mathbb{P}(V_2) + \mathbb{P}(U/V_3)\mathbb{P}(V_3) \\ &= 1 \cdot K + (0.5)2L + 0 = K + L = p_A,\end{aligned}$$

which proves the claim,

The proof in the general case only requires more notation. Let the possible genotypes be labelled $G_1, G_2, \ldots, G_K$. Each G_i may stand for a string of letters denoting a genotype at a set of loci. Let M be the total number of letters required to represent the genotype. Let n_i be the number of times letter A appears in the genotype G_i—if we are dealing with diploid individuals, which will always be the assumption in these notes, n_i will be either 0,1, or 2. If there are N individuals in the population, the number of individuals with genotype G_i is Nf_{G_i} and the size of the allele pool is MN. Each G_i contributes the numbe n_i of alleles $A's$ so the frequency of allele A in the allele pool is

$$p_A = \frac{1}{MN}\sum_{i=1}^{K} Nf_{G_i}n_i = \sum_{i=1}^{K}\frac{n_i}{M}f_{G_i}. \tag{2.5}$$

Now consider the experiment of first drawing an individual and then randomly drawing an allele. Then

$$\mathbb{P}\left(\text{outcome is } A/ \text{ individual of type } G_i \text{ is drawn}\right) = \frac{n_i}{M},$$

since an individual with genotype G_i has n_i alleles A out of a total of M alleles. But,

$$\mathbb{P}\left(\text{ individual of type } G_i \text{ is drawn}\right) = f_{G_i}.$$

So by the rule of total probabilities,

$$\begin{aligned}\mathbb{P}\left(\text{outcome is } A\right) &= \\ &\sum_{i=1}^{K}\mathbb{P}\left(\text{outcome is } A/\text{type } G_i \text{ is drawn}\right)\mathbb{P}\left(\text{type } G_i \text{ is drawn}\right) \\ &= \sum_{i=1}^{K}\frac{n_i}{M}f_{G_i} = p_A.\end{aligned}$$

We next return to the two locus/two allele case. Remember that there are 9 genotypes, $AABB$, $AAbb$, $AABb$, $aaBB$, $aabb$, $aaBb$, $AaBB$, $Aabb$, and $AaBb$. In the first eight cases, in which at least one locus is homozygous, the haploid gametes will be the same whatever the arrangement of genes on chromosomes. To illustrate, let ℓ_1 label the chromosome on which the locus for A or a occurs, and let ℓ_2 label the chromosome on which the locus for B or b occurs, and assume ℓ_1 and ℓ_2 are different. Consider an individual with $AABb$ genotype. Each of its gametes gets one copy of ℓ_1, which contains only the A allele, and one copy of the ℓ_2 allele, which can contain either the B or the b allele. Therefore, the possible genotypes of the gametes are AB or Ab and these will be generated in equal numbers, on average. Now suppose the two loci are on the same chromosome, ℓ. Then, in the $AABb$ individual one copy of ℓ will have AB and the other will have Ab. Again, its gametes will end up with one or the other of these copies, with equal probability on average, and so again AB and Ab will be the possible gametic genotypes. However, the possible gametes of the ninth genotype, the double heterozygote $AaBb$, will differ depending on whether the loci are on the same chromosome or not. If they are on two different chromosomes ℓ_1 and ℓ_2, then each loci can segregate independently; the possible gametes are AB, aB, Ab, and ab, all equally probable. However, suppose the two loci are on one chromosome, ℓ. Then there are two possibilites: i) copy 1 of ℓ contains AB and copy 2 contains ab; or, ii) copy 1 of ℓ contains Ab and copy 2 contains aB. If no recombination occurs, the gametes in case i) can only be AB or ab, while the gametes in case ii) can only be Ab or aB. In each case, gametes with the other two genotypes can only be created if recombination occurs; see chapter 1. Recombination is modelled by a parameter r, which denotes the probability of a recombination event that segregates the two loci ℓ_1 and ℓ_2. It is assumed that r is the same in both cases (i) and case (ii). This seems reasonable; it says that the alleles that do appear do not affect the recombination rate. If recombination occurs in a meiosis of an $AaBb$ parent, the four resulting gametes will be AB, Ab, aB, and ab, in both cases i) and ii). Therefore if $r = 1$, that is if recombination always occurs, four different possible gametes will be produced with equal probabilities, just as if the loci were on different chromosomes.

Exercise 1. We have just considered two situations: situation 1—the two loci are on different chromosomes; and situation 2—the two loci are on the same chromosome and $r = 1$. We have just pointed out the probability of each

possible gamete type is the same, that is 0.25 in both situations. However, there is still an important difference. What is it? Hint: consider the joint distribution of the number of each gamete type in one meiotic event.

Exercise 2. Consider the case in which both loci are on the same chromosome, and let r be the probability of recombination. Consider a population of gametes produced by a parent population of double heterozygotes. Let p be the probability that a randomly selected gamete has a type i) parent (i.e., one chromosome has AB, the other ab). Then $q = 1 - p$ is the probability that the parent is type ii) (one chromosome is Ab, the other aB). Find the probability that a randomly selected gamete is AB.

