Lecture Notes in Population Genetics

Stanley Sawyer — ©

Chapter 1

Table of Contents:

Vs. February 1, 2005

1.1. Introduction

One of the strongest influences on the growth and behavior of most living creatures is the genetic material or DNA located in their chromosomes. Bacteria tend to have a single circular chromosome. Most higher plants and animal have a number of pairs of long linear molecules. In either case, the chromosome or chromosomes can be thought of as a string of letters from the alphabet $\mathrm{T}, \mathrm{C}, \mathrm{A}, \mathrm{G}$, where each letter corresponds to a specific nucleotide. From this point of view, a mouse is the same as a tomato to a geneticist, since both have about the same amount of DNA.

By definition, a gene or genetic locus is a segment of a chromosome that is associated with a particular trait. This typically consists of one or more coding regions for a protein or RNA enzyme along with recognition sites for regulatory molecules. Proteins are composed of strings of amino acids. There are around 20 different amino acids, as opposed to 4 different nucleotides. Coding regions for genes are built up from consecutive triplets of nucleotides that are called codons. Codons are mapping to amino acids by RNA translation enzymes. Since the number of possible codons is $4^3 = 64$, there are enough codons to describe all amino acids with plenty of room to spare. Most amino acids are described by more than one codon. Codons that are not assigned to an amino acid tell the RNA translation enzyme to stop, so that the codon language has verbs as well as nouns.

Most genes are templates for enzymes or proteins that control or take part in biochemical processes. In most plants and animals, the DNA is arranged in a number of pairs of chromosomes. Such creatures are called *diploid*. Creatures that have non-paired chromosomes (such as bacteria) are called *haploid*. Some domesticated plants such as corn are *tetraploid*, which means that their chromosomes occur in groups of four. There are 23 chromosome pairs in man (and so 46 chromosomes), 4 pairs in *Drosophila* (fruit flies), and other numbers in other creatures.

In humans, 22 of the 23 chromosome pairs are composed of two chromosomes that are more-or-less the same size and have the same genetic loci. These are called *autosomal* loci. The remaining chromosome pair has two different types of chromosomes, one (type X) about six times the size of the other (type Y). The 23$^{\mathrm{rd}}$ chromosome pair in humans has two X chromosomes in females and one X and one Y in males. These are called *sex chromosomes* (or the *sex-chromosome pair*).

An Example: There is some evidence that blue vs. brown eye color in humans is governed by a single genetic locus with two possible types of genes, which we'll call U and W. (There is also some evidence that the genetics of human eye color is more complicated, but let's ignore this for the moment.) Since this locus occurs on both chromosomes of one of the chromosome pairs (that is, it is an *autosomal* locus), there are four possibilities (or *genotypes*) for individuals: UU, UW, WU, and WW. In almost all cases the action of genes is independent of the chromosome in which they occur, so that the number of possibilities reduces to three: UU, UW, and WW. In this case, both UW and WW carrying individuals have brown eyes. Brown eye color might be due to a particular protein pigment, and one copy of the W gene might produce enough pigment to cause the trait. Individuals without this pigment might have blue eyes.

In general, if there are two types of genes (say A and B) for which AB has the same effect as AA, then A is said to be *dominant* with respect to B, and B is *recessive* with respect to A. Typically this is because A produces a protein in sufficient quantities from one gene. The recessive gene is often a damaged version of the dominant gene. It is useful to use the word *allele* for a genetic type (as opposed to gene, which will refer to a piece of a particular molecule). In many allele classifications such as U/W, each 'allele' will be a collection of slightly different alleles with the same gross properties.

ABO Blood groups: As a second example, the ABO blood group in humans is governed by a single genetic locus with three alleles, which are called A, B, and O. Since this locus also occurs on both chromosomes of one of the chromosome pairs, there are six genotypes: AA, AB, BB, AO, BO, and OO. Both alleles A and B are dominant with respect to O, but neither are dominant with respect to the other. This leads to four 'blood types': A (AA or AO), B (BB or BO), O (OO) and AB (AB). The alleles A and B produce proteins which can cause dangerous immune reactions if blood containing that protein is given in a transfusion to an individual who does not have that protein. Thus an AB could receive blood from anyone without worrying about this immune reaction, while an AO would be endangered by a transfusion from a BB, BO, or AB.

By definition, a genotype is a *heterozygote* if it has two different alleles (such as AB, AO, or BO) and *homozygote* for genotypes with two copies of the same allele (such as AA, BB, and OO). Thus the six possible genotypes at the ABO locus are composed of three heterozygotes and three homozygotes.

Sickle-cell anemia: This is a disease controlled by a genetic locus with two alleles that are called S (for "sickling") and N (for "normal"), respectively. Since this locus occurs is also autosomal, individuals can be of three genotypes, NN, SN, and SS. Individuals with genotype SS have sickle-cell anemia and are usually very ill, often dying before the end of their twenties. Individuals of genotype SN have 'sickle-cell trait' and appear to suffer no ill effects, although there has been some concern about relatively anoxic environments such as airplane cockpits. However, SN people have a markedly higher resistance to malaria. Thus, in a malarial area, villages would have a tendency to have a high proportion of genotype SN, since the other genotypes would tend to die off. As we will see below, the children of two SN-individuals will be on the average 25% NN, 25% SS, and only 50% SN. Thus villages could not remain totally of genotype SN indefinitely. This is an example of what is called *heterozygote advantage* or *overdominance*, in which a variety of genotypes is kept at a particular locus due to the most successful genotype not breeding to form.

More about sex chromosomes: As mentioned before, the human X chromosome is about 6 times the length of the Y and is among the largest human chromosomes. The Y chromosome is one of the three smallest chromosomes. Human females have two X chromosomes (i.e., are XX) while males are XY. (Male frogs, however, are the analog of XX while females are XY, so that the situation is not uniform across species.) Human males (and presumably also female frogs) are then especially vulnerable to defective copies of X-chromosome genes for which there is no 'backup copy' on the other chromosome. The genes associated with some of the most common types of hemophilia and for the pigments necessary for perceiving red and green colors are located on the X-chromosome in humans. (The gene for the blue pigment is autosomal.) As expected, hemophilia and red/green color blindness are much more common in males than in females.

One sex or two? In most higher animals and some plants, the population is split into two sexes and mating occurs between members of opposite sexes. Such creatures are called *dioecious*. Most higher plants are diploid (that is, have chromosome pairs as opposed to a single chromosome) but are *monoecious*, which means that any individual can act as either sex and can even fertilize itself.

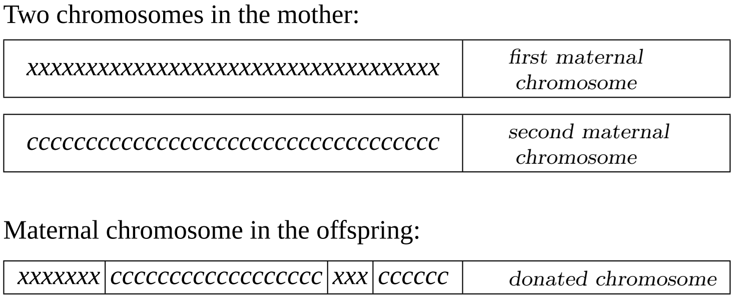
For all diploids, barring genetic accidents, the offspring have the same number of chromosome pairs as their parents. In an offspring, each chromosome pair is composed of one copy of one of the two maternal chromosomes and one copy of one of the two paternal chromosomes. (In monoecious plants, the “mother” and “father” might be the same individual, but “maternal” and “paternal” chromosomes come from different sources in the plant: Maternal chromsomes come from seeds and paternal chromosomes from pollen.)

Statistically, the two choices of which chromosome are usually independent with equal probability for the two parents and for different chromosome pairs. (This is the basic principle of “Mendelism”.) At an autosomal locus, this means that the offspring inherits one maternally-derived gene and one paternally-derived gene, with each chosen independently and at random as a copy of one gene in each parental genotype. Thus, for example, if both parents are of allelic type AB, an offspring will be AA, AB, or BB with probabilities respectively $1/4$, $1/2$, and $1/4$.

As mentioned before, sex chromosomes (in humans) are of types either XX or XY. Since all mating is between a male and a female, with both distributing one chromosome from each pair to each offspring, XX and XY are the only possibilities for offspring (barring genetic accidents).

Note that genes on the Y-chromosome pass directly from fathers to sons, avoiding all female intermediaries. In particular the Y-chromosome reproduces itself in a haploid rather than diploid manner, by cloning itself from generation to generation. There is also important genetic material in *mitochondria*, which are organelles that are carried inside cells but outside the cell nucleus. In humans and most animals, these are cloned from the mother. Such extranuclear loci would then be carried from mothers to daughters (and also sons) in a haploid manner. Higher plants also have *chloroplasts*, which are another type of extranuclear organelle that are passed maternally.

Figure 1.1. Illustration of linkage.


Two chromosomes in the mother:
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
first maternal chromosome
ccccccccccccccccccccccccccccccccccc
second maternal chromosome
Maternal chromosome in the offspring:
xxxxxxx
cccccccccccccccccc
xxx
cccccc
donated chromosome

More genetic scrambling: For most chromosome pairs, the actual situation is slightly more complicated. The chromosome donated from the mother (for example) can be a composite copy of the two maternal chromosomes, with the donated chromosome changing from one maternal chromosome to the other at one or more *crossover points*. This is illustrated by Figure 1.1, in which the parts of the donated chromosome that come from the first maternal chromosome are marked ‘*xx*’. Figure 1.1 illustrates three *crossover breaks* and reattachments (or *crossover events*) between the two maternal chromosomes in one chromosome pair as the donated chromosome was being formed. This process is called *recombination*. In humans, there is typically an average of about one crossover event per donated chromosome.

If we are following a single genetic locus, recombination doesn't matter unless there is a crossover point within the genetic locus. This is a rare event, since genes are generally much shorter than chromosomes, but it does happen. Crossover between X and Y chromosomes in humans can occur is rare in regions containing genetic loci. Otherwise, there would be genes that would occur on both X and Y chromosomes, which is rare except for genetic accidents.

1.2. Population Genetics

Population genetics is the study of the frequencies of alleles in populations and how they change over time or space. Three important effects that exert an influence on allele frequencies at a genetic locus are

(i) Selection (2.1)
(ii) Mutation
(iii) Genetic drift

Selection (sometimes called *Darwinian selection*) refers to changes in allele frequencies due to the effects of the gene on its host. Examples would be effects lowering or increasing the death rate of individuals carrying the gene, or lowering or increasing the number of its surviving offspring.

A gene undergoes *mutation* if it physically changes to another allele, as a result of an accident in replication during conception or some other cause.

Genetic drift is the result of probabilistic effects due to Mendelism or to the chance effects of mating and survival in a small population. A carrier of a particular allele may leave no surviving offspring for reasons which have nothing to do with that allele, for example accidental death. In general, the number of the surviving offspring of an individual can be thought of as a random variable, with a mean given by selection, but with still a positive probability of being zero. An allele that has a selective advantage over others may still be lost from the population due to random effects.

The *Wright-Fisher* model is an attempt to model these and similar effects. The Wright-Fisher model for dioecious populations assumes that the population is rigidly held at N_1 males land N_2 females over many generations. At the beginning of each generation, the population undergoes *random mating* (as defined below) to produce a large number offspring. Of these, N_1 males and N_2 females are chosen at random to to adulthood and replace the parents.

By "random mating" we mean the following. For each offspring, all possible male-female pairs are equally likely to be the parents, with the choices being independent for different offspring. This can be described by saying that the children choose their parents independently and at random.

The standard Wright-Fisher model assumes that the population is monoecious and that any parent can act as either mother or father (or both). The population is held at N individuals and, for each offspring, the parents are chosen independently and at random from the N individuals in the previous generation. In particular the probability that an offspring is the result of a self-fertilization is $1/N$.

At an autosomal locus, there is one maternally-derived gene and one paternally-derived gene in any offspring. Random mating and Mendelism together in a dioecious population imply that the maternally-derived gene in an offspring is a copy of a randomly chosen gene from among the $2N_2$ genes in the female adults in the preceding generation, and similarly the paternally-derived gene is a copy of a randomly chosen gene from the $2N_1$ genes of the male adults. In the monoecious or standard version of the Wright-Fisher model, each gene in an offspring is a copy of a randomly chosen gene from among the $2N$ genes of the preceding generation. In particular the two genes in an offspring (or any other two genes in the new generation) come from the same parent with probability

$1/N$, and are copies of the same parental gene with probability $1/2N$.

Selection can act either through genotype-dependent variation in the survival rate of juveniles until sexual maturity (this is called *viability* selection), or else through differences in the number of offspring produced as an adult, or both. The second type of selection is called *fertility* selection, and would depend on potential mating pairs rather than individuals. We will restrict ourselves to viability selection for simplicity, and assume that all surviving juveniles are equally fertile.

1.3. Random mating and the Wright-Fisher model

The ideas in the previous section can be summarized as follows: At birth, genes in individuals in the new generation can be found by independent sampling from the genes in the adults in the preceding generation. If the population is dioecious, the genes in each juvenile are found by independently sampling one gene from the female adults and one gene from the male adults. All sampling is with replacement.

The allelic types of genes may change (or *mutate*) during the process of sampling. There may also be *selection*, which is implemented by biased independent sampling based on the genotypes of the parents.

The Wright-Fisher model is usually assumed to be *monoecious*, with one set of N adults that can play the role of either sex. Sampling is still independent, so that (without selection) the probability that the two parents are the same individual is $1/N$.

Random mating is repeated binomial sampling: Note that this model is mathematically equivalent to the following: Consider the $2N$ genes in the N adults in the parental generation. Then the $2N$ genes in the N adults in the next generation are found by repeated binomial sampling (with replacement) from the $2N$ genes in the parental generation.

Suppose that we are following a single autosomal genetic locus in a monoecious population of size N with two types of genes, which we will call allele A and allele a.

Let $Q(n)$ be the number of genes of type A at the beginning of the n^{th} generation. If there is no mutation or selection, the, given $Q(n)$, the number of A-genes at birth in the next generation is probabilistically equivalent to the number of successes in $2N$ trials (corresponding to N juveniles each choosing their parents) with probability of success $p = Q(n)/2N$ at each trial. Mutation is modeled by assuming that each gene, as it is sampled from the preceding generation, changes to a different type with some probability between zero and one. For example, if there is a mutation rate of u from a to A and of v from A to a, then each sampled gene is A with probability $f(p) = (1-v)p + u(1-p)$. Selection can be modeled in a similar way (see below).

In general, we assume that the probability that an A gene is sampled for the next generation, assuming that the current proportion of A genes is p, is given by $f(p)$. If there is no selection or mutation (that is, pure random mating), then $f(p) = p$.

Symbolically, we can write the distribution of $Q(n+1)$ in the next generation as

$$\{\, Q(n+1) \ \mid \ Q(n) = k \,\} \ \approx \ B\big(2N,\ f(k/2N)\big) \tag{3.1}$$

where "|" means "given", $B(2N, p)$ denotes a binomial distribution based on $2N$ trials and probability of success p for reach trial, and '$\approx$' means "has the same distribution as". This is equivalent to saying

$$\Pr\big(Q(n+1) = j \mid Q(n) = k\big) = \binom{2N}{j} f(k/2N)^j \big(1 - f(k/2N)\big)^{2N-j}$$

for $j = 0, 1, 2, \dots, 2N$. In particular

$$E\big(Q(n+1) \mid Q(n) = k\big) = 2Nf(k/2N) \tag{3.2}$$

The Wright-Fisher model can be refined in many different ways. If more than one genetic locus is being followed, the chromosomes of the juveniles are determined by random sampling of recombined chromosomes from the preceding generation. If the adults in the preceding generation are more likely to mate with individuals of the same genotype as themselves (this is called *assortative mating*), this can be modeled as a correlated choice for the two genes within each new offspring.

The Probability of Fixation: In the simplest case, the model has genetic drift only; i.e.

$$\{\,Q(n+1)\ \mid\ Q(n)\,\}\ \approx\ B\biggl(2N,\,\frac{Q(n)}{2N}\biggr) \tag{3.3}$$

The process $\{Q(n)\}$ is a Markov chain with finite state space $0, 1, \dots, 2N$. The states 0 and $2N$ are traps, since, without mutation, if either allele is lost from the population then it is lost forever. The Markov chain is irreducible except for these two traps. By basic results from Markov chain theory, this means that eventually $\{Q(n)\}$ ends up at either 0 or $2N$, which means that eventually either the allele A or the allele a is lost. It is traditional to say that the population is then *fixed* at the allele which is now uniform at that locus.

Let's see if we can calculate the probability that the Wright-Fisher model is eventually trapped (fixed) at A. By (3.2), given $Q(n) = k$, the mean of the binomial variable $Q(n+1)$ is $2Np = 2N(k/2N) = k$, as should be expected from the absence of selection and mutation. Then $E\bigl(Q(n)\bigr) = E\bigl(Q(n-1)\bigr) = \ldots = Q(0)$ by induction for all n. Note that

$$\lim_{n\to\infty} Q(n)/2N = 1$$

if the population eventually fixes at A, and $\lim_{n\to\infty}\ Q(n)/2N\ =\ 0$ if the population fixes at a. Thus

$$\begin{aligned}&\text{Prob(Population is eventually all } A's) \qquad (3.4)\\ &\quad=\ E(\lim_{n\to\infty} Q(n)/2N) = \lim_{n\to\infty} E(Q(n)/2N) = Q(0)/2N\end{aligned}$$

In other words, if $Q(0) = k$, the probability that the population fixes at A is equal to $k/2N$, which is the same as the initial frequency ratio of A in the population. It can also be shown that, as $n \to \infty$, the population fixes at the descendents of exactly one of the genes in the population at time $n = 0$, with each of the initial genes being equally likely to be chosen. This also implies that the probability of fixation at A is equal to $k/2N$.

The Time to Fixation: Now let's see if we can estimate approximately how long the population takes to fix at one of the two alleles. Let

$$I(n) = \Pr(\text{Two randomly chosen genes in generation } n \text{ are identical})$$

The expression $I(n)$ is also called the *inbreeding coefficient* at time n. Let $p_n = Q(n)/2N$ be the proportion of the allele A in the n^{th} generation. Then $H(n) = 1 - I(n)$ is the probability that two randomly chosen genes will be of different types. Note that we can write

$$H(n)\ =\ E\bigl(2p_n(1-p_n)\bigr) \tag{3.5}$$

since, if D is the event that the two genes chosen are of different types,

$$\begin{aligned} H(n) \;=\; P(D) \;&=\; \sum_{k=0}^{2N} P(D \text{ and } p_n = k/2N) \\ &=\; \sum_{k=0}^{2N} P(D \mid p_n = k/2N)\, P(p_n = k/2N) \\ &=\; \sum_{k=0}^{2N} 2(k/2N)(1-(k/2N))P(p_n = k/2N) \\ &=\; E\big(2p_n(1-p_n)\big) \end{aligned}$$

The function $H(n)$ in (3.5) is called the *probability of heterozygosity*. This term comes from the fact that, by the properties of random mating, it is also the probability that a randomly chosen individual is heterozygous (that is, Aa, as opposed to AA or aa).

The two randomly chosen genes in generation $n+1$ came from the same parental gene in generation n with probability $1/2N$, and sampled two different randomly chosen genes in generation n with probability $1-1/2N$. This implies the equation

$$I(n+1) \;=\; 1/2N + (1-1/2N)I(n) \tag{3.6}$$

Since $H(n) = 1 - I(n)$

$$H(n+1) \;=\; (1-1/2N)H(n)$$

and

$$\begin{aligned} H(n) \;&=\; (1-1/2N)H(n-1) \;=\; \ldots \\ &=\; (1-1/2N)^n H(0) \;\approx\; \exp(-n/2N)H(0) \end{aligned} \tag{3.7}$$

if N is large. This suggests that the population begins to fix at around $n \approx 2N$ generations, in the sense that $H(n)$ is close to $H(0)$ if $n/2N$ is small and close to zero if $n/2N$ is large. More precisely, we can say that $n = 2N$ is the *relaxation time* of $H(n)$ for large N. This means that $n = 2N$ is the additional time required for $H(n)$ to decrease by a factor of e.

Equation (3.7) does not address the question of whether the population fixes at either all As or all as. Let T_{2N} be the number of generations until the population is fixed at either A or a. Then $T_{2N} > n$ if and only if $0 < p_n < 1$, so that $P(T_{2N} > n) = P(0 < p_n < 1)$. Since $p_n = k/2N$ where k is an integer, we have that $p_n(1-p_n) \geq (1/2N)(1-(1/2N))$ if $0 < p_n < 1$. (*Exercise:* Prove this.) This implies

$$(2N+2)p_n(1-p_n) \geq \left(1+\frac{2}{2N}\right)\left(1-\frac{1}{2N}\right) = 1 + \left(\frac{1}{2N}\left(1-\frac{1}{N}\right)\right) \geq 1$$

if $N \geq 1$ and $0 < p_n < 1$. Similarly $p_n(1-p_n) < 1$ if $0 < p_n < 1$. Thus by (3.5)

$$H(n) \leq 2P(0 < p_n < 1) \leq (2N+2)H(n) \leq (2N+2)(1-1/2N)^n \tag{3.8}$$

since $H(0) \leq 1$. This implies the approximate inequality

$$P(T_{2N} > n) \;=\; P(0 < p_n < 1) \;\leq\; (N+1)\exp(-n/2N)$$

This implies $P(T_{2N} > n) < 1/2N$ for $n \geq 4N\log(2N)$, which is only a slightly slower rate of fixation than $2N$. However, a stronger result can be proven. Let W_{2N} be the number of generations until a population that begins with every gene a different type fixes at one of the $2N$ types. Then it can be shown that

$$P(W_{2N} > n) \;\leq\; 3\left(1-\frac{1}{2N}\right)^n \tag{3.9}$$

for all n and N (Kingman 1980, Appendix II, pp63–66). Thus the time to fixation is of the order of $2N$ generations.

Comments about real populations: Fixation by random genetic drift can take a very long time for a population of reasonable size with a long generation time. For example, a random mating population of 100 individuals with a generation time of 17 years would take ≈ 3500 years to fix at any given locus. However, most biological populations have a large number of loci with deleterious recessive alleles, so that deleterious effects may show up considerable sooner, and, if population sizes and generation times are smaller, fixation may happen much faster.

If the population had two sexes—that is, was dioecious rather than monoecious—one has to distinguish between genes within the same individual and genes in separate individuals in defining the analog of the recursion (3.6). For example, genes in the same individual have probability zero of having a common parental gene in the previous generation. (See the next section.)

Offspring distributions in real populations: An unrealistic aspect of random mating (that is, of the Wright-Fisher model) is that individuals in real populations may have an unusually large number of offspring for reasons independent of genotype, for example as result of being the first-born in a litter. In this case, the choices of parents by various offspring would not be independent, since a parent that fathered (or mothered) one offspring would be more likely to parent others.

Statistically, this would show up as a larger variance in the offspring distribution. If choices of parents are independent, the number of surviving offspring of an individual is a binomial random variable, for which the ratio of variance to mean is always less than one. There is evidence that this ratio in real populations is usually greater than one, with ratios of three or more for fruit flies in one study (Crow and Morton 1955). Conjecture among some biologists is that, in fact, the female fruit fly that lays her eggs closest to the light bulb in the experimental cage has a disproportionate number of surviving offspring.