Population size. In modelling, we will distinguish between two cases, finite populations and infinite populations. What does the infinite population assumption mean? It does not mean literally an infinite collection of individuals. It refers instead to frequency limits, derived from the law of large numbers, as the population size increases to infinitiy. To illustrate, suppose we are building a population of red and blue marbles by the following artificial procedure. We let $X_1, X_2, \ldots$ be an infinite sequence of independent Bernoulli random variables with $\mathbb{P}(X_i = 1) = p$ for each i. If X_i is 0, we add a red marble to the population, if X_i, we add a blue one. When the first N marbles have been chosen, the frequency of blue marbles is $\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} X_i$. This is a random variable—in fact, we know $\sum_1^N X_i$ has the binomial distribution—and can take on any of the values $0, 1/N, 2/N, \ldots, (N-1)/N, 1$. However, as $N \to \infty$, the strong law of large numbers says that $\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} X_i \to p$. Or, invoking the weak law, for large N, the frequency of blue marbles is close to p with high probability. The infinite population assumption in this context means that we assume the frequency of blue marbles is identical to p. That is, we assume the population is large enough that the difference between the actual (random) frequency of blue marbles and p is negligeable

For population biology models, the infinite population assumption will be applied to genotype frequencies rather than marbles. In a truly finite population model, the genotype frequency will be random, of randomness inherent in mating and srvival. In a large population, these frequencies will be close to their average values. The infinite population assumption is an approximation in which frequencies are replaced by probabilities. Specifically,

The infinite population model means that the frequency of a genotype G in the offspring of a parent population is set equal to the probability that a genotype G is sired.

In a truly infinite population, it is not possible to define frequencies as in (2.1), since the denominator would be infinity. Instead, we resort to the probability definition: f_G is the probability of drawing genotype G in a random sampling of the population.

Non-overlapping generations. How mating occurs and how successive generations interact are also important modelling considerations. The simplest model assumes **non-overlapping generations**. We imagine a population at time 0. This population mates and their offspring constitute the next generation. The individuals of the parent population do not mate further and no mating across generations takes place. This model is convenient to describe mathematically because we just need to keep track of genotype frequencies in successive generations. To keep in mind the no-overlap constraint, it is easiest conceptually to imagine that each generations is born, grows to maturity, mates, and then dies, leaving the next generation. but this is not necessary, if one keeps in mind that no mating occurs across generations. We will work mostly with the non-overlapping assumption, but will consider briefly how to modify it.

Selection and mutation. Different genotypes might confer different levels of fitness on individuals, affecting their ability to survive and reproduce. To model this we introduce fitness coefficients. If G is a genotype, however complicated, we let w_G be the conditional probability,

$$w_G \stackrel{\triangle}{=} \mathbb{P}\left(\text{an individual survives to reproduce} \mid \text{individual has genotype } G\right). \tag{2.6}$$

For example, in the one locus/two allele case, w_{AA} would denote the fitness coefficient of genotype AA.

Mutations can also occur. For an example there might be a mutation introducing a completely new allele to the population. Or, it could happen that one known allele mutates to another; for example, in a locus with two alleles A and a, allele A might occasionally mutate to a and vice-versa. Such mutations will be considered random events and modelled by probabilities,

Random mating. A mating model specifies what the probabilites are for offspring to have the various genotypes, given the genotype structure of the

parent populations. Notice that it is a probabilistic model. In most populations, there is an inescapable random component in who mates with whom, You may think that your own choice of a mate is, or will be, far from random! Though you would probably not rule out the influence of luck, your choices are influenced by personal history, location, religion, class, and more. However, it is not sensible to attempt a predictive theory of individual mate choice, especially in order to understand an evolving population. Instead, the biologist identifies and tries to explain observed statistical regularities—that is probabilities.

Despite randomness being inherent to any population model, the term **random mating** is used to denote a very specific model—somehow the simplest one— which is in some sense the "most random." In a rough sense, *random mating* refers to the situation in which all possible matings are equally likely. To be more precise, consider a random selected individual I, and consider $I's$ mating pool. In the monoecious case this will be the whole population, while in the dioecious case it will be all members of the opposite sex. The first supposition of random mating is that the probability I mates with an individual of genotype G is equal the frequency of G in I's mating pool. This is equivalent to assuming I chooses a mate by random sampling from its mating pool.

A further supposition of random mating is that each genotype mates in proportion to its frequency. Equivalently, each individual has an equally likely chance of mating and producing offspring. Finally, it is assumed that when an organism mates, each of its gametes has an equally likely chance of being passed to a child. Putting all the suppositions together, random mating is equivalent to the following model.

The genotype of a a random offspring of a random mating can be viewed as being produced by the following experiment. Pick a parent from the population by random sampling, choose one of its gametes at random, and copy its chromosomes. Return the parent to the population pool and repeat independently to get the second set of chromosomes The two copies obtained combine to make the chromosomes of the offspring.

Notice that the parents are not physically removed from the population, after one mating event. They are returned to the mating pool—we do not wish to deprive them of the pleasure of mating again.

Example 1 (continued) In example 1, what is the probability of producing AA by random mating? In this case the contribution from each parent is

obtained by the experiment of choosing a random allele from a randomly selected parent. By Theorem 1, we know that the probability of selecting allele A in one such experiment is $p_A = 5/12$. Since the allele choices from the first and second parents are independent, the probability that both contribute A is $(5/12)^2$.