1.4. Random Mating with Two Sexes

The standard Wright-Fisher model assumes a population with only one sex (that is, monoecious) such that, in each generation, the probability of selfing (that is, that the same individual is both mother and father) is exactly $1/N$. How reasonable are the conclusions for a more realistic model of mating, for example a dioecious population?

One immediate difference is that, in any population in which selfing is excluded, the two genes in one individual have probability zero of coming from the same gene or individual in the previous generation. Thus we must consider *two different* inbreeding coefficients, one (call it $I_1(n)$) for a randomly chosen pair of genes from *different individuals* in the n^{th} generation and the second (call it $I_2(n)$) for the two genes in *one* randomly chosen individual.

Since individuals of different sexes choose their parents in the same way under random mating, $I_1(n)$ (for two individuals) will be the same whether the two individuals are both male, are both female, or are from different sexes, at least for generations $n \geq 1$. Similarly, $I_2(n)$ (for one individual) will be the same for individuals from either sex (also for $n \geq 1$).

Now assume that the population is held at exactly N_1 males and N_2 females in each generation. Let's derive equations for how the probabilities $\big(I_1(n+1), I_2(n+2)\big)$ depends on $\big(I_1(n), I_2(n)\big)$ from the previous generation. First, consider two randomly chosen genes from different individuals in generation $n+1$. It doesn't matter whether these are randomly chosen from males only, from females only, or from the entire population; the recurrence for $I_1(n+1)$ will be the same. With probability $1/4$, the two genes came from two male parents, with probability $1/4$ from two female parents, and with probability $1/2$ from parents of different sexes. If they came from the same sex (for example, males), then the (conditional) probability that they came from the same individual is $1/N_1$ and are the same gene is $1/2N_1$. If they came from different sexes, then they must have come from different individuals. Similarly, $I_2(n+1)$ is the same as the probability that two genes from different sexes in the previous generation are the same, so that $I_2(n+1) = I_1(n)$ (if $n \geq 1$). This leads to the recurrence

$$\begin{aligned} I_1(n+1) &= A + BI_1(n) + CI_2(n) \\ I_2(n+1) &= I_1(n) \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} A &= (1/4)(1/2N_1) + (1/4)(1/2N_2) \\ B &= (1/4)(1-1/N_1) + (1/4)(1-1/N_2) + (1/2) \\ &= 1 - (1/4)(1/N_1) - (1/4)(1/N_2) \\ C &= A \;=\; (1/4)(1/2N_1) + (1/4)(1/2N_2) \end{aligned} \tag{4.2}$$

In (4.1), $I_1(n)$ on the right-hand side of the equation is for a randomly chosen pair of gene from *different sexes* from the parents, since each individual in a doecious population has two parents, one of each sex. In contrast, $I_1(n+1)$ on the left-hand side is for any two individual offspring, whether two males, two females, one of each sex, or a random pair from the entire population. This means that $I_1(n)$ is the same for two male offspring, for two females, or for one of each sex for $n \geq 1$, assuming that the initial state is $n = 0$. Similarly, $I_2(n)$ for generations $n \geq 1$ is the same for the two genes in one male, in one female, or from a randomized choice from all offspring.

Continuing, we can write (4.1) in matrix notation as

$$\begin{pmatrix} I_1(n+1) \\ I_2(n+1) \end{pmatrix} = \begin{pmatrix} A \\ 0 \end{pmatrix} + \begin{pmatrix} B & C \\ 1 & 0 \end{pmatrix} \begin{pmatrix} I_1(n) \\ I_2(n) \end{pmatrix} \tag{4.3}$$

Since $A + B + C = 1$ in (4.2) and(4.3), we can also write

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} A \\ 0 \end{pmatrix} + \begin{pmatrix} B & C \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Let $H_1(n) = 1 - I_1(n)$ and $H_2(n) = 1 - I_2(n)$. Then by subtraction

$$\begin{pmatrix} H_1(n+1) \\ H_2(n+1) \end{pmatrix} = \begin{pmatrix} B & C \\ 1 & 0 \end{pmatrix} \begin{pmatrix} H_1(n) \\ H_2(n) \end{pmatrix} = M \begin{pmatrix} H_1(n) \\ H_2(n) \end{pmatrix}$$

where $M = \begin{pmatrix} B & C \\ 1 & 0 \end{pmatrix}$. Thus by induction

$$\begin{pmatrix} H_1(n+1) \\ H_2(n+1) \end{pmatrix} = M^n \begin{pmatrix} H_1(1) \\ H_2(1) \end{pmatrix} \tag{4.4}$$

This implies that $(H_1(n), H_2(n)) \to 0$ at a rate that is determined by the largest eigenvalue of M.

The characteristic polynomial of M is

$$\phi(\lambda) = \det(M - \lambda I) = (B - \lambda)(-\lambda) - C = \lambda^2 - B\lambda - C$$

for B, C in (4.2). The eigenvalues of M are

$$\lambda_1 = \frac{B + \sqrt{B^2 + 4C}}{2} \qquad \text{and} \qquad \lambda_2 = \frac{B - \sqrt{B^2 + 4C}}{2} \tag{4.5}$$

where $B > 0$ and $C > 0$. In particular, $\lambda_2 < 0$ and $|\lambda_2| < \lambda_1$. If N_1 and N_2 are large, then λ_1 is close to one and λ_2 is close to zero.

It follows from (4.4) that

$$\begin{pmatrix} H_1(n) \\ H_2(n) \end{pmatrix} \sim C_1\lambda_1 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \qquad \text{where} \qquad M\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \lambda_1 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \tag{4.6}$$

for some $C_1 > 0$. It is customary to say that a population genetics model with a property like (4.6) is like the standard Wright-Fisher model, but, in analogy with equation (3.7) $\big(H(n+1) = (1 - 1/2N)H(n)\big)$, has an *effective (Wright-Fisher) population size of* N_e instead of N, where N_e is defined by

$$\lambda_1 = 1 - 1/2N_e \tag{4.7}$$

(More precisely, N_e is the *inbreeding* effective population size.)

If N_1 and N_2 are large, then B in (4.2) and (4.5) is close to one and A and C are small. One can show that $\sqrt{1+x} = 1 + x/2 + O(x^2)$ for small x, where $O(x)$ denotes an arbitrary expression that is bounded by a constant times x. (This is called *Landau's big-O* notation.) By (4.2),

$$B = 1 + O(k_N), \quad A = O(k_N), \quad C = O(k_N)$$

for $k_N = (1/N_1) + (1/N_2)$. In particular, within terms that are $O(k_N^2)$,

$$\begin{aligned} \sqrt{B^2+4C} \ &= \ B\sqrt{1+4C/B^2} \ = \ B(1+2C/B^2) \\ &= \ B + 2C/B \ = \ B + 2C \end{aligned}$$

Thus

$$\lambda_1 \ = \ \frac{B+\sqrt{B^2+4C}}{2} \ = \ \frac{B+B+2C}{2} \ = \ B + C \ = \ 1 - A + O(k_N^2)$$

in the same sense. By the same reasoning,

$$\lambda_2 \ = \ \frac{B-\sqrt{B^2+4C}}{2} \ = \ \frac{B-(B+2C)}{2} \ = \ -C + O(k_N^2)$$

It follows that the effective population size for a dioecious population is approximately

$$N_e \ \approx \ \frac{1}{2(1-\lambda_1)} \ \approx \ \frac{1}{2A} \ = \ \frac{4}{1/N_1 + 1/N_2} \ = \ \frac{4N_1N_2}{N_1+N_2} \tag{4.8}$$

In particular, if the subpopulations of the two sexes are the same size with $N_1 = N_2 = m$, then $N_e = 4m^2/(2m) = 2m = N_1 + N_2$ and N_e is the same as the total population size. However, I claim that (4.8) is a more reasonable definition of the population size in general than $N = N_1 + N_2$ from the point of view of the random fixation of genes.

For example, assume that $N_1 \ll N_2$ (that is, N_1 is much smaller than N_2), as would be the case for animals in which only a few dominant males have most of the offspring and most males do not contribute to the next generation. Then

$$N_e \ \approx \ \frac{4N_1N_2}{N_1+N_2} \ = \ 4N_1\big(1 + O\big((N_1/N_2)\big)\big) \ \approx \ 4N_1$$

and N_e is essentially determined by the male population size. This is because, due to the smaller male population, most random fixation of genes occur in the male population, and fixations in the much larger female population can essentially be ignored. The factor of 4 in (4.8) corresponds to the fact that, in the ancestry of a pair of genes, the two ancestral genes will both be in the male population about 1/4 of the time. Of course, random fixations can still occur in the female population as long as $N_2 < \infty$. This shows up in (4.8) as the fact that N_e is slightly smaller than $4N_1$ if $N_1 \ll N_2$.

Exercise 4.1. Prove (4.6).

1.5. Mutation

Mutation is an error of replication between a gene in an offspring and the corresponding parental gene, which could be due to either a change occurring at conception or else a change in a germ cell while being carried by the parent.

Assume we are following an autosomal locus with two selectively equivalent alleles, A and a, in a diploid monoecious population held at N individuals. For each juvenile gene in each generation, assume that there is a probability u of a mutation from an a to an A (i.e., of a juvenile receiving an A in place of an original parental a), and a probability v of a mutation from an A to an a. Let $Q(n)$ be the number of genes of type A at the beginning of the n^{th} generation. As in Section 3, the distribution of the frequency ratio $p_{n+1} = Q(n+1)/2N$ given $p_n = Q(n)/2N$ can be represented symbolically as

$$\{\, p_{n+1} \mid p_n = p \,\} \;\approx\; \frac{B\big(2N, f(p)\big)}{2N} \tag{5.1}$$

where $B(2N,p)$ denotes a binomial distribution and

$$\begin{aligned} f(p) \;&=\; (1-v)p + u(1-p) \\ &=\; (1-u-v)p + u \end{aligned} \tag{5.2}$$

By (5.1), $\{p_n\}$ is a Markov chain with state space $S(2N) = \{0, 1/2N, \ldots, 1\}$. If $u, v > 0$, the Markov chain has no traps. Then $\{p_n\}$ is an irreducible Markov chain on $S(2N)$ and thus has a stationary distribution $\mu(2N, dp)$ on $S(2N)$. Stationarity means that if Z_N is a random variable with distribution $\mu(2N, dp)$ on S_N that is independent of p_n, then the conditional distribution of p_{n+1}

$$\{\, p_{n+1} \mid p_n = Z_N \,\} \;\approx\; Z_N \;\approx\; p_n \tag{5.3}$$

is the same as p_n. Since $S(2N)$ is a subset of the unit interval [0,1], the stationary distribution $\mu(2N, dp)$ can be viewed as a probability measure on [0,1].

Let p_∞ be the fixed point of $f(p)$ in (5.2); i.e. the solution of

$$p_\infty = f(p_\infty) = (1-u-v)p_\infty + u \tag{5.4}$$

which is $p_\infty = u/(u+v)$. For an infinitely large population, that is in the limit as $2N \to \infty$, the frequency ratios p_n in (5.1) become a deterministic sequence with $p_{n+1} = f(p_n)$. Then by subtracting (5.4) from (5.2)

$$\begin{aligned} p_{n+1} - p_\infty &= (1-u-v)(p_n - p_\infty) \\ &= (1-u-v)^{n+1}(p_0 - p_\infty) \end{aligned}$$

In particular, $p_n \to p_\infty$ for $p_\infty = u/(u+v)$ in (5.4). The time scale of this convergence is of order $n \approx 1/(u+v)$ with relaxation time $n = 1/(u+v)$ for small u, v, exactly as in (3.7).

There are $\approx 2N(u+v)$ mutations on the average in each generation. The mutation rates u, v for a genetic locus may be in the range $10^{-4} - 10^{-6}$, depending on the organism. Mutation rates per site can be much smaller, for example $u, v \approx 10^{-10}$ for fruit flies. The population size $2N$ can be of order $10^3 - 10^6$ for endangered or semi-endangered species or $\approx 10^9$ for a local population of a bacterium like *Escherichia coli*. In most cases, $2N$ is large, u, v are small, but $2N(u+v)$ (the number of mutations per generations) is of order one. Thus it is natural to scale the mutations rates u, v by the population size, as in

$$u \sim \frac{\alpha}{2N} \quad \text{and} \quad v \sim \frac{\beta}{2N} \quad \text{as} \quad 2N \to \infty \quad \text{where} \quad \alpha, \beta > 0 \tag{5.5}$$

Mutation and genetic drift then act on the same time scale, and so should have about the same strength. Thus, for large $2N$, the stationary distribution $\mu(2N, dp)$ for (5.1) should show the effects of both mutation and genetic drift.

1.6. Approximating $\mu(2N, dp)$ for large N

The stationary distribution $\mu(2N, dp)$ of (5.1) can be found explicitly only for small values of $2N$. However, it turns out that, if the mutation rates are scaled by (5.5), the probability measures $\mu(2N, dp)$ converge as $2N \to \infty$ to a limiting continuous probability measure $\mu(dp)$ on $[0, 1]$ that one can calculate.

Consider stationary probability measures $\mu(2N, dp)$ for a sequence of values $\big(N(k), u(k), v(k)\big)$ that satisfy (5.5) for $N(k) \to \infty$ as $k \to \infty$. By the Helly-Bray theorem, some subsequence of these measures will have a limiting probability measure $\mu(dp)$ on [0,1]. If the limiting probability measure $\mu(dp)$ is the same for all sequences $\big(N(k), u(k), v(k)\big)$ (that is, if $\mu(dp)$ is unique), then the sequence $\mu(2N, dp)$ will itself converges to $\mu(dp)$.

The limiting probability measure $\mu(dp)$ will turn out to be a continous measure in this case; specifically, $\mu(dp) = f(p)dp$ for some function $f(p)$. This means that the probability distribution of the proportion p of genes of type A for large N will be approximately $\mu(dp) = f(p)dp$. The next step will be to derive an equation for $f(p)$.

Let Z_N be a random variable with distribution $\mu(2N, dp)$. Then, by (5.3), if $g(p)$ is an arbitrary three-times continuously differentiable function on [0,1],

$$\begin{aligned} &E\big(E(g(p_{n+1}) \mid p_n \approx Z_N)\big) - E\big(g(Z_N)\big) \;=\; 0 \\ &= \int_0^1 E\big(g(p_{n+1}) - g(p_n) \mid p_n = p\big)\; \mu(2N, dp) \end{aligned} \tag{6.1}$$

It follows from (5.1), (5.2), (5.5), and the identity $E(X^2) = \mathrm{Var}(X) + E(X)^2$ that as $N \to \infty$ (or $k \to \infty$)

$$\begin{aligned} 2N\, E(\;\; p_{n+1} - p \;\;\mid p_n = p) \;&=\; 2N\big(f(p) - p\big) \\ \to\; m(p) \;&=\; \alpha - (\alpha + \beta)p \\ 2N\, E\big((p_{n+1} - p)^2 \mid p_n = p\big) \;&=\; f(p)\big(1 - f(p)\big) + 2N\big(f(p) - p\big)^2 \\ \to\; a(p) = p(1-p), &\qquad \text{and} \\ 2N\, E\big(|p_{n+1} - p|^3 \mid p_n = p\big) \;&\to\; 0 \end{aligned} \tag{6.2}$$

uniformly in p as $2N \to \infty$. If we expand $g(p)$ in the identity (6.1) in a four-term Taylor expansion about p and apply (6.2), we conclude

$$\begin{aligned} &2N \int_0^1 E\big(g(p_{n+1}) - g(p_n) \mid p_n = p\big)\; \mu(2N, dp) \\ &\to \int_0^1 \Big({}^1\!/_2 a(p) g''(p) + m(p) g'(p) \Big)\, \mu(dp) = 0 \end{aligned} \tag{6.3}$$

Equation (6.3) holds for all functions $g(p)$ on [0,1] that are three times continuously differentiable. The integration by parts formula

$$\int_0^1 g'(p)m(p)\mu(dp) \;=\; g'(1)\int_0^1 m(x)\mu(dx) - \int_0^1 g''(p)\int_0^p m(x)\mu(dx)\, dp \tag{6.4}$$

holds for arbitrary measures $\mu(dp)$ on [0,1], providing that the integrals in (6.4) are over closed intervals to allow for possible atoms of $\mu(dp)$ at the endpoints. Substituting (6.4) in (6.3),

$$\int_0^1 g''(p)\left({}^1\!/_2 a(p)\mu(dp) - \int_0^p m(x)\mu(dx)\, dp \right) \;+\; g'(1)\int_0^1 m(x)\mu(dx) = 0 \tag{6.5}$$

The substitution $g(p) = \int_p^1 (y-p)f(y)dy$ yields

$$\int_0^1 f(p)\,\kappa(dp) = 0$$

for arbitrary smooth functions $f(p)$ on [0,1] where

$$\kappa(dp) \;=\; {}^1\!/\!_2 a(p)\mu(dp) - \int_0^p m(x)\mu(dx)\,dp$$

is the measure in the first integral in (6.5). Since this holds for all smooth functions on [0,1], the measure $\kappa(dp) = 0$. After dividing by ${}^1\!/\!_2 a(p)$, the measure

$$\mu(dp) = f(p)\,dp \qquad \text{for} \qquad f(p) = \frac{2}{a(p)}\int_0^p m(x)\mu(dx)$$

for $a(p)$ and $m(p)$ in (6.2) ($0 < x < 1$). Similarly

$$f(p) = \frac{2}{a(p)}\int_0^p m(x)f(x)\,dx$$

and hence $f(x)$ is a solution of the differential equation

$${}^1\!/\!_2 \big(a(p)f(p)\big)' - m(p)f(p) = 0$$

again for $a(p)$ and $m(p)$ in (6.2). If we solve this equation for $f(p)$ with the constant of integration determined by $\int_0^1 \mu(dp) = \int_0^1 f(p)dp = 1$, we conclude

$$f(p) \;=\; \frac{\Gamma(2\alpha{+}2\beta)}{\Gamma(2\alpha)\Gamma(2\beta)}\, p^{2\alpha-1}(1-p)^{2\beta-1} \tag{6.6}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$ is the gamma function.

The density in (6.6) is called the beta distribution with parameters 2α and 2β. We have now shown that any limiting distribution of $\mu(2N(k), dp)$ as $2N(k) \to \infty$ with $u(k), v(k)$ subject to (5.5) must be the beta distribution (6.6). Hence, by uniqueness, the full sequence $\mu(2N, dp)$ converges weakly to the beta distribution (6.6).

Examples: (i) If $\alpha, \beta = 2Nu, 2Nv \ll 1$ and $2N$ is large, the beta density (6.6) (and *a posteriori* the stationary distribution $\mu(2N, dp)$ for large N) is concentrated near the endpoints. In that case the frequency p_n of the allele A is most likely to be close to either zero or one—i.e., the population spends most of its time essentially fixed at either one allele or the other (but not always the same allele). In general

$$\lim\; E(p_n) \;=\; \int_0^1 ph(p)dp \;=\; \frac{\alpha}{\alpha+\beta} \tag{6.7}$$

by (6.6), where the limit is taken as $n \to \infty$ and then $N \to \infty$. This suggests that, for large n and N, the population will be nearly fixed for A with probability $\alpha/(\alpha+\beta)$ and nearly fixed for a with probability $\beta/(\alpha+\beta)$.

(ii) If $2Nu, 2Nv \gg 1$ and $2N$ is large, most of the time is spent near $p_\infty = u/(u+v)$. Note that this is also consistent with (6.7).

Migration and Mutation: The results in this section can be viewed in another way. Suppose we have an island which is situated between two mainlands. One mainland is fixed at the allele A; the other is fixed at a. Juveniles from both mainlands migrate to the island in sufficient number so that fractions u, v of the total number of juveniles on the island are migrants from the two mainlands respectively. In each generation, a total of $2N$ haploid juveniles survive to become adults.

These represent a random choice from the juveniles which are present at the beginning of the generation. Assume for definiteness that the frequency of genes of type A at the beginning of the $n^{\rm th}$ generation is p. Then the total number of A-genes at the beginning of the next generation is a binomial variable based on $2N$ trials with probability of success $(1-u-v)p+u$, which is exactly the same as (5.2). Thus this kind of migration is mathematically equivalent to mutation.

Thus, if u, v satisfy (5.5) and $2N$ is large, the frequency of the subpopulation of the island which originally came from the first mainland randomly varies and has the beta distribution (6.6). Note that (5.5) implies that, on the average, $\alpha \approx 2Nu$ juvenile immigrants from the first mainland survive to adulthood, and $\beta \approx 2Nv$ from the second. Thus, no matter how large $2N$ is, a few surviving immigrants per generation keep the island population from fixing at one allele or the other. If we had two islands rather than one island and two mainlands, this would suggest

General Principle. *The exchange of one or two migrants per generation is enough to keep the genetic structure of two random-mating populations from becoming quite different by genetic drift.*

Exercise 6.1. Prove the relations (6.2). (The third relation is the hardest.)

Exercise 6.2. Verify the integration by parts formula (6.4).

1.7. The Method of Diffusion Approximation

Another way of viewing the Kolmogorov relations (6.2) is that they relate the Markov chain p_n with a *diffusion process* X_t for $t \approx n/2N$ whose *infinitesimal generator* is

$$\mathcal{L}_X\, g(p) = {}^1\!/_2\, a(p)\, g''(p) + m(p)\, g'(p) \tag{7.1}$$

There is a general theory about diffusion processes of this type. In this case, the diffusion process X_t with generator $\mathcal{L}_X$ can be shown to have the beta distribution (6.6) as its stationary distribution. Quantities such as the expected time to fixation of one allele in the absence of mutation, or the relative probabilities of fixation as a function of the initial frequency, can also be approximated by the limiting diffusion process. This is called the method of *diffusion approximation* (of Markov chains), and is widely used in population genetics and in applied mathematics generally.

The argument from (6.3)–(6.6) is a special case of a more general procedure. If the relation (6.3)

$$\int_0^1 \Big({}^1\!/_2 a(p) g''(p) + m(p) g'(p) \Big)\, \mu(dp) = 0 \tag{7.2}$$

holds for all smooth functions $g(p)$ that vanish near the endpoints, then the measure $\mu(dp)$ is said to be a *weak solution* of the equation

$$\begin{aligned} &\mathcal{L}_X^* \mu(dp) = 0 \qquad \text{where} \\ &\mathcal{L}_X^* h(p) = {}^1\!/_2 \big(a(p)h(p)\big)'' - \big(m(p)h(p)\big)' \end{aligned} \tag{7.3}$$

Note that if we did have $\mu(dp) = h(p)\, dp$ where $h(p)$ was smooth, then (7.2) and integration by parts would imply

$$\int_0^1 g(p) \mathcal{L}_X^* h(p)\, dp = 0$$

for all smooth $g(p)$, which would imply $\mathcal{L}_X^* h(x) = 0$.

If $a(p)$ and $m(p)$ are smooth, then *Weyl's Lemma* states that all weak solutions of (7.3) are in fact of the form $\mu(dp) = h(p)dp$ where $h(p)$ is a smooth solution of $\mathcal{L}_X^* h(p) = 0$ (see e.g. McKean 1969, p85). (In fact, Weyl's Lemma holds in an arbitrary number of dimensions.) In our case, (7.2) holds for all smooth functions on [0,1], which allowed us to eliminate one of the constants of integration in (7.3).