2.3 One locus/two allele models

2.3.1 Basic allele probability update equations

A population genetics model is essentially a set of equations that specifies how genotype and allele frequencies evolve in successive generations. This section constructs models for the analysis of two alleles at one locus.

It is best to start with modelling assumptions that lead to the simplest model equations. These are:

Random mating.	$(A.1)$
Generations do not overlap.	$(A.2)$
The population is monecious and self fertilization is allowed.	$(A.3)$
No selection, mutation, or migration.	$(A.4)$

The meaning of these assumptions was explained in the previous section, except for the assumption of no migration. This just means that no individuals of whatever genotype from another population can migrate to the one under study.

The first goal is work out the mathematical consequences of assumptions (A.1)–(A.4) for genotype frequencies. That is the purpose of this first subsection. The second subsection develops and analyzes a complete model under the additional assumption of an infinite population. Further subsections state the Wright-Fisher model for finite populations, and models in which assumption (A.4) is modified to allow selection or mutation.

Because generations do not overlap, we can talk about generation t, where t takes on the values 0 (for the starting population), 1, 2, etc. Then, as in the previous section, $K(t)$ shall denote the frequency of genotype AA, $2L(t)$ the frequency of genotype Aa, and $M(t)$ the frequency of aa in generation t **at sexual maturity**. Similarly, $p_A(t)$ is the frequency of allele A, and $p_a(t) = 1 - p_A(t)$ the frequency of allele a in generation t at sexual maturity.

The proviso *at sexual maturity* means that the frequencies are those of the mating population. It is added here to distinguish between the probabilities with which the different genotypes are produced in mating and the frequency of those which survive to mate. This distinction is necessary later to model selection. But for the moment we are working under the assumption of no selection (A.4). We interpret this to mean that the probabilities with which generation $t-1$ produces a given genotype is the same as the probability that the genotype appears in generation t at sexual maturity. Recall from (2.4), that

$$p_A(t) = K(t) + L(t) \qquad \text{and} \qquad p_a(t) = L(t) + M(t). \tag{2.7}$$

Next, we derive the rules for how genotype and allele frequencies evolve from generation to generation. Consider, for example, the probability that a random mating in generation t produces an AA individual. Call this probability $k(t+1)$. According to the definition of random mating, this genotype is produced by two independent random draws, one for each parent. In each draw, a parent is selected at random, and then one of its two alleles is drawn at random. According to Theorem 1 of the previous section, the probability of choosing allele A in this way is simply the allele frequency $p_A(t)$. Therefore, using independence, the probability to get A from both parents is $k(t+1) = p_A^2(t)$. Similarly the probability to get aa, which we denote by $m(t+1)$ is, $m(t+1) = p_a^2(t)$. Finally, to be Aa, either an A somes from parent 1 and an a from parent 2, which has probability $p_A(t)p_a(t)$, or an a comes from parent 1 and an A from parent 2, which again has probability $p_A(t)p_a(t)$; thus, $2\ell(t)$, the probability to be born Aa, is $2p_A(t)p_a(t)$. To summarize:

$$k(t+1) = p_A^2(t); \qquad 2\ell(t) = 2p_A(t)p_a(t); \quad m(t+1) = p_a^2(t). \tag{2.8}$$

Notice that *the probabilites of genotype production in generation $t+1$ depend only on the allele frequencies in generation t.* This is a general feature of random mating.

2.3.2 The infinite population case and Hardy-Weinberg equilibrium.

We now add the assumption:

$$\text{The population is infinite.} \tag{A.5}$$

As explained in the previous section, the infinite population assumption means that we set the frequency of a genotype in generation $t+1$ equal to the probability that it is produced by generation t. Mathematically, this means, for example, that $K(t)$ is set equal to $k(t)$, where, as in the previous section, $k(t)$ is the probability that AA results from a random mating. Similarly $L(t+1) = \ell(t+1)$ and $M(t+1) = m(t+1)$ for $t \geq 0$. The full mathematical expression of the model for the evolution of genotype frequencies under assumptions (A.1)–(A.5) now follows from equation (2.8)

$$\begin{aligned} K(t+1) &= p_A^2(t) = (K(t) + L(t))^2 \\ L(t+1) &= p_A(t)p_a(t) = (K(t) + L(t))\,(L(t) + M(t)) \\ M(t+1) &= p_a^2(t) = (L(t) + M(t))^2 \left(= (1 - p_A(t))^2\right) \end{aligned} \tag{2.9}$$

Notice that this is a deterministic system of difference equations. Although the model is based on simple probabilistic calculations, the infinite population assumption means that the frequencies in successive generations are deterministic.