1.8. Selection and the Hardy-Weinberg Law

Consider an autosomal genetic locus with two alleles A, a in a population which is held at N_1 males and N_2 females. Darwinian selection means that the population proportion of an allele in the adults of the next generation can depend on the effects of that allele on the individuals which carry it.

This can be modeled in the Wright-Fisher model by assuming that genes in offspring are found by biased independent random sampling from genes in the parental generation, where the bias depends on the genotypes at birth in the parental generation. Let $p(z, AA)$, $p(z, Aa)$, and $p(z, aa)$ be the population proportions of the genotypes AA, Aa, and aa at birth in the two sexes, where z = 'm' or 'f' for 'mother' or 'father'. In general, the *fitness* of a gene or genotype is the relative number of surviving offspring in the next generation that are descendents of that gene or genotype. In this context, we implement selection by defining *fitness constants* $w(AA)$, $w(Aa)$, and $w(aa)$ for the three genotypes and assume that a particular juvenile has a mother of type AA with probability $w(AA)p(m, AA)/C(m)$, where

$$C(m) = w(AA)p(m, AA) + w(Aa)p(m, Aa) + w(aa)p(m, aa)$$

is a normalizing constant. (For simplicity, we assume that the fitness constants $w(AA), w(Aa), w(aa)$ do not depend on sex.)

If the mother is Aa, the probability that the maternally-derived gene for that juvenile is A is 1/2 by Mendelism. In general, let $pn(z, G)$ be the probability that the z-derived gene in a newborn is G, where G = A or a. Then

$$pn(m, A) = \big(w(AA)p(m, AA) + {}^1\!/\!_2 w(Aa)p(m, Aa)\big)/C(m) \qquad (8.1)$$

We assume that the choices of the two parents are independent. This amounts to assuming that selection acts on potential parents individually rather than as mating pairs, as well as a lack of mating preferences that depend on genotype. In particular, this model would apply to both viability and fertility selection, as long as either type of selection depends on individual parents and not on mating pairs.

Since a juvenile genotype has a maternally-derived gene and a paternally-derived gene, the probability distribution of the genotypes of newborns of either sex is then

$$\begin{aligned} pn(AA) &= pn(m, A)pn(f, A) \\ pn(Aa) &= pn(m, A)\,pn(f, a) + pn(m, a)pn(f, A) \\ pn(aa) &= pn(m, a)\,pn(f, a) \end{aligned} \qquad (8.2)$$

If the frequencies of the parental genotypes are independent of sex, then by (8.1)

$$\begin{aligned} pn(AA) &= \quad pn(A)^2, \\ pn(Aa) &= 2\,pn(A)\,qn(A), \qquad qn(A) = 1 - pn(A), \\ pn(aa) &= \quad qn(A)^2 \end{aligned} \qquad (8.3)$$

where $pn(A) = pn(f, A) = pn(m, A)$ and $qn(A) = pn(f, a) = pn(m, a)$. The relations (8.3) are known as the *Hardy-Weinberg* law.

If $pn(m, A) \neq pn(f, A)$, then (8.3) has no obvious meaning since it is not clear what $pn(A)$ might mean if the sexes have different population sizes. In general, we say that genotype frequencies $pn(AA), pn(Aa), pn(aa)$ are in *Hardy-Weinberg proportions* if

$$\begin{aligned} pn(AA) &= \ r^2 \\ pn(Aa) &= 2rs \\ pn(aa) &= \ s^2 \end{aligned} \tag{8.4}$$

for numbers $r, s \geq 0$. It follows from (8.4) that

$$pn(AA) + pn(Aa) + pn(aa) = 1 = r^2 + 2rs + s^2 = (r + s)^2 = 1$$

so that $r + s = 1$. Then $r = pn(AA) + {}^1\!/\!_2 pn(Aa)$ by addition in (8.4). However, if (8.4) and

$$\begin{aligned} pn(AA) &= pn(m, A)pn(f, A) \\ pn(Aa) &= pn(m, A)\,pn(f, a) \,+ pn(m, a)pn(f, A) \\ pn(aa) &= pn(m, a)\,pn(f, a) \end{aligned} \tag{8.5}$$

both hold, then one can prove $pn(m, A) = pn(f, A) = r$ so that the Hardy-Weinberg law (8.3) holds unambiguously. (See Exercise 8.1 below.)

The Hardy-Weinberg law (8.3) implies that genotype frequencies are a function of gene frequencies, and that one can follow the fate of genes in populations and not worry about the genotypes that carry them. If the Hardy-Weinberg laws hold for the parental genotype frequencies as well, so that $p(m, AA) = p(f, AA) = p^2$, $p(m, Aa) = p(f, Aa) = 2p(1 - p)$ etc. in (8.1), then $pn(A)$ in (8.3) equals

$$pn(A) \ = \ p\,\frac{w(AA)p + w(Aa)q}{w(AA)p^2 + w(Aa)2pq + w(aa)q^2} \tag{8.6}$$

where $q = 1 - p$.

If the population size is infinite, the probabilities $pn(AA), \ldots$ in (8.2) are the frequencies in the next generation. Then if the parental genotype frequencies $p(z, AA), \ldots$ are independent of sex, the Hardy-Weinberg law (8.3) holds after one generation. If $p(z, AA)$ etc. do depend on sex, then by (8.2) they are independent of sex after one generation (assuming sex-independent fitnesses), and (8.3) holds after two generations.

If the fitnesses $w(AA), w(Aa), w(aa)$ refer to actual parental survival and not to differences in fertility, note that the Hardy-Weinberg law (8.3) does not hold in general for the parental genotypes at the mating stage; i.e., after viability selection has been applied. These adult population proportions are

$$p(AA) = \frac{w(AA)p^2}{C}, \quad p(Aa) = \frac{w(Aa)2pq}{C}, \quad p(aa) = \frac{w(aa)q^2}{C}$$

These will satisfy the Hardy-Weinberg law (8.3) if and only if the fitnesses are multiplicative functions of the alleles in the genotype; that is, if

$$w(AA) = w(A)^2, \quad w(Aa) = w(A)w(a), \quad \text{and} \quad w(aa) = w(a)^2 \tag{8.7}$$

The case (8.7) is called *genic selection* as opposed to *genotypic selection*.

For finite populations, the offspring probabilities $pn(AA)$, $pn(a)$, etc. are probabilities and not actual frequencies. The true genotype and gene frequencies among the juvenile population will be random variables with these numbers as expected values given the previous generation. Thus (8.2)–(8.3) will almost never be true for the genotype frequencies themselves. Similarly, although the genotype frequencies for the different sexes may have the same means given the preceding generation, they will usually be different.

While this model of viability selection is conceptually simple and easy to work with, it is not realistic for many situations in which selection occurs. The fitness of an adult may depend not only on its genotype, but also on the genotypes of the other adults. This would lead to frequency-dependent fitnesses that depend on the population proportions of genotypes.

Examples:
Altruism: One example of frequency-dependent fitnesses is altruism. Suppose that fitnesses depend on the behavior of pairs of individuals, either of which can help the other. If a helpful individual links with a non-helpful individual, the fitness of the helpful individual is decreased and that of the non-helpful individual is increased. However, if both are helpful, the fitness of both individuals is greatly increased. Suppose that 'helpfulness' is governed by a locus with two alleles, where AA-individuals are always helpful, aa-individuals are never helpful, and, for definiteness, Aa's are sometimes helpful and sometimes not. Then, if the allele A is rare, individuals carrying it will be selected against, since they will be taken advantage of. However, if genes of type A are common, then AA and Aa individuals will often find valuable liaisons and have a great selective advantage.

Toxic bacteria: A similar example occurs in bacteria. Suppose that one genotype excretes a particular poison or toxin to which it itself is immune, but, due to the effort involved in making and defending against the toxin, is otherwise less fit than normal bacteria. If this genotype is common, this strategy may succeed. However, if it is rare, there will be a lower level of toxin in the environment. The lower level of toxin may not cause enough damage to the normal bacteria to compensate for the upkeep in making the toxin, so that the toxin-producing genotype will be at a net selective disadvantage. (This is also a kind of altruism.)

In both these examples, geographical distribution may be quite important if genes of type A are rare. By chance (e.g. genetic drift) the non-normal genotypes may find themselves in the majority in a localized area, and take it over. They may then spread throughout the habitat by having a sufficient frequency in border areas, and in this way an allele which is disadvantageous when rare may eventually dominate the population. As an example, many ecologists feel that new species of animals or plants cannot arise without geographical effects, essentially because the first representatives of an emerging species would have too much difficulty finding mates.

Sickle-cell anemia: A final example concerns sickle-cell anemia. Assume that the sickling gene S is relatively rare in a malarial area. By itself, an individual of type SS will have a very low fitness. However, the two parents of the individuals were probably both SN, which suggests that a type-S individual has a relatively larger percentage of close relatives who are of the relatively rare type SN. Thus, if the individual is able to live long enough to have children, and the aunts and uncles of these children are willing to help look after them, then offspring of type SS individuals may have a much higher probability of surviving than the offspring of NN individuals. This may result in a higher fitness of SS individuals than of normal NN individuals, even though most SS individuals may be sickly throughout life and die at a relatively young age. All three of these examples show the difficulty of applying naïve Darwinian arguments to social species.

Exercise 8.1. (i) Show that Hardy-Weinberg proportion (8.4) is equivalent to the relation $pn(Aa)^2 = 2pn(AA)pn(aa)$.

(ii) Prove that equations (8.4) and (8.5) together imply that $pn(m, A) = pn(f, A) = r$ and $pn(m, a) = pn(f, a) = s$.

Exercise 8.2. A more general model of selection would allow fitnesses $w(z, AA)$ (z = 'f' or 'm') that could depend on the sex of the individual carrying that genotype. Examples would be genes active in one sex only, for example genes affecting courtship behavior or plumage in males, or genes affect response to courtship behavior in females. Find the analogs of (8.1)–(8.2).

Using these equations, show that, in general, the Hardy-Weinberg law (8.3) or (8.4) will never occur for population frequency proportions at birth unless the parental fitnesses are in fact proportional for both sexes. That is, if

$$\frac{w(m,AA)}{w(f,AA)} = \frac{w(m,Aa)}{w(f,Aa)} = \frac{w(m,aa)}{w(f,aa)} = C \tag{8.8}$$

Exercise 8.3. What happens if the locus is on the X-chromosome?

(i) Find the analog of (8.2). (Note that males have haploid genotypes A and a!)

(ii) Show that, in the absence of selection, (a) the female genotype frequencies $pn(AA)$ etc. may never satisfy the Hardy-Weinberg law (8.4) for any r, although (b) the deviation from Hardy-Weinberg proportions converges to zero exponentially fast in the number of generations.

Exercise 8.4. (Part (iii) below is difficult.) Suppose that the A-allele is represented by only one gene in a large monoecious population of N individuals, but that it carries a selective advantage over the allele a in heterozygote form. (I.e., $w(Aa) = w > 1,\ w(aa) = 1$.) The number of A-type genes in the next generation has a binomial distribution with parameter p given by (8.1). Let $N \to \infty$, keeping the number of genes of type A initially at one.

Note that in the second and all succeeding generations, all A-type genes will be carried in individuals of genotype Aa because individuals who both carry A's will be too rare to meet. Thus all matings will be either $aa \times aa$ or $Aa \times aa$.

Prove that

(i) In the limit as $N \to \infty$, the number of copies X of the A-gene in the second generation has a Poisson distribution with mean w.

(ii) If there are r different A genes in a subsequent generation, prove that the number of descendents $X_1, X_2, \ldots, X_r$ of the r different A-genes in the next generation have a joint distribution for which, in the limit as $N \to \infty$, the $X_1, X_2, \ldots, X_r$ are independent Poisson.

Part (ii) means that number $Q(n)$ of A-genes in the n^{th} generation forms a *branching process* with a Poisson offspring distribution. Using the theory of branching processes (Karlin and Taylor 1975, Athreya and Ney 1972, Crow and Kimura 1970, pp419–423), prove that

(iii) If the relative fitness $w = w(Aa) = 1 + s$ where $s > 0$, where $w(aa) = 1$, the probability that this branching processes does not eventually become extinct equals $2s$ within terms of order s^2 for small s.

This is a result attributed to R. A. Fisher: Specifically, that if A is a new, selectively advantageous allele whose heterozygote has relative fitness $w = 1 + s$, then the probability that A does not become lost in the first few generations due to genetic drift is $2s + O(s^2)$. In fact, one can show more exactly that the probability is $2s - (8/3)s^2 + (28/9)s^3 + O(s^4)$ (A. Amei, personal communication).

1.9. Driving Bad Genes Out of a Large Population

As mentioned earlier, one might expect that most severely deleterious genes would be recessive. However, there do exist dominant deleterious genes in humans, for example the genes that produce achondroplastic dwarfism, Huntington's chorea, and certain types of muscular dystrophy. A model that includes both dominant and recessive deleterious alleles has

$$w(AA) = w = 1 - s, \qquad w(Aa) = 1 - hs, \qquad w(aa) = 1 \tag{9.1}$$

where $s > 0$ and the relative fitnesses have been normalized by setting $w(aa) = 1$. Dominant deleterious alleles correspond to $h = 1$, and recessives to $h = 0$. If the fitnesses (9.1) hold for both sexes in a large population, the genotype frequencies at birth will be in Hardy- Weinberg equilibrium.

Let p_n be the frequency of the deleterious allele A at the beginning of the $n^{\rm th}$ generation. By (8.6)

$$\begin{aligned} p_{n+1} \ &= \ f(p_n) \ = \ p\,\frac{1-s(p+hq)}{1-sp(p+2hq)} \qquad (9.2)\\ &= \ p\,\frac{1-s\big(h+p(1-h)\big)}{1-sp\big(2h+p(1-2h)\big)} \end{aligned}$$

where $p = p_n$ and $q = q_n = 1 - p_n$. Thus $p_{n+1} = f(p_n) = f^{[n+1]}(p_0)$, where $f^{[n]}$ denotes the $n^{\rm th}$ iterate of $f(p)$ in (9.2). If A is dominant (i.e. $h = 1$),

$$f(p) \ = \ p\,\frac{1-s}{1-sp(2-p)} \ = \ \frac{wp}{wp(2-p)+(1-p)^2} \qquad (9.3)$$

If A is recessive (i.e. $h = 0$),

$$f(p) \ = \ p\,\frac{1-sp}{1-sp^2} \qquad (9.4)$$

In both cases $0 < f(p) < p < 1$ for $0 < p < 1$ and $p_{n+1} = f(p_n)$. Hence $p_n \downarrow$, and $f(p)$ has no fixed points in the open interval (0,1). Hence $p_n \downarrow 0$ as $n \to \infty$, and the frequency of the deleterious allele A converges to zero in either case.

Intuitively, the rate at which $p_n \to 0$ for $p_n = f(p_{n-1})$ will be influenced by the value of $f'(0)$. Note $f'(0) = \lim_{p\to 0} f(p)/p = 1 - s < 1$ in (9.3) if A is dominant and $f'(0) = 1$ in (9.4) if A is recessive. (In fact, $f'(0) = 1$ and $f''(0) = -2s$ in (9.4).)

Biologically, a dominant deleterious allele should be lost faster than a deleterious recessive for the following reason. By the Hardy-Weinberg law (8.3), an allele with frequency p will be found in homozygote form with frequency p^2 and in heterozygote form with frequency $2pq$. Hence rare alleles will be found mostly in heterozygotes, which have normal fitness for a deleterious recessive. Thus one would expect that a deleterious recessive would show more long-term resistance to selection than a dominant.

Theorem 9.1. *Assume* $0 < p < 1$*, and that* $f(p)$ *is a three-times continuously differentiable function of the unit interval* [0,1] *into itself such that*

$$\text{(i) } f(0) = 0, \ f(1) = 1,$$
$$\text{(ii) } 0 < f(p) < p < 1 \quad \text{for} \quad 0 < p < 1$$

If $0 < f'(0) = w < 1$*, then for any fixed* p *with* $0 < p < 1$*,*

$$p_n = f^{[n]}(p) \sim C_1 w^n \qquad \text{for some} \quad C_1 > 0 \qquad (9.5)$$

If $f'(0) = 1$ *and* $f''(0) = -\alpha < 0$*,*

$$p_n = f^{[n]}(p) \sim 2/(\alpha n) \qquad \text{as } n \to \infty \qquad (9.6)$$

Corollary 9.1. *If* A *is a deleterious dominant allele as in (9.3), then* $p_n \sim C(1-s)^n \to 0$ *exponentially fast. If* A *is deleterious recessive allele as in (9.4),* $p_n \sim 1/(sn) \to 0$ *at a slower rate.*

Thus a deleterious dominant will be lost fairly quickly unless it is continually renewed by mutation, while eliminating a deleterious recessive can take a very long time. Note that there has only been at most 150 human generations since Classical Greek times, assuming 17 years per generation.

Proof. The arguments in Theorem 9.1 are standard in the theory of branching processes, but are fairly easy to obtain. In general, since $p_{k+1} = f(p_k)$,

$$\begin{aligned} \log p_n \ &= \ \log p_0 \ + \ \sum_0^{n-1} (\log p_{k+1} - \log p_k) \\ &= \ \log p_0 \ + \ \sum_0^{n-1} \log\big(f(p_k)/p_k\big) \end{aligned}$$

To prove (9.5), assume $f'(0) = w < 1$. A three-term Taylor expansion about 0 yields

$$\begin{aligned} \log p_n \ &= \ \log p_0 \ + \ \sum_0^{n-1} \log\big(w + {}^1\!/\!_2 p_k f''(\theta_k)\big) \qquad (9.7) \\ &= \ \log p_0 + n \log w \ + \ \sum_0^{n-1} \log\big(1 + {}^1\!/\!_2 p_k f''(\theta_k)/w\big) \end{aligned}$$

where $0 < \theta_k < p_k$. By (ii), $p_n \downarrow p_\infty$ where $f(p_\infty) = p_\infty$, so that $p_n \downarrow 0$. Since $f(0) = 0$, $p_{k+1} = f(p_k) = f'(\theta_k)p_k$ by the mean-value theorem, where $0 < \theta_k < p_k$ and $f'(0) = w < 1$. Hence $p_{k+1} \le w_1 p_k \le C w_1^{k+1}$ for all $k \ge 1$ and sufficiently large C, where $w < w_1 < 1$. In particular, $\sum_{k=1}^\infty p_k < \infty$.

Since $f''(\theta)$ is bounded by assumption, the last sum in (9.7) is absolutely convergent. This implies that $\log p_n = n \log(w) + C_n$ where $\lim_{n\to\infty} C_n = C_0$ converges. This, in turn, implies $p_n \sim C_3 w^n$ as $n \to \infty$ for $C_3 = \exp(C_0)$.

To prove (9.6), assume $f'(0) = 1$ and $f''(0) = -\alpha < 0$. A four-term Taylor expansion implies

$$p_{n+1} = f(p_n) = p_n\big(1 - {}^1\!/\!_2 \alpha p_n + {}^1\!/\!_6 f'''(\theta_n)p_n^2\big)$$

where $0 < \theta_n < p_n$. The identity $1/(1-x) = 1 + x + O(x^2)$ implies

$$1/p_{n+1} \ = \ 1/p_n + {}^1\!/\!_2\alpha + \text{terms of order } p_n \qquad (9.8)$$

Since $p_n \to 0$, the sum of the last two terms in (9.8) eventually becomes bounded from below, so that $1/p_n \ge \gamma n$ for sufficiently large n. Thus $p_n \le C/n$ for some C. Then by (9.8)

$$\big|\,\big(1/p_{n+1} \ - \ 1/p_n\big) \ - \ {}^1\!/\!_2\alpha\,\big| \ \le \ C/n \qquad (9.9)$$

By summation in (9.9)

$$1/p_n = {}^1\!/\!_2\alpha n + \text{terms of order } \log n$$

This completes the proof of Theorem 9.1.

Exercise 9.1. The case of *intermediate dominance* is

$$w(AA) < w(Aa) < w(aa)$$

or $0 < h < 1$ in (9.2). At what rate is the allele A driven out? If the rate is exponential, what is the replacement for w in (9.5)? Could you have guessed it in advance from the fact that rare genes are found mostly in heterozygotes?

1.10. Heterozygote Advantage and Selection-Mutation Balance

A genetic locus is said to be *polymorphic* in a population if there are two or more alleles in the population at that locus. One possible cause would be if heterozygotes were more fit than homozygotes, as in the sickle-cell anemia example. For definiteness, consider an autosomal locus with two possible alleles, A and a, and assume the heterozygote Aa is selectively favored over both homozygotes AA and aa. If we normalize by setting the heterozygote relative fitness equal to one, we would have

$$w(AA) = 1 - s, \quad w(Aa) = 1, \quad w(aa) = 1 - t \tag{10.1}$$

where $s, t > 0$. Let p_n be the frequency of the A-allele in the n^{th} generation. Then by (8.6)

$$p_{n+1} \;=\; \frac{(1-s)p^2 + pq}{(1-s)p^2 + 2pq + (1-t)q^2}$$

where $p = p_n = 1 - q$. Hence $p_{n+1} = f(p_n) = f^{[n+1]}(p_0)$ for

$$f(p) \;=\; \frac{p(1-sp)}{1 - sp^2 - tq^2} \tag{10.2}$$

As before $f(0) = 0$ and $f(1) = 1$, but now $f(p) = p$ has an additional root

$$p_{cent} = t/(s+t)$$

Simple algebraic computations show

$$\begin{aligned} 0 < p < p_{cent} \quad &\text{implies} \quad 0 < p < f(p) < p_{cent}, \quad \text{and} \\ p_{cent} < p < 1 \quad &\text{implies} \quad p_{cent} < f(p) < p < 1 \end{aligned} \tag{10.3}$$

Moreover

$$f'(p_{cent}) = R = (s + t - 2st)/(s + t - st) < 1$$

If $0 < p_0 < 1$, (10.3) implies $p_n \to p_{cent}$. The same argument as in Theorem 9.1 now implies

Theorem 10.1. *Assume $p_{n+1} = f(p_n)$ for $f(p)$ in* (10.2). *Then $p_n \to p_{cent}$ if $0 < p_0 < 1$, and also*

$$p_n - p_{cent} \sim C_2\, R^n \qquad \text{as } n \to \infty, \quad \text{where } R < 1$$

Thus, if heterozygotes are most fit and both alleles are initially present, the frequency of A converges to an intermediate value and neither allele is eliminated from the population. However, at equilibrium the allele with the fitter homozygote has proportionately the higher frequency.

Since an individual has two loci at a typical autosomal locus, it may be to his advantage to have one allele that is good during cold weather, for example, and another that is advantageous during warm weather. This may make the heterozygote selectively more fit than either homozygote. Effects such as this may be an important cause of polymorphism in nature. Due to its importance, there are many synonymous terms for heterozygote advantage; two common ones are *overdominance* and *heterosis*. A related but not quite identical situation occurs when two separate populations have both been subject to a great deal of inbreeding. If two individuals, one from each population, are mated, the offspring are often observed to be healthier than either parent (this is called "*hybrid vigor*"). This is most likely due to the separate populations fixing at recessive deleterious alleles at different loci, which then become heterozygote in the offspring.