Given an initial set of genotype frequencies $K(0)$, $2L(0)$, and $M(0)$, these equations may be solved iteratively to obtain the frequencies for all future generations. In fact, the solution is particularly simple. To derive it, we use the fact expressed in (2.9) that genotype frequencies in generation $t+1$ depend only on allele frequencies in generation t. Thus we need only study the evolution of allele frequencies. But a simple calculation (left to the reader) gives

$$p_A(t+1) = K(t+1) + L(t+1) = p_A(t); \qquad p_a(t+1) = L(t+1) + M(t+1) = p_a(t) \tag{2.10}$$

(The second equality is of course a consequence of the first and $p_a(t) = 1 - p_A(t)$.) This has a very important consequence, which is a general feature of infinite population, random mating. *Allele frequencies do not change from generation to generation:*

$$p_A(t) = p_A(0) \text{ and } p_a(t) = p_a(0) \text{ for all generations } t \geq 0. \tag{2.11}$$

But then, substituting this result into (2.9), it follows that for all $t \geq 1$,

$$\begin{aligned} K(t) &= p_A^2(0) = (K(0) + L(0))^2 \\ L(t) &= p_A(0)p_a(0) = (K(0) + L(0))\,(L(0) + M(0)) \\ M(t) &= p_a^2(0) = (L(0) + M(0))^2 = (1 - p_a(0))^2. \end{aligned} \tag{2.12}$$

In other words, after the $t = 0$ generation mates, the genotype frequencies of all future generations are fixed. **Conclusion: In the absence of selection, random mating maintains genetic variation in the infinite population model.** As this observation was first most clearly enunciated by Hardy and Weinberg, it is standard to use the following terminology, motivated by (2.12).

Definition: Allele frequencies K, $2L$, and M, with $K + 2L + M = 1$, are said to be in **Hardy-Weinberg** equilibrium if there exists a p, $0 \leq p \leq 1$, such that

$$K = p^2, \qquad 2L = 2p(1-p), \qquad M = (1-p)^2.$$

Using this definition, we summarize the analysis so far in the next theorem, which is fundamental to population genetics.

Theorem 2 Hardy-Weinberg Theorem *Assume (A.1)–(A.5). After the $t = 0$ generation mates, the genotype frequencies for AA, Aa, and aa are in Hardy-Weinberg equilibrium with $p = p_A(0) = K(0) + L(0)$. In particular, the allele frequencies are constant from generation to generation.*

Despite its simplicity, this is an extremely important result, and the Hardy-Weinberg equilibrium is a fundamental concept in population genetics. In a natural population, single locus genotype alleles not in Hardy-Weinberg equilibrium indicate that one of the assumptions (A.1)–(A.5) does not hold. If the population is large and isolated and random mating seems reasonable, it is natural to expect selective pressure.

There is an easy test for Hardy-Weinberg equilibrium:

Exercise 3. Genotype frequencies K, $2L$ and M are in Hardy-Weinberg equilibrium if and only if $L^2 = KM$. (Remember, $K + 2L + M = 1$.)

2.3.3 The finite population Wright-Fisher Model.

In this section, it is assumed that assumptions (A.1)–(A.4) hold but that the population is a fixed, finite and of constant size N for all generations. The frequency of allele A in the population can take on any one of the values $0, 1/(2N), 2/(2N), \ldots, (2N-1)/(2N), 1$. Let $X(t)$ be the frequency of allele A in generation t; previously, this was called this $p_A(t)$, but the notation $X(t)$ will now be used to emphasize that for finite populations, the allele frequency will be a random variable. The total number of A alleles in generation t

is thus $NX(t)$. The Wright-Fisher model prescribes rules for determining how the probability mass function of $X(t)$ evolves. Imagine going through the population at time $t+1$ one individual at a time The random mating model says that the each allele in an individual is A with probability A and a with probability $p_a(t) = 1 - p_A(t)$. The Wright-Fisher model adds one more assumption: the random samplings that determines the alleles in each individual are independent from individual to individual. This means that the allele pool in generation t+1 may be viewed as being formed by $2N$ random draws, with replacement, from the allele pool of generation t. Suppose that we know $X(t) = i/(2N)$. As each draw results in an A with probability $X(t) = i/(2N)$, the number $2NX(t+1)$ of A alleles in generation t+1, will be a binomial random variable with parameters $n = 2N$ and $p = i/(2N)$. That is, the Wright-Fisher models states that,

$$\mathbb{P}\left(X(t+1)=\frac{k}{2N} \;\middle|\; X(t)=\frac{i}{2N}\right) = \binom{2N}{k}\left(\frac{i}{2N}\right)^k\left(1-\frac{i}{2N}\right)^{2N-k}, \quad 0 \le k \le 2N. \tag{2.13}$$

Observe a crucial and important difference between the infinite population model and the Wright-Fisher model. In the infinite population models, the allele frequencies are fixed; if the population starts out with a mix of A and a alleles at a locus, these alleles remain in the population in all generations in the same proportion. In the Wright-Fisher model, the allele frequencies are random, and will fluctuate around their expected values. This fluctuation is called **random genetic drift.** The fluctuation makes possible a dramatic difference from the infinite population case. Suppose the $t = 0$ generation starts out with a mix of both A and a alleles; in other words, the initial frequency of A alleles is $X(0) = i/(2N)$, where $i > 0$ and $i < 2N$. Then in the next generation there is a positive probability $(1 - (i/2N))^{2N}$ that there are **no** A alleles, and a positive probability $(i/2N)^{2N}$ that there are no a alleles. In the first case, one says that allele A becomes fixed, in the second, that allele a becomes fixed, because, after fixation once it becomes fixed, it remains the only allele in all future generations. If no allele is fixed in the $t = 1$ generation, there is still a probability it gets fixed in the $t = 2$ generation, and so on. Can a population obeying the Wright-Fisher model remain forever in a state in which both alleles are present? The answer is no. With probability one one allele or the other must eventually become fixed through random genetic drift. Which allele becomes fixed is a matter

of chance, but fixation will eventually happen. An argument for this will be given later in the treatment of Markov chains.

The evolution of the probability mass function of $X(t)$ can be derived from the conditional probability rule in (2.13) and a second assumption, called the Markov chain property, of $\{X(t)\}_{t\geq 0}$. (Markov chains will be discussed in detail later; the condition is not necessary to understand right now.) Howevery, it is not possible to derive a simple expression for the probability mass function at arbitrary t, and it is necessary in general to find approximations. For the moment, to illustrate how the Wright-Fisher models works, we work an example computing the probability mass function for the first three generations, starting from a given initial frequency of A alleles.

Example 2. Consider a population of 10 individuals evolving according to the Wright-Fisher model. Assume that in the $t = 0$ generation, the frequency of A alleles is 0.5. According to equation (2.13), the probability mass function of $X(1)$ is

$$p(k,1) \triangleq \mathbb{P}\left(X(1)=\frac{k}{20} \;\middle|\; X(0)=0.5\right) = \binom{20}{k}(.5)^k(.5)^{20-k} = \binom{20}{k}(.5)^{20},$$

for integers k, $0 \leq k \leq 20$. What about $p_2(k)$, the probability mass function of $X(2)$? Using the rule of total probablities,

$$\begin{aligned} p(k,2) &= \sum_{j=1}^{20} \mathbb{P}\left(X(2)=\frac{k}{20} \;\middle|\; X(1)=\frac{j}{20}\right)\mathbb{P}\left(X(1)=j\right) \\ &= \sum_{j=1}^{20} \binom{20}{k}\left(\frac{j}{20}\right)^k\left(\frac{20-j}{20}\right)^{20-k}\binom{20}{j}(.5)^{20}. \end{aligned}$$

This is already unpleasant. $p(k,3)$, the probability mass function of $X(3)$, can be computed in terms of $p(k,2)$ by applying the rule of total probabability again:

$$p(k,3) = \sum_{j=1}^{20} \mathbb{P}\left(X(2)=\frac{k}{20} \;\middle|\; X(1)=\frac{j}{20}\right) p(j,2).$$

We could plug in the expressions derived above $p(j,2)$, but we will just get a very complicated formula. However, the principle extends; for each t, $p(k,t{+}1)$ can be computed in terms of $p(k,t)$ using the total probability rule.

As an application, suppose we want to know the probability that allele A is fixed by the $t = 2$ generation. This is

$$p(20, 2) = \sum_{j=1}^{20} \left(\frac{j}{20}\right)^{20} \binom{20}{j} (.5)^{20}.$$

2.3.4 Infinite population model with selection.

We shall now modify the infinite population model to account for selection. The effects of selection will be incorporated in the model using the selection coefficients w_{AA}, w_{Aa} and w_{aa}. As explained in section 2.2, these are interpreted as conditional probabilities of survival to reproductive maturity of the different genotypes. For example w_{AA} is the probability that an individual survives to reproductive maturity, given that it is AA. The precise assumptions of this subsection are therefore (A.1), (A.2), (A.3), (A.5) and

No mutation or migration. Selection coefficients w_{AA}, w_{Aa}, w_{aa}. $(A.4)'$

To state the model, it is necessary to modify the analysis that led to equation (2.8). Consider generation $t+1$. Let U denote the event that an individual born to generation t parents survives to reproductive maturity. Let E_{AA} be the event that an individual born to t generation parents is AA; define E_{Aa} and E_{aa} similarly. The probability of E_{AA} is just the probability $k(t+1)$ computed in equation (2.8), and likewise $\mathbb{P}(E_{Aa}) = 2\ell(t+1)$ and $\mathbb{P}(E_{aa}) = m(t+1)$. By the total probability rule,

$$\begin{aligned} \mathbb{P}(U) &= \mathbb{P}(U \mid E_{AA})\mathbb{P}(E_{AA} + \mathbb{P}(U \mid E_{Aa})\mathbb{P}(E_{Aa}) \\ &\qquad +\mathbb{P}(U \mid E_{aa})\mathbb{P}(E_{aa}) \\ &= w_{AA}p_A^2(t) + w_{Aa}2p_A(t)(1 - p_A(t)) + w_{aa}(1 - p_A(t))^2. \end{aligned} \tag{2.14}$$

For later purposes, it is convenient to define the **fitness function**

$$W(p) = p^2 w_{AA} + 2p(1 - p)w_{Aa} + (1 - p)^2 w_{aa}, \qquad 0 \le p \le 1.$$

With this notation, equation (2.14) is written,

$$\mathbb{P}(\text{survival to reproductive maturity in generation } t+1) = W\left(p_A(t)\right).$$

The probability that an individual is AA at time of reproductive maturity in generation $t+1$ is the conditional probability that an individual born

to generation t has genotype AA *given that it has survived to reproductive maturity.* This is

$$\begin{aligned}\mathbb{P}\left(E_{AA} \mid U\right) &= \frac{\mathbb{P}(E_{AA} \cap U)}{\mathbb{P}(U)} = \frac{\mathbb{P}(U \mid E_{AA})\mathbb{P}(E_{AA})}{\mathbb{P}(U)} \\ &= \frac{w_{AA}p_A^2(t)}{W(p_A(t))}. \end{aligned} \tag{2.15}$$

In going from the second expression to the third in this derivation, we used the equality '$\mathbb{P}(E_A A \cap U) = \mathbb{P}(U \mid E_{AA})\mathbb{P}(E_{AA})$, which is a direct consequence of the definiton of conditional probability; this step is really just a derivation of Bayes rule for the original conditional probability on the left.