A more common source of polymorphism is when a deleterious allele is replenished by mutation from normal alleles, and so remains in the population in spite of selection. This is called *selection-mutation balance*. For definiteness, assume there are two alleles A and a, where A is deleterious with respect to a. Assume that the effect on the frequency p of A of one generation of selection is given by $f(p)$, where, for example, $f(p)$ is (9.3) or (9.4). Let u be the mutation rate per generation from a to A. Since deleterious alleles are usually rare, mutation from A to a will be ignored. Depending on whether mutation follows selection or selection follows mutation in each generation, the frequency of A after one generation will be

$$\begin{aligned} &\text{(a) } f(u,p) \;=\; f(p) + u\big(1 - f(p)\big) \;=\; u + f(p)(1-u) \qquad (10.4) \\ \textit{or} \quad &\text{(b) } f(u,p) = f\big(p + u(1-p)\big) \end{aligned}$$

In general, mutation occurs either in the germ line or basic reproduction tissue of an individual (and then would generally not affect the individual otherwise) or occurs in the process of meiosis (the formation of a sperm or egg cell) or fertilization. If individuals are scored as AA, Aa, or aa at birth, then mutation during that generation would appear to follow selection, since later germline mutations within an individual generally do not affect a trait undergoing selection. Then $f(u,p)$ would be given by (10.4a).

If individuals are scored at sexual maturity instead of at birth, then selection would follow mutation and (10.4b) would apply, although the genotype frequencies of the individuals would not be in Hardy-Weinberg equilibrium except in the case of genic selection (that is, unless (8.7) holds).

To make things even more complicated, mutation can also occur in *somatic cells*; that is, cells not in the germ line. These could affect the *phenotype* (or *physical form*) of an individual but not its genotype (as determined by its germ line), so that an individual with one (germ-line) genotype might be subject to the selective effects of a different genotype. A somatic-cell mutation during gestation might also cause the genotype of an individual to be scored incorrectly at birth. To have a significant effect, a somatic-cell mutation would have to happen within the first few cell divisions of an embryo, and might even affect some parts of the body and not others. While somatic-cell mutations in the first few embryonic cell divisions do occur (and can sometimes lead to tragic situations), they are rare and will be ignored here.

Generally, u is in the range $10^{-6} - 10^{-4}$, with for example $s \geq 0.01$. Thus it is usually safe to assume $0 < u \ll s < 1$. In the following, subscripts denote partial derivatives, and $O(u^2)$ means "up to terms that can be bounded by u^2 for small u".

Theorem 10.2. *Assume $f(u,p)$ is a three-times continuously differentiable function of $u, p \in [0,1]$ such that*

$$\begin{aligned} &\textit{(i) } 0 \leq f(u,p) \leq 1, && u,p \in [0,1], \qquad (10.5) \\ &\textit{(ii) } 0 < f(0,p) < p < 1, && p \in (0,1), \\ &\textit{(iii) } f_u(0,0) > 0 \end{aligned}$$

and assume one of the two conditions

$$\begin{aligned} &\textit{(a) } f_p(0,0) = w = 1 - s < 1 \qquad \textit{or} \qquad (10.6) \\ &\textit{(b) } f_p(0,0) = 1, \;\; f_{pp}(0,0) = -\alpha < 0, \qquad \textit{and} \\ &\qquad f_p(u,p) < 1 \;\; \textit{for} \;\; 0 < u < \delta, \; 0 < p < \epsilon \end{aligned}$$

for some $\delta > 0$, $\epsilon > 0$. Let $p_{n+1} = f(u, p_n)$ for $n \geq 0$, where $0 < p_0 < 1$. Then, there exists $u_0 > 0$ such that $\lim_{n\to\infty} p_n = p(u)$ exists if $0 < u < u_0$, where, if (10.6a) holds,

$$p(u) \;=\; \left(\frac{f_u(0,0)}{1 - f_p(0,0)}\right) u \qquad + \; O(u^2). \qquad (10.7)$$

If, instead, (10.6b) holds,

$$p(u) \;=\; \sqrt{-2u\frac{f_u(0,0)}{f_{pp}(0,0)}} \qquad +\; O(u). \tag{10.8}$$

Deferring the proof of Theorem 10.2 for the moment, note that (10.5) and (10.6a) hold if A is dominant (i.e. (9.3) with (10.4a) or (10.4b)), and (10.5)-(10.6b) hold if A is recessive (i.e. $f(p)$ is given by (9.4)). Thus

$$\begin{aligned} p(u) = \quad & \frac{u}{s} && +\; O(u^2) \qquad A \text{ dom., mut. follows sel.} \\ = \quad & \frac{u(1-s)}{s} && +\; O(u^2) \qquad A \text{ dom., sel. follows mut.} \\ = \quad & \sqrt{\frac{u}{s}} && +\; O(u^2) \qquad A \text{ recessive} \end{aligned} \tag{10.9}$$

where we have abbreviated several key words by their first three letters.

Since we are assuming $u \ll s$, the deleterious allele A will be rare in either case. The frequency of Aa individuals will be be approximately twice the values in (10.9), since each individual carries two genes at that locus (see (8.3)). Heterozygote individuals will be affected if A is dominant, while only homozygotes will be affected if A is recessive. Thus the frequency of affected *individuals* is $\approx 2u/s$ if A is dominant and $p(u)^2 \approx u/s$ if A is recessive. Since $u/s \ll \sqrt{u/s}$ if $u \ll s$, the allele A will be much more common (although mostly hidden) in the recessive case.

An intuitive explanation for the first two formulas in (10.9) can be given as follows. Since A is deleterious, A genes are constantly being lost from the population, and each A gene must have been created by mutation at some time in the past. The set of A genes that were created n generations ago have been exposed to selection n times. Since $f(p) \approx wp$ by (9.3) if A is rare, each generation of selection will decrease the frequency of these genes by a factor of w. If mutation follows selection, the sum of the present frequencies of these genes is $u + uw + uw^2 + \ldots = u/(1-w) = u/s$. If selection follows mutation, the sum is $uw + uw^2 + \ldots + = uw/s$.

Note that this argument allows us to estimate u and s separately. The frequency at birth of Aa homozygotes whose parents are unaffected (i.e., both are aa) is $2u$, since there are two genes that could mutate. Hence the ratio of the frequency of Aa's with unaffected parents (that is, the frequency of Aa's that are due to fresh mutations) to the frequency of all newborn Aa's is $\approx 2u/(2u/s) = s$. In particular, if $w = s = 1/2$, then approximately half of observed Aa's are due to fresh mutations and about half have inherited their A gene from their parents.

Proof of Theorem 10.2: For any $\epsilon > 0$, (10.5ii) implies $p - f(u,p) \ge \delta > 0$ for all $p \in [\epsilon, 1-\epsilon]$ for $0 < u < \delta$ and some $\delta > 0$. Since we can choose $\epsilon < 1 - p_0$, it follows that $\limsup_{n\to\infty} p_n \le \epsilon$ if $u < \delta$. First, assume (10.6a). Then $|f_p(u,p)| \le \psi < 1$ for $u < \delta,\ p < \epsilon$, and

$$|p_{n+1} - p_n| = |f(u,p_n) - f(u,p_{n-1})| \le \psi\,|p_n - p_{n-1}| \le \psi^n |p_1 - p_0|$$

by the mean value theorem. Since $\psi < 1,\ p_n \to p(u)$ where $f(u,p(u)) = p(u)$, and by Taylor's Theorem

$$\begin{aligned} p(u) \;&=\; f_u(0,0)u + f_p(0,0)p(u) + O(u^2 + p(u)^2) \\ &=\; \frac{f_u(0,0)u}{1 - f_p(0,0)} + O(u^2 + p(u)^2) \end{aligned} \tag{10.10}$$

It only remains to prove $p(u) = O(u)$. Since $u \le 1$, (10.10) implies $p(u) \le Au + Bp(u)^2$ for constants A, B. If $p(u) \le \epsilon < 1/2B$, we conclude $p(u) \le Au/\big(1 - Bp(u)\big) \le 2Au$. Hence $p(u) = O(u)$ in (10.10) and (10.7) follows.

If, instead, (10.6b) holds,

$$\begin{aligned} p_{n+1} &= f(u,p_n) = f(u,p_n) - f(0,p_n) + f(0,p_n) \qquad (10.11) \\ &= u\big(f_u(0,0) + O(u+p_n)\big) + p_n - {}^1\!/\!{}_2\alpha p_n^2\big(1 + O(p_n)\big) \end{aligned}$$

In particular $p_n \ge \gamma u > 0$ where $\gamma > 0$. If $u > 0$ is assumed fixed and small, $|f_p(u,p)| \le \psi < 1$ by (10.6b) for $0 < \gamma u \le p \le 3\epsilon$, where $\limsup_{n\to\infty} p_n \le \epsilon$. Hence $p_n \to p(u)$ as before, and by (10.11)

$$1/2\alpha p(u)^2 = uf_u(0,0)\big(1 + O(u + p_n)\big)$$

The relation (10.8) follows.

If we had only assumed $f_p(0,0) = 1$ and $f_{pp}(0,0) = -\alpha < 0$ in (10.6b), we wouldn't be able to show convergence of $\{p_n\}$. However, (10.11) does imply that (10.8) holds with p_n in place of $p(u)$ for all sufficiently large n, and so (10.8) holds in this sense.

Exercise 10.1. Suppose $1 - s$ and $1 - t$ in (10.1) are replaced by $1 + s$ and $1 + t$ respectively, where $s, t > 0$. Then the heterozygote is less fit than either homozygote. This situation is called, appropriately enough, *heterozygote disadvantage* or *negative heterosis* or sometimes even *underdominance*.

Whichever term you prefer, prove that there exists a critical frequency p_{crit} such that if $0 < p_0 < p_{crit}$, then $p_n \to 0$, whereas if $p_{crit} < p_0 < 1$, then $p_n \to 1$.

Can you calculate the exponential rates at which these convergences occur, using Theorem 9.1? Could you have guessed these rates in advance?