Similar reasoning gives

$$\mathbb{P}\left(E_{Aa} \mid U\right) = \frac{2w_{Aa}p_A(t)(1-p_A(t))}{W(p_A(t))}, \tag{2.16}$$

$$\mathbb{P}\left(E_{aa} \mid U\right) = \frac{2w_{aa}(1-p_A(t))^2}{W(p_A(t))}, \tag{2.17}$$

The infinite population assumption means that population frequencies are set equal to probabilities. So, for example, $K(t{+}1)$ is identified with the first conditional probability, the probability of being AA given the probability of survival to reproductive maturity. The update equations for genotype frequencies follow as an immediate consequence:

$$K(t{+}1) = \frac{p_A(t)^2 w_{AA}}{W(p_A(t))} \tag{2.18}$$

$$2L(t{+}1) = \frac{2p_A(t)(1-p_A(t))w_{Aa}}{W(p_A(t))} \tag{2.19}$$

$$M(t{+}1) = \frac{(1-p_A(t))^2 w_{aa}}{W(p_A(t))}. \tag{2.20}$$

As in the model with no selection, the genotype frequencies in generation $t{+}1$ depend only on the allele frequencies in generation t. Therefore, it is possible to summarize (2.18)–(2.20) by one equation for the allele frequency p_A. Using $p_A(t{+}1) = K(t{+}1) + L(t{+}1)$, equations (2.18)–(2.20) imply

$$p_A(t{+}1) = \frac{p_A(t)^2 w_{AA} + p_A(t)(1-p_A(t))w_{Aa}}{W(p_A(t))} \tag{2.21}$$

Analysis of this equation will reveal how the allele frequency evolves under the influence of selection on the genotypes. From the scientific point of view, it is interesting to calculate $\lim_{t\to\infty} p_A(t)$ and to show how it changes as the relative values of w_AA, w_{Aa} and w_{aa} change. The mathematics is quite a bit more complicated than in the case of no selection, and the theory is defered to section 2.4.

Exercise 4. Using the selection coefficients w_{AA}, w_{Aa}, w_{aa}, modify the Wright-Fisher model to include selection. Your final answer should be a formula that specifies how to compute the conditional probabilities on the left side of equation (2.13).

2.3.5 Infinite population with mutation, but no selection

In this section, we derive a model assuming (A.1)–(A.3) and (A.5) and no selection or migration. However, we will allow for mutations between A and a forms of the alleles. Specifically, we replace assumption (A.4) with

$(A.4)''$: *No selection or mutation. In reproductive transmission, an A allele mutates into an a allele with probability u and an a allele mutates into an A allele with probability v. The probabilities u and v satisfy $0 < u+v < 2$.*

The case in which $u = 0$ and $v = 0$ is excluded in $(A.4)''$ to impose some mutation. The case in which $u + v = 2$, or equivalently, $u = v = 1$, is excluded because, since it implies that A always mutates to a and a to A, it is unrealistic.

Assumption $(A.4)''$ means that the probability that a parent contributes an A in a random mating is

$$(1 - u)p_A(t) + vp_a(t).$$

(Derive this from the total probability rule.) We learned in analysis of the infinite population with random mating that the frequency $p_A(t+1)$ of allele A in generation $t+1$ is just the probability the probability of drawing A in a random sample of the allele pool of generation t. With no mutation or selection, this fact leads directly to the constancy of allele frequency derived in equation (2.11). But with mutation, it yields the model:

$$p_A(t+1) = (1 - u)p_A(t) + v(1 - p_A(t)) = (1 - (u + v))p_A(t) + v. \quad (2.22)$$

(The genotype frequencies will be given by $K(t+1) = p_A(t+1)^2$, $2L(t+1) = 2p_A(t+1)(1 - p_A(t+1))$, and $M(t+1) = (1 - p_A(t+1))^2$.)

Equation (2.22) is an example of a first order, linear, inhomogeneous difference equation and can be solved explicitly. The first step is to define $\tilde{p}_A(t) = p_A(t) + c$ where c is a constant. By plugging this into (2.22),

$$\tilde{p}_A(t+1) = (1-(u+v))\tilde{p}_A(t)+v+c-c(1-(u+v)) = (1-(u+v))p_A(t)+v+c(u+v).$$

Now choose c to make the constant term $v + c(u + v) = 0$; in other words $c = -v/(u + v)$. Then

$$\tilde{p}_A(t) = (1 - (u + v))\tilde{p}_A(t).$$

Exercise 5. Show by induction that $\tilde{p}_A(t)$ is

$$\tilde{p}_A(t) = (1 - (u + v))^t\tilde{p}_A(0).$$

Exercise 6. What is the limiting behavior of $p_A(t)$ as $t \to \infty$? The condition $0 < u+v < 2$ in assumption $(A.4)''$ implies that $-1 < 1-(u+v) < 1$. Show that

$$\lim_{t\to\infty} p_A(t) = \frac{v}{u + v}.$$

Conclusion. With mutation $p_A(t)$ tends to the limiting value $v/(u + v)$, which is independent of the initial frequency $p_A(t)$. If $v > 0$ and $u > 0$, variation between A and a alleles is maintained. If $v = 0$, while $u > 0$, allele A disappears in the long run limit, while if $u = 0$ and $v > 0$, allele a disappears.

Exercise 7. Analyze the solution of (2.22) when $u = 1$ and $v = 1$ and interpret.

Exercise 8. Let $p_A(0) = 0.5$, Let $p_A^{(1)}(t)$ denote the allele frequency in generation t in the the case in which $u = 1/4$ and $v = 1/2$. Let $p_A^{(2)}(t)$ be the allele frequency in generation t in the case that $u = 1/16$ and $v = 1/8$.

a) Show that the limiting frequency, as $t \to \infty$, is the same in both cases.

b) Denote the limiting frequency found in a) by $\bar{p}$. Find the smallest value T_1 such that $|p_A^{(1)}(T_1) - \bar{p}| \leq 0.01$. Find the smallest value T_2 such that $|p_A^{(2)}(T_2) - \bar{p}| \leq 0.01$. (Note that $p_A^{(1)}(t)$ and $p_A^{(1)}(t)$ are both increasing in t.) Compare T_1 and T_2 and explain why your result is to be expected on intuitive grounds, considering the mutation rates in both cases.

2.4 Extensions

This section is a sequence of guided exercises in which, following the analysis of section 2.3, you are to set up and analyze extensions to the one locus/two allele models. Throughout, only infinite population models are considered. You will see that the one locus/two allele assumption and the assumption that the organism is monecious are not really important in the principles that give the Hardy-Weinberg equilibrium.

2.4.1 One locus/multiple alleles

In this section you will set up a model for one locus with many alleles, assuming (A.1)-(A.5). The general model will have alleles $A_1, A_2, \ldots, A_m$ all at one locus. The allele frequencies in generation t are denoted by $p_1(t), p_2(t), \ldots, p_m(t)$, where p_i is shorthand for p_{A_i}. Of course $p_1(t) + \cdots + p_m(t) = 1$. The genotype frequencies in generation t are denoted by $f_{A_iA_j}(t)$, where $1 \leq i \leq j \leq m$, in accordance with the notation of section 2.2. For convenience, $f_{ij}(t)$ can be used as shorthand for $f_{A_iA_j}(t)$.

You may do the exercises that follow either for general m, or, for notational simplicity, $m = 3$.

Exercise 9. Express $p_i(t)$ in terms of $f_{ij}(t)$.

Exercise 10. Apply random mating to express $f_{ij}(t+1)$ in terms of the allele frequencies in the previous generation.

Exercise 11. Generalize the Hardy-Weinberg theorem to the multi-allele case. Show that the allele frequencies are constant and that the genotype frequencies reach equilibrium values in generation $t = 1$ that remain fixed for all future time. Express those equilibrium genotype frequencies in terms of the allele frequencies in the time $t = 0$ population.

Exercise 12. (See Exercise 3.) Show that the set of genotype frequencies $f_{ij}(t)$, $1 \leq i \leq j \leq m$, is in Hardy-Weinberg equilibrium if and only if for every $i \neq j$, $f_{ij}^2 = 4f_{ii}f_{jj}$.

2.4.2 One locus/two alleles for dioecious populations

In this section, we show that the Hardy-Weinberg equilibrium also extends to dioecious organisms, at least for a locus on an autosomal chromosome.

Our model will assume conditions (A.1), (A.2), (A.3), (A.4), that is, random mating in an infinite population with no selection, mutation, or migration, and non-overlapping generations. We add one more assumption:

Sex and locus genotype are transmitted independently (*A*.6)

To adapt the model to dioecious organisms and derive the equations for the evolution of allele and genotype frequencies, it is necessary to consider the male and female portions separately, at least initially. Let $K^m(0)$, $2L^m(0)$, $M^m(0)$ be the frequencies of genotypes AA, Aa, and aa, respectively, among the males in the $t = 0$ generation. Let $K^f(0)$, $2L^f(0)$, $M^f(t)$ be the corresponding frequencies among the females. Denote the allele frequencies by $p_A^m(0)$ and $p_A^f(0)$. That next exercise shows that the Hardy-Weinberg principle remains true. The only difference is that it takes two generations to reach Hardy-Weinberg equilibrium. Because of assumption (A.6), the first mating equalizes genotype and allele frequencies between the male and female subpopulations. Once this occurs, future matings produce Hardy-Weinberg equilibrium.