1.11. Selection in Finite Populations

Consider a monoecious population of N individuals with two alleles A and a at a particular locus, with fitnesses

$$w(AA) = 1 + s, \qquad w(Aa) = 1 + hs, \qquad w(aa) = 1$$

and mutation rates $u\colon a \to A$ and $v\colon A \to a$ per gene per generation. Let p_n be the frequency of A at the beginning of the n^{th} generation. By Section 2, the distribution of p_{n+1} given p_n can be represented symbolically as

$$\{p_{n+1} \mid p_n = p\} \;\approx\; \frac{B(2N, f(p))}{2N} \qquad (11.1)$$

where $B(2N, p)$ denotes a binomial distribution, and $f(p)$ is the probability of sampling an A-gene in generation $n + 1$ if the frequency of A was p at the beginning of the n^{th} generation. Let $fs(p)$ denote the change in p due to selection alone. Then by (9.1)–(9.2)

$$fs(p) \;=\; p\,\frac{1 + s\big(h + p(1-h)\big)}{1 + sp\big(2h + p(1-2h)\big)}.$$

If frequencies are measured at birth, and if mutation is assumed to act when sperm and eggs are generated, then mutation appears to follow selection, and in (11.1)

$$\begin{aligned} f(p) &= (1-v)fs(p) + u\big(1 - fs(p)\big) \\ &= (1-u-v)fs(p) + u \qquad (11.2) \end{aligned}$$

Assume selection, mutation, and genetic drift all have about the same strength. I.e.,

$$s \sim \frac{\sigma}{2N}, \qquad u \sim \frac{\alpha}{2N}, \qquad v \sim \frac{\beta}{2N}, \quad \text{as} \quad 2N \to \infty \tag{11.3}$$

where α, $\beta > 0$. As in Section 3, $\{p_n\}$ defines a recurrent Markov chain on $\{0, 1/2N, \dots, 1\}$ if u, $v > 0$. We now try to find the limit of the stationary distribution $\mu(2N, dp)$ as $N \to \infty$. Note

$$\begin{aligned} fs(p) - p \;&=\; p\,\frac{s\big(h + p(1-h)\big) - sp\big(2h + p(1-2h)\big)}{1 + sp\big(2h + p(1-2h)\big)} \\ &=\; sp(1-p)\,\frac{h + p(1-2h)}{1 + sp\big(2h + p(1-2h)\big)} \end{aligned}$$

Thus by (11.2)

$$\begin{aligned} 2N\,E(\;\; p_{n+1} - p \;\;|\; p_n = p) \;&=\; 2N\big(f(p) - p\big) \\ \to\; m(p) \;&=\; \sigma p(1-p)\big(h + p(1-2h)\big) + \alpha - (\alpha+\beta)p \\ 2N\,E\big((p_{n+1} - p)^2 \;|\; p_n = p\big) \;&=\; f(p)\big(1 - f(p)\big) + 2N\big(f(p) - p)\big)^2 \\ \to\; a(p) = p(1-p), &\qquad \text{and} \\ 2N\,E\big(|p_{n+1} - p|^3 \;|\; p_n = p\big) \;&\to\; 0 \end{aligned} \tag{11.4}$$

As in (6.1)–(6.6), the stationary distribution $\mu(2N, dp)$ converges as $2N \to \infty$ to $h(p)dp$, where $h(p)$ is a solution of

$$\tfrac{1}{2}\big(a(p)h(p)\big)' - m(p)h(p) = 0$$

Substituting $m(p)$ and $a(p)$ from (11.4)

$$h(p) \;=\; C\,p^{2\alpha-1}(1-p)^{2\beta-1}e^{\sigma p\left(2h+p(1-2h)\right)}$$

The simplest case $h = 1/2$ corresponds to the heterozygote having fitness halfway between the two homozygotes. In that case

$$h(p) \;=\; h(p, \alpha, \beta, \sigma) \;=\; C\,e^{\sigma p}p^{2\alpha-1}(1-p)^{2\beta-1} \tag{11.5}$$

where we write $h(p, \alpha, \beta, \sigma)$ to emphasize the dependence on those parameters. We can find the constant C in (11.5) in closed form:

$$\begin{aligned} C^{-1} \;&=\; \int_0^1 e^{\sigma p}\,p^{2\alpha-1}(1-p)^{2\beta-1}\,dp \\ &=\; \sum_0^\infty \frac{\sigma^n}{n!}\int_0^1 p^{n+2\alpha-1}(1-p)^{2\beta-1}\,dp \;=\; \sum_0^\infty \frac{\sigma^n}{n!}\,\frac{\Gamma(2\alpha{+}n)\Gamma(2\beta)}{\Gamma(2\alpha{+}n{+}2\beta)} \\ &=\; \frac{\Gamma(2\alpha)\Gamma(2\beta)}{\Gamma(2\alpha{+}2\beta)}\sum_0^\infty \frac{(2\alpha)^{(n)}}{(2\alpha{+}2\beta)^{(n)}}\,\sigma^n \\ &=\; \frac{\Gamma(2\alpha)\Gamma(2\beta)}{\Gamma(2\alpha{+}2\beta)}\;{}_1F_1(2\alpha;\, 2\alpha{+}2\beta;\, \sigma) \end{aligned} \tag{11.6}$$

where $x^{(n)} = x(x+1)\dots(x+n-1)$, In (11.6), ${}_1F_1(a;\, c;\, z)$ is called the confluent hypergeometric function or Kummer's function (Magnus et al 1966, Chapter VI).

The same argument can be used to find the moments of the limiting distribution $h(p)dp$ in terms of ${}_1F_1$:

$$\begin{aligned}\int_0^1 & p^m(1-p)^n h(p,\alpha,\beta,\sigma)\,dp \\ &= \frac{(2\alpha)^{(m)}(2\beta)^{(n)}}{(2\alpha+2\beta)^{(m+n)}}\;\frac{{}_1F_1(2\alpha+m;\; 2\alpha+2\beta+m+n;\; \sigma)}{{}_1F_1(2\alpha;\; 2\alpha+2\beta;\; \sigma)}\end{aligned} \tag{11.7}$$

While confluent hypergeometric functions are not as easy to deal with as trigonometric or exponential functions, an immense number of identities and asymptotic formulas are known for them. In particular, by (11.7) and Magnus et al (1966, p289),

$$\int_0^1 p^m(1-p)^n\, h(p,\alpha,\beta,\sigma)\,dp \;\sim\; \frac{(2\beta)^{(n)}}{\sigma^n} \qquad \text{as} \quad \sigma\to\infty \tag{11.8}$$

Exercise 11.2 shows how (11.8) can be used to find the limiting distribution of the frequency of the allele a for large σ.

As mentioned earlier in Section 3, it can be shown that (11.4) implies that $\{p_n\}$ can be approximated by a continuous-time diffusion process $\{X_t\}$ with $t \approx n/2N$ and 'infinitesimal generator'

$$\mathcal{L}f(x) \;=\; {}^1\!/\!{}_2 a(p)f''(x) + m(p)f'(x)$$

Now, assume there is no mutation; i.e. $u = v = 0$. Then 0 and 1 are traps for $\{p_n\}$. Similarly, it can be shown that with probability one X_t is trapped for some $t < \infty$ at either 0 or 1. Let

$$s(x) = P(X_t \text{ eventually trapped at } 1 \mid X_0 = x) \tag{11.9}$$

If $k/2N \to x$ as $N \to \infty$, it can be shown

$$\lim_{N\to\infty} P(p_n \text{ event. trapped at } \ 1 \mid p_0 = k/2N) \;=\; s(x) \tag{11.10}$$

Moreover, $s(x)$ in (11.9) is the unique solution of $\mathcal{L}s(x) = 0$ with $s(0) \;=\; 0$ and $s(1) \;=\; 1$. If $\alpha = \beta = 0$ and $h = 1/2$ in (11.4), then $m(p) = {}^1\!/\!{}_2\sigma p(1-p)$ and $a(p) = p(1-p)$. Solving

$$\mathcal{L}s(x) = {}^1\!/\!{}_2 p(1-p)\big(s''(x) + \sigma s'(x)\big) = 0$$

and applying (11.10),

$$\begin{aligned}\lim_{N\to\infty} & P(\text{population eventually fixes at } A \mid p_0 = k/2N) \\ &= \; s(x) \;=\; \frac{1-e^{-\sigma x}}{1-e^{-\sigma}}.\end{aligned} \tag{11.11}$$

For example, suppose that the initial frequency of A is $p_0 = 0.25$, and $\sigma = 12$. Since $w(AA) = 1 + s \approx 1 + \sigma/2N$, this could be very weak selection if $2N$ is large. By (11.11), $s(0.25) \approx 0.950$, and the population will eventually fix at A with $\approx 95\%$ probability.

As another example, consider the problem of the fate of a new, selectively advantageous gene in a large population. We would like to take $p_0 = X_0 = 1/2N$ in (11.11) for fixed $s > 0$. However, if $s > 0$ is fixed, $\sigma = 2Ns \to \infty$, while σ, α, and β are assumed fixed and finite in (11.3)–(11.11).

Also, $p_0 = 1/2N$ implies $p_0 \to x = 0$ and $s(0) = 0$, so a literal use of (11.11) would give no information. While the theory does not exactly apply to what we want, we might try $p_0 = 1/2N$ and $\sigma = 2Ns \to \infty$ in (11.11) as an approximation and see what happens. By (11.11),

$$\begin{aligned} &P(\text{population eventually fixes at } A \mid \text{initially one copy}) \\ &\approx\; s(1/2N) = \frac{1-e^{-\sigma/2N}}{1-e^{-\sigma}} \;\approx\; \frac{1-e^{-s}}{1-e^{-2Ns}} \;\approx\; s + O(s^2) \end{aligned} \tag{11.12}$$

The survival probability in (11.12) can be written $s + O(s^2) = 2hs + O(s^2)$, where $1 + hs$ is the fitness of heterozygotes Aa. Recall that it was found by branching process methods in Exercise 8.4 that the probability of survival was $2s + O(s^2)$, where $1 + s$ was the fitness of heterozygotes. Hence (11.12) is consistent with Exercise 8.4.

Suppose now that (11.1) holds but s, u, and v are fixed. Suppose further that $f(p)$ in (11.2) has a stable fixed point γ; i.e.

$$|f(p) - \gamma| \;\le\; \psi\,|p - \gamma| \qquad \text{for} \quad \psi < 1, \quad 0 \le p \le 1$$

(this can usually be weakened to $|f'(\gamma)| < 1$.) If $2N$ is large, p_n will spend most of its time near γ, since the deterministic forces of selection and mutation are now much stronger than genetic drift. However, one can still ask for the error in approximating p_n by γ, or ask how long it will take for p_n to vary from a neighborhood of γ. It turns out that

$$p_n \;\approx\; \gamma + \frac{Y_n}{\sqrt{2N}}$$

where $\{Y_n\}$ is a sequence of correlated Gaussian random variables. Moreover, if $p_0 = \gamma$ and c is any preassigned constant, p_n will remain within $O(\sqrt{\log 2N/2N})$ of γ for $0 \le n \le N^c$ with probability $1 - O(N^{-c})$ (Sawyer 1983). If $\epsilon > 0$ is fixed and T_N is the first n such that p_n is outside of $(\gamma - \epsilon, \gamma + \epsilon)$, there exists a constant $b > 0$ (for which an expression is found) such that

$$\lim_{N\to\infty} P(e^{(b-\delta)N} \le T_N \le e^{(b+\delta)N}) \;=\; 1 \qquad \text{for all } \delta > 0$$

(Morrow and Sawyer 1987). Also, if $\{p_n\}$ has an equilibrium distribution in [0,1], the equilibrium probability that p_n is outside $(\gamma - \epsilon, \gamma + \epsilon)$ is exponentially small as a function of $2N$ (ibid.).

There a large literature of "diffusion approximation" results in continuous time in which deterministic forces are asymptotically much stronger than random forces. In these results, the approximating process is typically a fixed path plus a Gaussian process times a small parameter. See e.g. Kurtz (1971, 1981), Norman (1975b), Nagylaki (1986), Sawyer (1983), and Morrow and Sawyer (1987).

Exercise 11.1. Explain in words why it is more likely that the power of σ in (11.8) should depend on n and not m or $n + m$.

Exercise 11.2. A random variable $Z = Z(\theta, \lambda)$ is said to have a *gamma distribution with parameters θ and λ* if

$$P\big(Z(\theta,\lambda) \le z\big) \;=\; \frac{\lambda^\theta}{\Gamma(\theta)} \int_0^z x^{\theta-1} e^{-\lambda x}\,dx \qquad \text{for all } z \ge 0 \tag{11.13}$$

In particular $E\big(Z(\theta,\lambda)\big) = \theta/\lambda$ and $Z(\theta,\lambda) \approx (1/\lambda)Z(\theta,1)$. (The latter means that these two random variables have the same probability distribution.)

Let $p_\infty(a)$ be the limiting equilibrium distribution under (11.3) of the allele a. Use (11.8) to prove that

$$\lim_{\sigma\to\infty} P\big(\sigma p_\infty(a) \le t\big) \;=\; P\big(Z(2\beta,1) \le t\big) \tag{11.14}$$

for all $t \ge 0$, for $Z(2\beta,1)$ as in (11.13). (*Hint:* Write $E\big(p_\infty(a)^n\big)$ in terms of (11.7). Use (11.8) to show that the moments of $\sigma p_\infty(a)$ converge to those of Z and apply the method of moments to prove (11.14).)