Exercise 13.

a) Calculate the frequencies in the *entire* population of the genotypes AA, Aa, and aa in the first generation. (Remember that the frequency of a genotype is interpreted as the probability of getting the genotype in a random draw. Assume that males and females are drawn with equal probability.)

b) Random mating under assumption (A.6) is equivalent to producing a new offspring by a random draw of a genotype from the male population and an independent random draw of a genotype from the female population and then, independent of the genotype, choosing a sex at random, with male and female choice being equally likely. Since the choice of sex is independent of choice of genotype, the frequency of each genotype in the $t = 1$ generation will be the same in males and females, and equal to the overall genotype frequency. The same will be true in all subsequent generations. Find the common genotype frequencies in the $t = 1$ generation in terms of the allele frequencies in the $t = 0$ generation. Show that the frequency of allele A in the $t = 1$ generation, which is the same if restricted to either males or females, is $(1/2)(K^m(0) + K^f(0) + L^m(0) + L^f(0))$.

c) Show that the genotype frequencies in all following generations are in Hardy-Weinberg equilibrium, and determine the equilibrium values in terms of the frequencies of the $t = 0$ generation.

Consider next, a model for a locus on the X chromosome of a human population. Males will have genotypes consisting only of a single letter, either A or a; the frequency of genotype A among the males of generation t is simply the frequency of allele A among males and we are denoting it by $p_A^m(t)$; there are no frequencies $K^m(t), L^m(t), M^m(t)$ to contend with. Again we assume random mating. The probability that random mating results in a female is the probability that the male parent passes on its X chromosome, which we assume occurs with probability one-half. The males in generation $t+1$ get their A chromosomes only from females of the previous generation.

Exercise 14. a) Show that random mating implies that $K(1)^f = p_A^m(0)p_A^f(0)$, $2L^f(1) = p_A^m(0)(1-p_A^f(0))+p_A^f(0)(1-p_A^m(0))$, $M(1)^f = (1-p_A^m(0))(1-p_A^f(0))$, and $p_A^f(1) = (p_A^m(0) + p_A^f(0))/2$. Show also that $p_A^m(1) = p_A^f(0)$.

b) The same arguement as in a) shows that for any t,

$$\begin{aligned} p_A^m(t+1) &= p_A^f(t) \\ p_A^f(t+1) &= \frac{1}{2}(p_A^m(t) + p_A^f(t)) = \frac{1}{2}(p_A^f(t-1) + p_A^f(t)) \end{aligned}$$

To solve $p_A^f(t+1) = \frac{1}{2}(p_A^f(t-1) + p_A^f(t))$, look for a solution of the form $p_A^f(t) = \alpha + \beta r^t$. By plugging this into the equation show that there is a value r, not equal to 1, such that $\alpha + \beta r^t$ will be a solution. (Hint; find a quadratic equation for r and find a root not equal to 1.) Determine the values of α and β in terms of $p_A^f(0)$ and $p_A^m(0)$ by setting $\alpha + \beta = p_A^f(0)$ and $\alpha + \beta r = p_A^f(1) = (p_A^m(0) + p_A^f(0))/2$. As a result, find a formula for $p_A^f(t)$ for general t.

c) Observe that the overall frequency of A alleles in generation $t = 0$ is $p \stackrel{\triangle}{=} (2/3)p_A^f(0) + (1/3)p_A^m(0)$. Using the results of b), show that $\lim_{t\to\infty} p_A^f(t) = p$ and $\lim_{t\to\infty} p_A^m(t) = p$, and conclude that the limiting genotypes in the female population are in Hardy-Weinberg equilibrium with the original overall frequency of A's.

d) Use the results of part b) to show $(2/3)p_A^f(t) + (1/3)p_A^m(t)$ is constant from generation to generation. This extends the principle that random mating implies allele frequencies are constant.

2.4.3 One locus/two alleles with overlapping generations

Consider an infinite population models in which individuals are continually dying and being born. It will be assumed that the death rate and birth rate are the same, and time will be measured in units so that on the average $\triangle t$ are born and $\triangle t$ individuals die in a time interval of length $\triangle t$. There is no preference or selection for genotype in death. This means that genotypes die in proportion to their frequency in the population. Mating is random, so genotypes are created by random selection of alleles from the current population. Let us translate these conditions into mathematics. Let p denote the frequency of allele A at time $t = 0$. Because of random mating, this frequency will remain constant in time. Let $K(t)$ denote the frequency of genotype AA at time t. Consider $K(t + \triangle) - K(t)$. In time $\triangle t$, $K(t)\triangle t$ individuals of genotype AA die and $p^2\triangle$ AA's are born. Thus

$$K(t + \triangle) - K(t) = \triangle t(-K(t) + p^2).$$

Divide by $\triangle t$ and let it tend to 0. Then

$$K'(t) = \lim_{\triangle t \downarrow 0} \frac{K(t + \triangle) - K(t)}{\triangle t} = -K(t) + p^2. \tag{2.23}$$

Exercise 15. Let $K(0) = K$. Solve equation (2.23) in terms of K and t. (Hint: consider $\tilde{K}(t) = K(t) - p^2$.) Show that $\lim_{t\to\infty} K(t) = p^2$. Thus $K(t)$ will tend to the Hardy-Weinberg equilibrium value associated to allele frequency $p = p_A(0)$, as time increases.

It is easy to see that the results of Exercise 15 generalize to the other genotype frequencies. **Conclusion:** In this infinite population, overlapping generation model, the Hardy-Weinberg equilibrium is achieved in the infinite time limit.

If the turn-over rate is small, that is , if individuals are born and others die frequently, it will not take long for the actual frequencies to become close to the Hardy-Weinberg equilibrium values.